DOI:10.11918/202101029

组合动作空间深度强化学习的人群疏散引导方法

薛怡然,吴 锐,刘家锋

(模式识别与智能系统研究中心(哈尔滨工业大学),哈尔滨 150001)

摘 要: 人群疏散引导系统可在建筑物内发生灾害时有效保护生命安全,减少人员财产损失。针对现有人群疏散引导系统需要人工设计模型和输入参数,工作量大且容易造成误差的问题,本文提出了基于深度强化学习的端到端智能疏散引导方法,设计了基于社会力模型的强化学习智能体仿真交互环境。使智能体可以仅以场景图像为输入,通过与仿真环境的交互和试错自主学习场景模型,探索路径规划策略,直接输出动态引导标志信息,指引人群有效疏散。针对强化学习深度Q网络(DQN)算法在人群疏散问题中因为动作空间维度较高,导致神经网络复杂度指数增长的"维度灾难"现象,本文提出了将Q网络输出层按动作维度分组的组合动作空间DQN算法,显著降低了网络结构复杂度,提高了系统在多个引导标志复杂场景中的实用性。在不同场景的仿真实验表明本文方法在逃生时间指标上优于静态引导方法,达到人工构造模型方法的相同水平。说明本文方法可以有效引导人群,提高疏散效率,同时降低人工构造模型的工作量并减小人为误差。

关键词: 神经网络;强化学习;疏散引导;人群仿真;深度Q网络

中图分类号: TP183 **文献标志码:** A **文章编号:** 0367-6234(2021)08-0029-10

Crowd evacuation guidance based on combined action-space deep reinforcement learning

XUE Yiran, WU Rui, LIU Jiafeng

(Pattern Recognition and Intelligent System Research Center (Harbin Institute of Technology), Harbin 150001, China)

Abstract: Crowd evacuation guidance systems are of great significance for protecting lives and reducing personal and property losses during disasters in buildings. Existing crowd evacuation guidance systems require the manual design of models and input parameters, incurring significant workloads and potential errors. An end-to-end intelligent evacuation guidance method based on deep reinforcement learning was proposed, and an interactive simulation environment based on the social force model was designed. The agent could automatically learn a scene model and explore the path planning strategy by interacting with simulation environment and through trial and error with only scene images as input, and then directly output dynamic signage information, thus achieving the crowd evacuation guidance efficiently. Aiming to solve the "dimension disaster" phenomenon of deep Q network (DQN) algorithm caused by high dimension action space and complex network structure in crowd evacuation, a combined action-space DQN algorithm was proposed. The algorithm grouped the output layer nodes of the Q network according to action dimensions, significantly reduced the network complexity, and improved the practicality of the system in complex scenes with multiple guidance signs. Experiments in different simulation scenes demonstrate that the proposed method is superior to the static guidance method in evacuation time and on par with the manually designed model method. It shows that the proposed method can effectively guide the crowd, improve the evacuation efficiency, and reduce the workload and artificial errors of manually designed models.

Keywords: neural network; reinforcement learning; evacuation guidance; crowd simulation; deep Q network (DQN)

大型商场、写字楼等多功能建筑在满足人们多种需求的同时,建筑复杂程度逐渐提高。在发生地震、火灾等灾害时,建筑内复杂的结构对人群疏散逃生形成阻碍,对生命安全形成新的威胁。灾害发生时,人群由于对建筑物环境不了解、视野受限、心理恐慌等因素,难以准确找到最优逃生路线[1]。在从众心理的影响下,逃生者容易形成拥堵甚至踩踏,造成更大损失[2]。如何引导人群以最有效的路径疏散,对灾害中保护生命安全,减少人员财产损失具有重要意义。

为了在灾害发生时引导人群有效疏散,研究者开发了多种基于动态引导标志的人群疏散引导系统[3-6]。此类系统可以对建筑场景建模,收集灾害位置和人群分布等实时信息,用路径规划算法找出

收稿日期: 2021-01-10
基金项目: 国家自然科学基金(61672190)
作者简介: 薛怡然(1992—),男,博士研究生
通信作者: 吴 锐,simple@hit.edu.cn

最优逃生路径,通过动态引导标志诱导人群的运动状态,有效地提高了紧急情况下人群逃生效率。但是,现有的人群疏散引导系统都离不开人工设计基于拓扑图或者网格形式的场景模型、根据场景特征手动输入模型参数等工作,人工工作量较大并且容易引入人为因素造成的误差,对后续路径规划等计算步骤造成干扰。

针对此问题,本文提出了基于深度强化学习算法的端对端的人群疏散引导方法。即训练一种仅以建筑平面图为输入,在与环境的交互和反馈中自动探索学习场景模型和路径规划方法,发现最优动作策略,直接输出动态引导标志信息的疏散引导智能体。为实现此方法,设计了基于社会力模型人群动力学仿真的强化学习智能体仿真交互环境,并针对深度强化学习中典型深度 Q 网络(DQN)[7]方法应用于人群疏散引导时出现的"维度灾难"问题,提出了组合动作空间的 DQN 方法,降低了网络结构复杂度,提高了算法在复杂建筑场景中的实用性。

1 相关工作

1.1 人群仿真与疏散引导

人群运动仿真是人群疏散研究中分析人群行为特征、自组织等现象的重要基础。人群仿真研究可分为宏观模型与微观模型。宏观模型主要考察人群整体的运动状态,一般采用元胞自动机等栅格模型[8]。例如用流体力学方法计算速度场,再作用于个体的高密度人群仿真方法[9],和基于格子玻尔兹曼模型的人群异常检测方法等[10]。微观模型用动力学方法仿真每个个体的运动特征,典型方法有引入人的主观因素的社会力模型[11-12]。

在仿真研究中,研究者希望提高疏散效率,使人群运动更贴近现实,因此人群疏散中的路径规划问题受到研究者的关注。有研究者分别利用群体智能的布谷鸟算法[13]和结合心理因素的 A * 算法[14]改进路径搜索方法。也有研究者结合多种传感器信息,例如构建威胁态势信息场的路径优化方法[15]、感知灾害位置的路径选择方法[16]、根据路径和出口容量优化的路径选择模型[17]等。仿真环境的路径规划方法可以综合环境信息,计算使全局疏散效率最高的逃生路径。在实际场景中,逃生者由于视野和经验受限,只能掌握自身周边信息,建筑物监测系统即使可以掌握优化的逃生路径,也需要专门途径告知逃生者。

为了指示逃生路线,大型建筑内一般设置有应急逃生标志。应急标志可分为静态引导标志和动态引导标志两类[18]。在真实场景实验[19]和基于社会力模型的仿真实验[20]中,静态引导标志都对疏散效率起到了重要的正面作用。不同于静态引导标志仅能指示一种预设的疏散路线,动态引导标志可根据灾害场景中人群分布等实时条件显示不同的引导信息。研究表明在某个出口不可用时,动态引导标志可以有效诱导人群从其他出口疏散[21],在路径中发生危险时,动态标志也能引导人群避开不安全的路线[22]。

将上述人群仿真环境、路径规划算法和动态引导标志相结合,研究者开发出了多种人群疏散引导系统。此类系统以建筑物环境模型为基础,实现了从场景信息感知、疏散路径规划到人群运动诱导的闭环反馈,具备一定的实用价值[3]。例如在拓扑图模型上基于网络流路径规划的动态引导方法[4]、使用仿真摄像机采集人群密度信息,应用实时最短路算法的动态疏散系统[5]。还有研究与现实建筑系统相结合,建立平行应急疏散系统框架,取得了更大现实意义[6]。

此类系统基本流程包含输入场景平面图、人工构建拓扑图或网格模型、根据通道容量等因素输入模型参数、应用路径规划算法和设置动态引导标志信息等几个步骤。其中构建模型和填写参数几个步骤的人工参与度高,工作量大,容易由于人为失误造成误差并在后续步骤中放大,使系统疏散效率受到影响。针对此问题,本文利用深度强化学习方法,提出端到端的动态人群疏散引导系统。

近年来,强化学习方法在人群疏散研究中得到了一些应用。研究者开发了数据驱动的强化学习人群仿真方法,用智能体模拟和预测个体的运动[23]。在路径选择问题上,有研究利用逆向强化学习方法使机器人模仿人类行动轨迹[24]。这些仿真研究目标是接近真实场景,而非优化疏散效率。对于疏散引导问题,一些研究者开发出多种使用强化学习智能体输出机器人运动方向,控制机器人在人群中运动,从而干涉人群运动状态,提高疏散效率的方法[25-27]。此类方法在单个路口的仿真实验中取得了一定效果,但在实际应用中存在加剧人群拥挤、引发踩踏事故等隐患。现有基于强化学习的研究将逃生者个体或机器人个体定义为智能体。与此不同的是,本文将疏散引导系统定义为强化学习智能体,其以场景图像为观测输入,输出遍布场景的多组动态引导标志信号,从而诱导人群运动,提高疏散效率。

1.2 深度强化学习

强化学习[28]是人工智能领域的重要组成部分之一,是一种通过与环境的交互和试错,学习从环境状态到动作的映射,发现最优行为策略,以使从环境

获得的积累奖赏最大的学习方法。结合深度神经网络，深度强化学习智能体能直接以图像作为输入，将特征提取和值函数估计等过程内化在网络结构中，显著拓展了智能体的感知和决策能力。深度强化学习的标志性成果包括在 Atari 视频游戏中超越人类玩家水平的 DQN 方法[7]、在围棋中战胜人类顶级选手的 AlphaGo[29] 和在星际争霸 2 游戏在线对战中打入大师级排行的 AlphaStar[30] 等。

强化学习模型[28] 基于马尔可夫决策过程（MDP），可描述为四元组 (S,A,P_a,R_a)，其中 S 为所有状态的集合，即状态空间，A 为动作空间，状态转移函数 $P_a(s,s')=P(s_{t+1}=s'\mid s_t=s,a_t=a)$ 表示在状态 s 时智能体执行动作 a，环境进入状态 s' 的概率，奖励函数 $R_a(s,s')$ 表示在状态 s 执行动作 a 进入状态 s' 时所获得的即时奖励。智能体在每个离散的时间步 t，观测环境状态 s_t，根据策略 π: $S\rightarrow A$ 选择动作 $a_t=\pi(s_t)$ 作用于环境，环境反馈给智能体奖励 r_t，并转移到下一个状态 s_{t+1}。智能体与环境的交互过程见图 1。

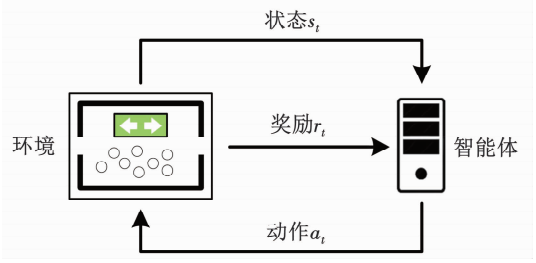

图 1　强化学习模型示意

Fig. 1　Schematic of reinforcement learning model

在强化学习的 MDP 模型基础上，定义状态 - 动作值函数，也可称作动作值函数

$$Q_\pi(s_t,a_t)=E_\pi[r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots] \quad (1)$$

表示按策略 π，在状态 s_t 时执行动作 a_t 之后获得的期望累积奖励，其中 γ 为奖励衰减系数。谷歌 DeepMind 团队开发的 DQN 方法[7] 用深度神经网络表示动作值函数，且分为参数为 θ 的当前 Q 网络和参数为 θ^- 的目标 Q 网络，每隔一定时间将当前 Q 网络参数复制到目标 Q 网络中。DQN 的策略 π 为贪婪策略，总是选择当前状态下 Q 值最大的动作，训练时加入一定概率选择随机动作作为探索过程。DQN 使用经验池存储和管理样本，对于一个时间步的样本 $e_t=(s_t,a_t,r_t,s_{t+1})$，计算时序差分（temporal difference，TD）误差

$$\delta_t=r_t+\gamma\max_a Q(s_{t+1},a;\theta^-)-Q(s_t,a_t;\theta) \quad (2)$$

对于从经验池中采样得到的一个批次样本数据 $B=\{e_1,\cdots,e_t\}$，网络损失函数定义为平方误差损失：$L(\theta)=E_B(\delta_t^2)$，然后用误差反向传播算法更新网络参数。

DQN 在以图像为输入的 Atari 视频游戏等任务上取得突破。研究者在 DQN 的基础上，提出了用当前 Q 网络进行目标动作选取的 Double DQN（DDQN）方法[31]、用 TD 误差区分经验池中样本优先级的优先经验回放[32] 等改进方法。

然而，DQN 输出的动作空间是离散的，并且对每一种可能的动作组合使用一个输出层节点进行评价，因此当动作维数增加时，网络复杂度将以指数方式增长。在人群疏散引导问题中，智能体以动态引导标志的显示状态作为输出动作，每个标志的离散动作形成独立的动作维度。在复杂建筑场景中动态引导标志数目较多时，DQN 的输出层规模将变得过于庞大而使算法无法实现。

2　基于组合动作空间 DQN 的疏散引导

2.1　人群疏散引导的强化学习模型

人群疏散引导问题涉及 3 类对象，包括建筑场景、逃生者和智能疏散引导系统。现有研究常将每个逃生者个体定义为一个智能体，研究个体的行动策略和运动状态，或添加可动机器人个体作为智能体。与此不同的是，本文将疏散引导系统看作一个智能体，如图 2 所示，则智能体所处的环境包括实际建筑场景和其中运动的人群。建筑场景由平面图表示，摄像机等多种传感器收集人群运动状态，绘制进场景平面图，此图像即包含了当前环境中所需的信息，连续多帧图像的灰度位图组合成（width × height × depth）三维张量，定义为 MDP 的环境状态 $s_t\in S$。对于多层建筑，可以将不同楼层平面图拼接成整体场景图像输入系统，从而实现多层建筑中的疏散引导。疏散引导系统通过动态引导标志显示信号，诱导、干涉人群运动，因此智能体动作 $\boldsymbol{a}_t\in A$ 对应引导标志信号，$\boldsymbol{a}_t$ 是离散向量，每个维度对应一个引导标志，取值为此标志显示状态（向左、向右等）之一。由于环境和人群运动较为复杂，状态转移函数 $P_a(s,s')$ 是未知的，需要智能体在交互过程中学习和适应。奖励函数的设计决定了智能体的优化方向和学习目的，在人群疏散问题中，应根据成功撤离的人数或疏散所用时间等因素设计奖励函数。本文定义 $R_a(s,s')=-1$，即每个时间步固定给予惩罚，智能体的学习目标是使累积惩罚最小，即全体人群疏散时间最短。

强化学习智能体的训练过程需要与环境不断交互，在探索和试错中学习。其所需的交互规模十分庞大，一般在数万个周期、百万个时间步以上。并且训练初期智能体知识不足，可能造成更多潜在危险。

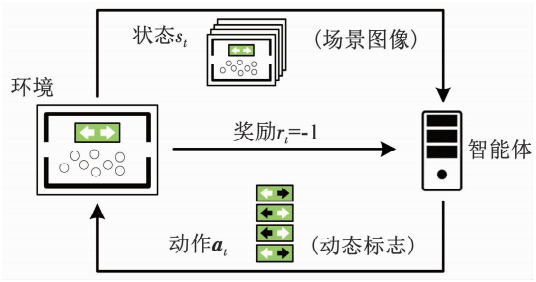

图 2 人群疏散引导的强化学习模型

Fig. 2 Reinforcement learning model for evacuation guidance

因此用于智能疏散引导系统的强化学习智能体必须在仿真环境中进行训练,训练完成后再部署到实际建筑内。

疏散引导系统智能体通过与仿真环境的大量交互进行探索与学习,最终得到神经网络形式的优化策略 $\pi(s)$。学习过程中不需要人工设计建筑通道拓扑图或网格模型,智能体能自主发现和优化引导策略,不需要另外设计路径规划等中间算法。实际应用中,每个时刻 t 传感器收集人群运动信息,将人群位置分布、当前引导标志显示状态等信息绘制进场景平面图。利用卷积神经网络对图像的感知能力,多帧场景灰度位图组成三维张量传入神经网络,作为智能体输入的观测状态 s_t,智能体根据训练完成后包含优化策略的神经网络计算动作向量 $\boldsymbol{a}_t = \boldsymbol{\pi}(s_t)$,由动态引导标志显示对应信号,实现对人群疏散的有效引导。

2.2 组合动作空间 DQN

在网络结构上,DQN 采用多层卷积神经网络处理图像输入,然后连接多层全连接神经网络,输出层每一个神经元对应一种可能的离散动作组合。对于动作中相互独立的成分,总的动作空间是各个独立动作空间的笛卡尔积。当动作空间有 n 个相互独立的维度,每个维度有 m 个离散动作时,DQN 网络需要 m^n 个输出层节点,以对应输入状态 s 时不同动作 $Q(s,\boldsymbol{a})$ 的值。因此,随着独立动作数目的增长,DQN 的网络结构复杂度以指数速度增长,从而使算法不可实现。同时输出层过多也会导致样本利用率降低和网络参数更新困难。这个现象被称为 DQN 的"维度灾难"问题。

在人群疏散引导的应用当中,智能体动作定义为引导标志的显示状态。即使每个引导标志只有向左和向右两个状态,对于 n 个引导标志,总的动作空间容量也会达到 2^n 之多,引发"维度灾难"。本文针对此问题,提出组合动作空间的 DQN 方法(CA-DQN)。如图 3 所示,对于相互独立的动作维度,每个维度对应 Q 函数网络输出层一组节点,每组包含这个维度上的所有离散动作。这个改变可看作对每个动作维度 d 设置了各自的值函数 $Q_d(s,a^{(d)};\theta)$,并且共用一套网络参数。此时网络输出层节点数目是各个维度上离散动作数之和,随独立动作数目的增长速度从指数增长降为线性增长,例如 n 个引导标志所需输出层节点为 $2n$。

在 CA-DQN 中,动作 $\boldsymbol{a}_t = (a_t^{(1)}, a_t^{(2)}, \cdots, a_t^{(D)})$ 为 D 维组合动作向量。对每个维度 d,智能体用贪婪策略选择动作

$$a_t^{(d)} = \underset{a}{\operatorname{argmax}} Q_d(s_t, a; \theta) \tag{3}$$

对于一个样本 $e_t = (s_t, \boldsymbol{a}_t, r_t, s_{t+1})$ 定义每个维度上的 TD 误差

$$\delta_t^{(d)} = r_t + \gamma \max_a Q_d(s_{t+1}, a; \theta^-) - Q_d(s_t, a_t^{(d)}; \theta) \tag{4}$$

结合研究者提出用当前 Q 网络选择 $t+1$ 时间动作,以避免过高估计的 DDQN 算法[31],TD 误差进一步定义为

$$\delta_t^{(d)} = r_t + \gamma Q_d(s_{t+1}, \underset{a}{\operatorname{argmax}} Q_d(s_{t+1}, a; \theta); \theta^-) - Q_d(s_t, a_t^{(d)}; \theta) \tag{5}$$

则从经验池中采样所得一组样本 $B = \{e_1, \cdots, e_t\}$,神经网络的损失函数定义为平方误差损失的算术平均值

$$L(\theta) = E_B\left(\frac{1}{D}\sum_d (\delta_t^{(d)})^2\right) \tag{6}$$

神经网络按式(6)定义的损失函数用误差反向传播算法进行训练。此时,对于每个样本,动作的每个维度都有一个输出层节点被选择并参与 TD 误差的计算和网络误差的反向传播,则共有 D 个输出层节点可以得到更新。相比 DQN 中每个样本只能更新一个输出层节点,CA-DQN 方法提高了样本的利用效率。

2.3 组合动作空间 DQN 的优先经验回放

DQN 以随机方式从经验池中采样,不考虑样本差异,样本利用效率较低。采用优先经验回放方法[32],用式(2)定义的样本 TD 误差,将样本采样优先级定义为 $p_t = (|\delta_t| + \varepsilon)^\alpha$,其中 ε 和 α 为常数。TD 误差绝对值越大的样本意味着所包含的有效信息越多,对其赋予更高采样优先级,可提高样本利用率和训练效率。

CA-DQN 中一个样本的 TD 误差按式(5)定义,是一个向量 $\boldsymbol{\delta}_t = (\delta_t^{(1)}, \cdots, \delta_t^{(d)})$,应用优先经验回放方法时,样本优先级定义为各维度 TD 误差绝对值的平均值

$$p_t = \left(\frac{1}{D}\sum_d |\delta_t^{(d)}| + \varepsilon\right)^\alpha \tag{7}$$

样本优先级定义为平均值可能使样本重要性被其他动作维度稀释,但有助于保持训练过程的稳定性。

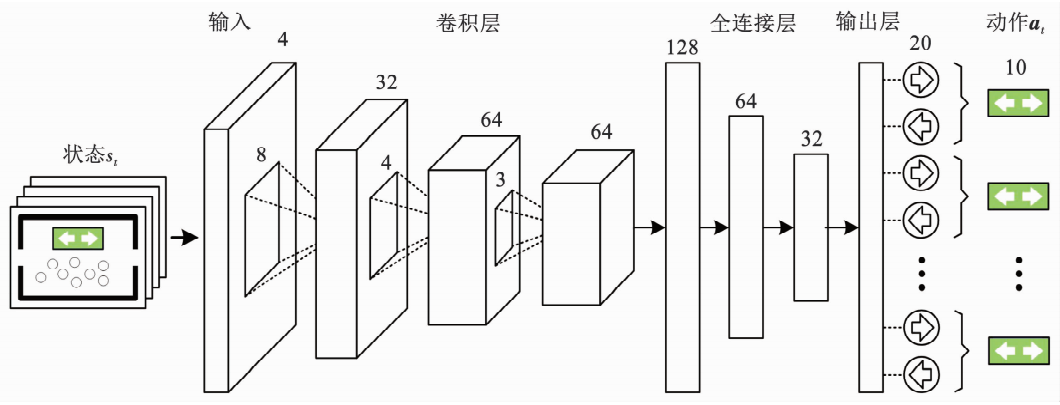

图 3　CA-DQN 神经网络结构

Fig. 3　Network structure of CA-DQN

结合以上优先经验回放方法，CA-DQN 的训练过程如下：

算法 1：智能体训练过程

输入： 仿真环境 env

输出： 神经网络参数 θ^*

随机初始化神经网络参数 θ，θ^-
初始化经验池 pool
while steps < total_steps **then**
　state，reward，terminate ← env. RandomInit( ) // 随机初始化仿真环境
　while not terminate **then**
　　action ← AgentPolicy(state，θ) // 按式(3)选择动作
　　state_new，reward，terminate ← env. Step(action)
　　td_error ← CalcTDError(state，action，reward，state_new，θ，θ^-) // 按式(5)计算 TD 误差
　　priority ← CalcPriority(td_error) // 按式(7)计算样本优先级
　　pool. Append(state，action，reward，state_new，priority)
　　state ← state_new
　　steps ← steps + 1
　　s，$\boldsymbol{a}$，r，s' ← pool. RandomSample(batch_size) // 按优先级随机采样
　　td_error ← CalcTDError(s，$\boldsymbol{a}$，r，s'，θ，θ^-) // 按式(5)计算 TD 误差
　　loss ← CalcLoss(td_error) // 按式(6)计算损失函数
　　θ ← BackPropagation(θ，loss) // 更新网络参数
　　每隔一定步数 $\theta^- \leftarrow \theta$
　end
　每隔一定周期数计算平均周期回报，若性能提升 $\theta^* \leftarrow \theta$
end

3　实验与分析

3.1　实验设计与实现

本文采用基于社会力模型的人群动力学仿真系统[5]作为智能体交互环境，构造了典型的多房间、双出口室内场景，以下称为“场景 1”。多层建筑场景可通过拼接各层平面图输入疏散引导系统，本文为直观起见，采用单层仿真场景。仿真系统计算每个个体的运动状态，并加入个体心理因素对运动造成的影响。仿真系统基于 C + + 语言和 Qt 库编写。

如图 4 所示，仿真场景大小为 29.2 m × 19.7 m，平面图像素为 499 × 337，场景内包含左右 2 个出口和 6 个房间，上下 2 个通道连接房间和出口，每个通道设置 5 个动态引导标志，标志可显示相对两个方向之一。人群数量为 200 人，初始位置以圆形范围随机分布，分布中心和半径取值范围为 $x \in (100, 140)$，$y \in (60, 280)$，$r \in (100, 200)$。场景图像中，蓝色直线表示墙壁，绿色矩形表示出口位置，绿色箭头表示动态引导标志，每个标志有相反两个方向的显示状态，蓝色圆点表示逃生者个体，灰色部分为不可到达区域。个体最大运动速度为 5 m/s。仿真个体在没有看到疏散引导标志时，选择距离最近的出口，按照静态最短路线逃生，看到疏散引导标志时，按照引导标志指示的方向逃生。仿真系统动力学计算的每个时间步为 40 ms，仿真时间上限为 100 s。

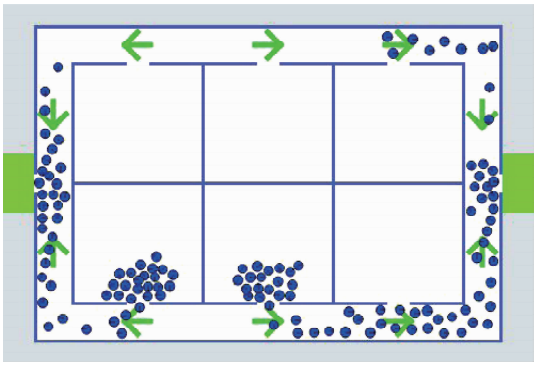

图 4 仿真场景 1

Fig. 4 Simulation scene 1

同时，本文也采用原交互环境中基于实际建筑平面图的仿真场景[5]进行实验，如图 5 所示，以下称为“场景 2”。场景图像中符号含义与场景 1 相同。场景大小为 47.0 m × 28.8 m，图像像素为 805 × 494，人群数量为 200 人，分布中心和半径取值范围为 $x \in (100,700)$，$y \in (60,440)$，$r \in (300,500)$，场景内共有 2 个出口和 6 个动态引导标志。不同仿真场景的强化学习智能体由于输入输出定义不同，疏散策略不同，需要分别进行训练。

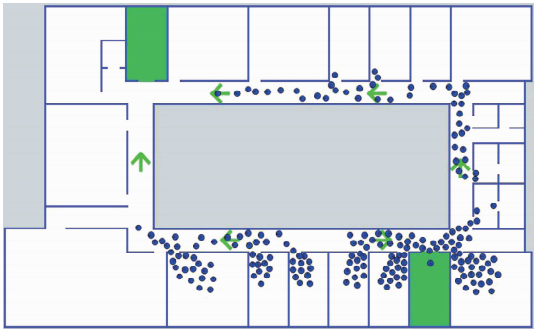

图 5 仿真场景 2

Fig. 5 Simulation scene 2

CA-DQN 方法基于 Python 语言、TensorFlow 平台和 OpenAI/baseline 库实现。实现过程与超参数的选择参考了 baseline 库中用于 Atari 视频游戏的 DQN 方法，并针对本文方法进行适当调整。强化学习智能体的每个时间步中，首先由仿真系统进行 5 步计算，即仿真 200 ms 内人群的运动状态，将获得的最后 4 帧图像下采样为 1/2 大小的灰度图，以场景 1 为例，组合成像素为 249 × 168 的 4 通道图像，作为智能体的状态 s_t 输入值函数 Q 网络。Q 网络结构如图 3 所示，由三层卷积神经网络和三层全连接神经网络组成。第一层由 32 组 8 × 8 卷积核组成，输入为 249 × 168 × 4 的三维张量，第二层由 64 组 4 × 4 卷积核组成，第三层由 64 组 3 × 3 卷积核组成。卷积神经网络的激活函数为 ReLU。三层全连接层神经元数目分别为 128、64、32，激活函数为 ReLU。输出层激活函数为恒等函数，20 个神经元分为 10 组，每组 2 个中取输出值较大的作为一个动态引导标志的显示信号，共同组成 10 维离散输出向量作为智能体动作 $\boldsymbol{a}_t$。$\boldsymbol{a}_t$ 作用于仿真系统，改变 10 个引导标志显示的方向，从而指引人群运动方向，此时智能体与仿真环境的交互完成一个循环。智能体每步的奖励固定为 −1，即每秒获得 −5 的奖励，智能体训练目标为减少总体疏散时间。

训练超参数中，批量大小为 64，学习率为 10^{-5}，总时间步为 10^7，经验池样本容量为 10^5，每 2×10^4 步将当前 Q 网络参数复制到目标 Q 网络。实验硬件平台为 AMD Threadripper 2990WX CPU、NVIDIA RTX 2080Ti GPU、128 GB 内存。

3.2 实验结果与分析

由于原 DQN 方法用于本文实验时，以场景 1 为例，需设置 2^{10} = 1 024 个输出层节点，相比 CA-DQN 的 20 个节点，DQN 网络规模过大，在现有条件下难以实现。因此本文选择基于静态引导标志的方法和基于拓扑图建模和动态 Dijkstra 最短路方法的疏散引导算法[5]作为对比。静态引导标志方法中，用自动或人工的最短路方法计算，每个标志指向距离最近的出口，每个场景仅计算一次，不考虑人群实时分布，模拟过程中标志不发生变化。动态 Dijkstra 最短路方法需要专家人员根据地图内通道结构人工建立拓扑图模型，并且设置多个虚拟摄像头节点，统计通道不同位置的人群密度，实时调整拓扑图各边权值，用 Dijkstra 算法进行路径规划，实现人群的有效疏散。实验结果中，每 1 s 疏散时间对应 −5 的周期奖励。

由图 6 的训练曲线看出，对于场景 1，智能体在约 3×10^4 个训练周期后达到最优策略，此时智能体与仿真环境交互次数约为 6.4×10^6 个时间步。图 7 中，对于场景 2，智能体在约 4.5×10^4 个训练周期后

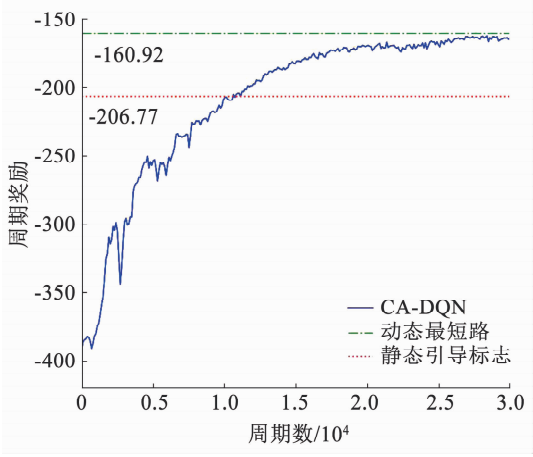

图 6 场景 1 智能体训练曲线

Fig. 6 Training curve of agent in scene 1

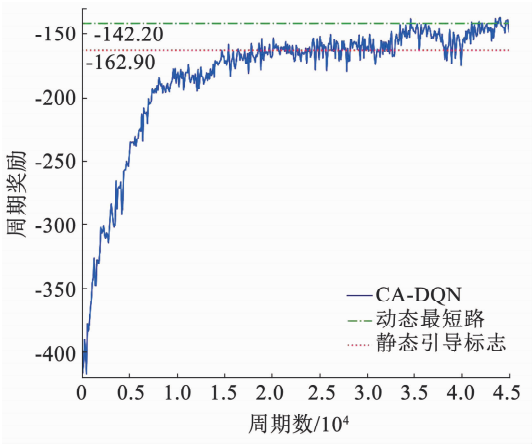

图 7　场景 2 智能体训练曲线

Fig. 7　Training curve of agent in scene 2

达到最优策略。如表 1 所示，对不同疏散方法使用新的随机人群分布参数进行 100 个周期的疏散仿真，场景 1 中智能体训练所得最优策略的平均周期奖励为 −158.25，即平均疏散时间为 31.65 s，优于使用静态引导标志的 41.35 s 和动态 Dijkstra 最短路方法的 32.18 s。场景 2 中智能体训练所得最优策略平均疏散时间为 27.33 s，优于静态引导标志和动态最短路方法。说明本文基于 CA-DQN 的智能疏散引导智能体可以有效引导人群疏散。

表 1　不同方法疏散时间

Tab. 1　Evacuation time under different methods　　s

引导方法	场景 1 疏散时间	场景 2 疏散时间
静态引导标志	41.35	32.58
动态最短路	32.18	28.44
CA-DQN	**31.65**	**27.33**

图 8 展示了场景 1 中一个典型的疏散过程（图中符号含义请参考 3.1 节）：图 8(a)是人群的初始分布，人群主要分布于左侧 4 个房间，若没有动态指引，人群按到出口距离最短的静态标志疏散策略，将造成左侧出口拥堵，右侧出口得不到有效利用。在图 8(b)到图 8(d)时刻，智能体感知到人群分布，将左上方房间人群引向左侧出口，其余人群引向右侧出口。图 8(e)时刻，左侧出口拥堵已得到缓解，右侧出口预期撤离人数较多，因此智能体将左下区域剩余人群引向左侧出口。最终在图 8(f)时刻，人群

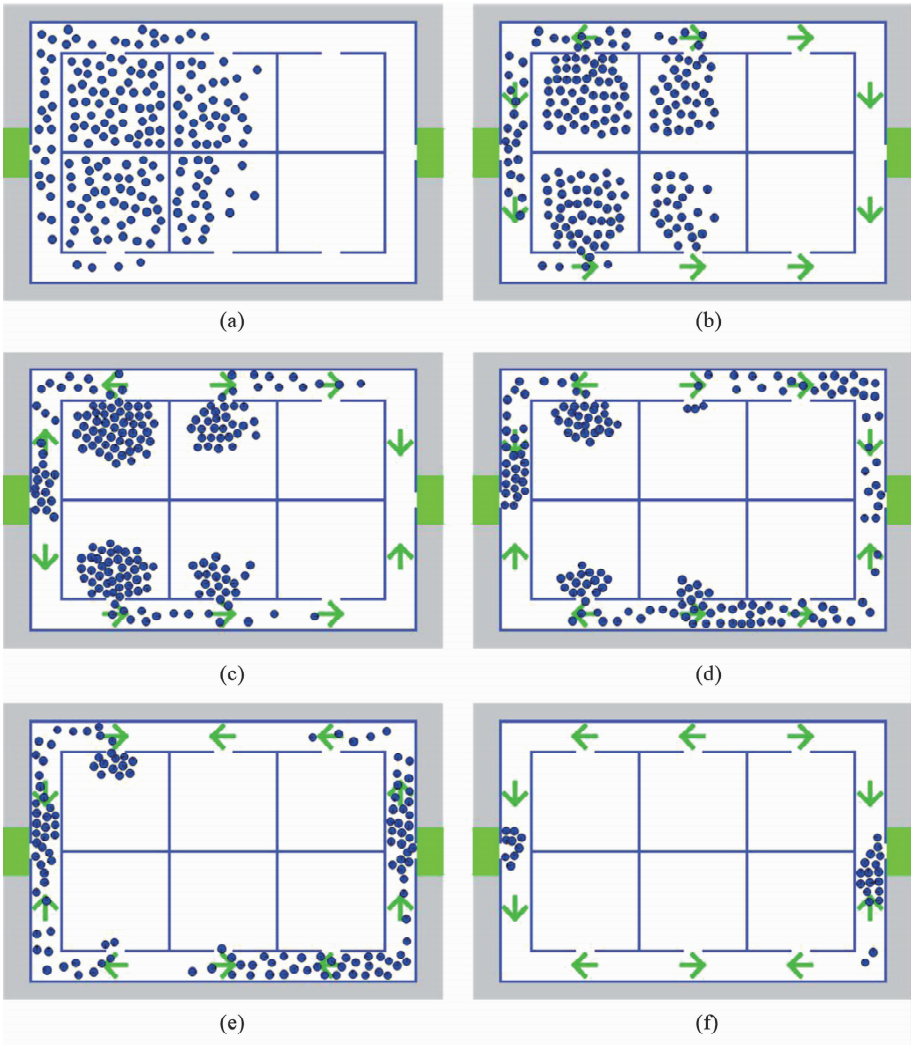

图 8　场景 1 一个周期的疏散过程

Fig. 8　Typical evacuation process in scene 1

基本同时从两侧出口完成疏散,表明人群疏散引导智能体实现了人群疏散效率的最大化。

类似地,图 9 展示了场景 2 中典型的一个疏散过程。图 9(a)中,人群初始化分布主要集中在场景上方。图 9(b)时刻,智能体感知人群分布,将左上方房间以外的大部分区域人群向右下方出口诱导。图 9(c)到图 9(d)时刻,一部分人群有效地转移至右侧通道,避免了左上方出口进一步拥堵。最终,在图 9(e)到图 9(f)时刻,人群基本同时从两个出口完成疏散,说明智能体的引导实现了人群疏散效率最大化。

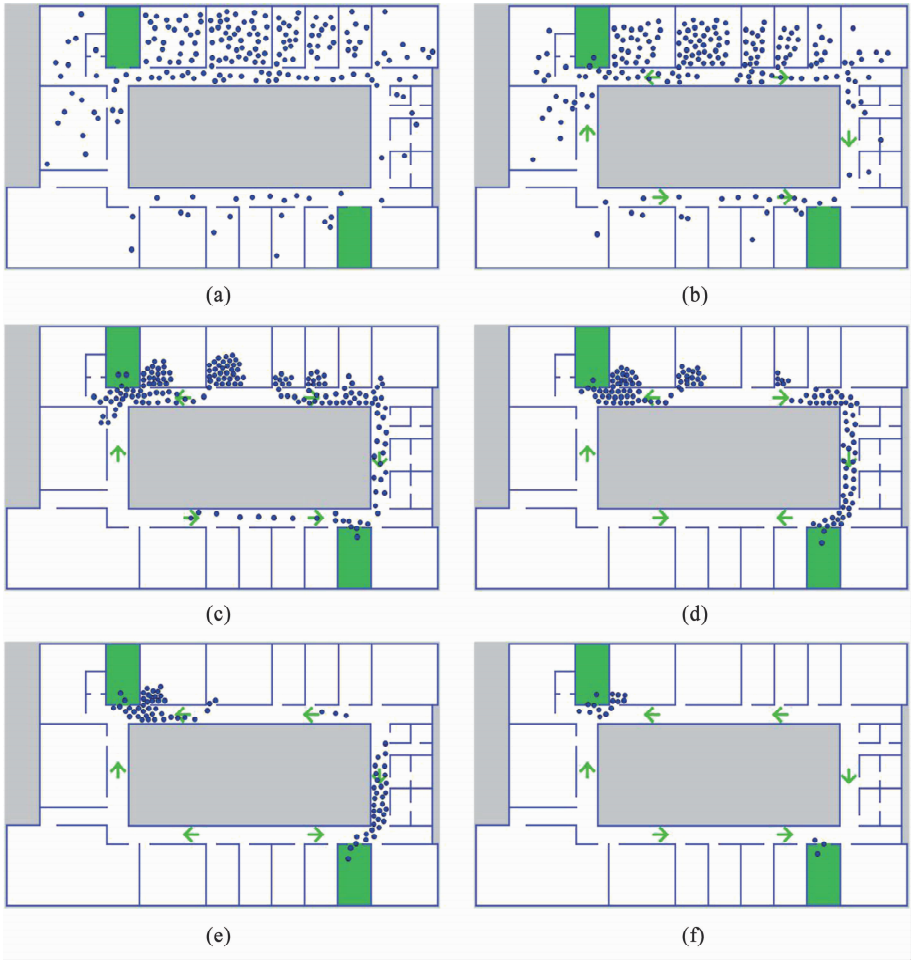

图 9 场景 2 一个周期的疏散过程

Fig. 9 Typical evacuation process in scene 2

改变仿真场景初始化人数,分别进行 100 个周期的疏散仿真,不同方法的疏散效果对比见图 10。场景 1 中,在人数较少时,各个通道都能保持通畅,静态引导方法效果较好。人群数量增加时,静态引导方法受影响较大,CA-DQN 和动态最短路方法可以避免人群拥堵。人群数量增加到 80 人以上时,两种动态方法疏散效果优于静态方法,其中本文 CA-DQN 方法实现了最优疏散引导效率。场景 2 的实验也显示出类似结果,由图 11 看出,本文方法在不同人群数量下均能取得较好效果。

实验结果显示,相比静态标志不能感知人群分布信息,本文基于 CA-DQN 的强化学习人群疏散引导方法能动态地调整引导标志的显示信号,有效提高人群疏散效率。与基于拓扑图建模的动态 Dijkstra 最短路方法相比,本文方法取得了更好的疏散引导效率,同时避免人工构造拓扑图的工作量和潜在的人工误差。

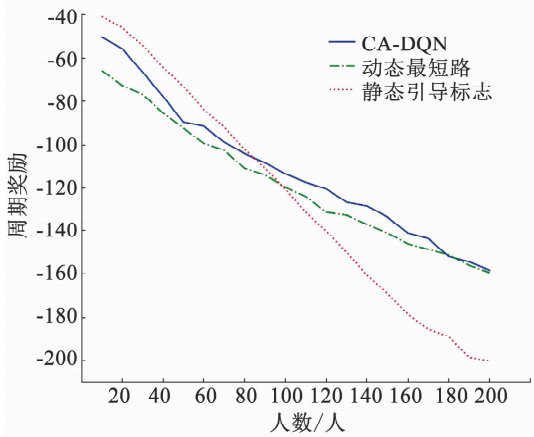

图 10 场景 1 中不同人数的周期奖励

Fig. 10 Period reward with varying number of persons in scene 1

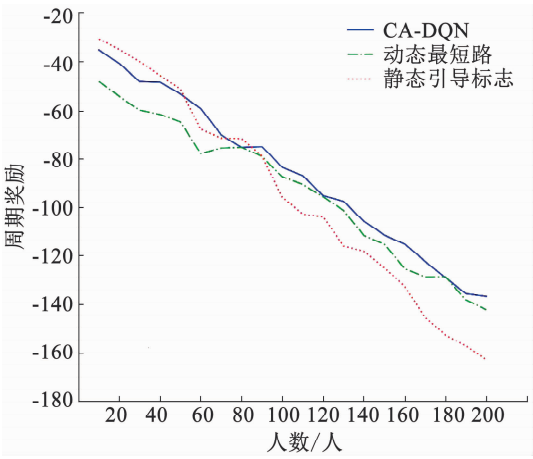

图 11　场景 2 中不同人数的周期奖励

Fig. 11　Period reward with varying number of persons in scene 2

4　结　论

本文分析了使用动态引导标志的人群疏散引导问题，针对现有方法需要人工设计拓扑图模型或网格模型，配合独立的路径规划算法，导致人工工作量大、容易引入人为误差等不足之处，提出了基于组合动作空间深度强化学习的人群疏散引导方法。通过端对端的深度学习，由智能体在训练过程中自行探索学习建筑结构和路径规划方法，通过环境反馈自动修正认知误差，从而找到最优的疏散引导策略。

针对深度强化学习中典型的 DQN 方法应用于人群疏散问题时因输出的动态引导标志数量较多而出现的“维度灾难”问题，本文提出 CA-DQN 网络结构，将关于输出动作维度的网络结构复杂度从指数级增长降低为线性增长，提高了强化学习方法在复杂场景和大规模人群疏散问题中的可用性。在基于社会力模型的人群动力学仿真系统中的实验表明，本文方法相对静态引导标志有效提升了人群疏散效率，减少疏散时间，达到与基于人工建模的动态最短路方法相同水平。

未来工作将进一步提升强化学习智能体在复杂场景中的训练效率，对输出信号变更频率等加以更多限制，使其在真实场景中更易理解。

参考文献

[1] 刘翠娟，刘箴，柴艳杰，等. 人群应急疏散中一种多智能体情绪感染仿真模型[J]. 计算机辅助设计与图形学学报，2020，32(4)：660
LIU Cuijuan，LIU Zhen，CHAI Yanjie，et al. A multi-agent emotional contagion model in crowd emergency evacuation[J]. Journal of Computer-Aided Design & Computer Graphics，2020，32(4)：660

[2] 苟成秋，余瀚游，徐梓桉，等. 基于信息非对称性的小群组紧急疏散行为模拟[J]. 计算机辅助设计与图形学学报，2018，30(3)：524
GOU Chengqiu，YU Hanyou，XU Zian，et al. Simulation of small social group behaviors in emergency evacuation based on information asymmetry[J]. Journal of Computer-Aided Design & Computer Graphics，2018，30(3)：524

[3] RANH，SUN L，GAO X. Influences of intelligent evacuation guidance system on crowd evacuation in building fire[J]. Automation in Construction，2014，41：78

[4] DESMETA，GELENBE E. Capacity based evacuation with dynamic exit signs[C]//IEEE Annual Conference on Pervasive Computing and Communications Workshops. Budapest：IEEE，2014：332

[5] 赵巍，刘畅，廉兴宇，等. 人群运动仿真和疏散优化方法设计与实现[J]. 系统仿真学报，2014，26(3)：523
ZHAO Wei，LIU Chang，LIAN Xingyu，et al. Simulation of crowd movement and design and implementation of evacuation optimization method[J]. Journal of System Simulation，2014，26(3)：523

[6] 周敏，董海荣，徐惠春，等. 平行应急疏散系统：基本概念、体系框架及其应用[J]. 自动化学报，2019，45(6)：1074
ZHOU Min，DONG Hairong，XU Huichun，et al. Parallel emergency evacuation systems：basic concept，framework and applications[J]. Acta Automatica Sinica，2019，45(6)：1074

[7] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529

[8] MURAMATSUM，IRIE T，NAGATANI T. Jamming transition in pedestrian counter flow[J]. Physica A：Statistical Mechanics and its Applications，1999，267(3/4)：487

[9] 孙立博，孙晓峰，秦文虎. 基于连续模型和动力学仿真模型的高密度人群仿真算法[J]. 计算机学报，2016，39(7)：1375
SUN Libo，SUN Xiaofeng，QIN Wenhu. Dense crowd simulation based on continuum model and aggregate dynamics model[J]. Chinese Journal of Computers，2016，39(7)：1375

[10] XUE Y，LIU P，TAO Y，et al. Abnormal prediction of dense crowd videos by a purpose-driven lattice Boltzmann model[J]. International Journal of Applied Mathematics and Computer Science，2017，27(1)：181

[11] HELBING D，MOLNAR P. Social force model for pedestrian dynamics[J]. Physical Review E，1995，51(5)：4282

[12] 王爱丽，董宝田，王泽胜. 基于社会力的行人交通微观仿真模型研究[J]. 系统仿真学报，2014，26(3)：662
WANG Aili，DONG Baotian，WANG Zesheng. Modeling and simulation of microscopic pedestrian based on social force[J]. Journal of System Simulation，2014，26(3)：662

[13] 董崇杰，刘毅，彭勇. 改进布谷鸟算法在人群疏散多目标优化中的应用[J]. 系统仿真学报，2016，28(5)：1063
DONG Chongjie，LIU Yi，PENG Yong. Improved cuckoo search algorithm applied to multi-objective optimization of crowd evacuation[J]. Journal of System Simulation，2016，28(5)：1063

[14] 任治国，盖文静，彭群生. 疏散仿真中关注个体心理的路径规划[J]. 计算机辅助设计与图形学学报，2015，27(9)：1775
REN Zhiguo，GAI Wenjing，PENG Qunsheng. Ego-centered path planning in evacuation simulation[J]. Journal of Computer-Aided Design & Computer Graphics，2015，27(9)：1775

[15] 丁雨淋，何小波，朱庆，等. 实时威胁态势感知的室内火灾疏散路径动态优化方法[J]. 测绘学报，2016，45(12)：1464
DING Yulin，HE Xiaobo，ZHU Qing，et al. A dynamic optimization method of indoor fire evacuation route based on real-time situation awareness[J]. Acta Geodaetica et Cartographica Sinica，2016，45(12)：1464

[16] CUESTA A, ABREU O, BALBOA A, et al. Real-time evacuation route selection methodology for complex buildings[J]. Fire Safety Journal, 2017, 91: 947

[17] 韩延彬, 刘弘. 一种基于疏散路径集合的路径选择模型在人群疏散仿真中的应用研究[J]. 计算机学报, 2018, 41(12): 2653
HAN Yanbin, LIU Hong. Research on route choice model based on evacuation route set and its application in crowd evacuation simulation[J]. Chinese Journal of Computers, 2018, 41(12): 2653

[18] ZHOU M, DONG H, IOANNOU P A, et al. Guided crowd evacuation: approaches and challenges[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(5): 1081

[19] FU L, CAO S, SONG W, et al. The influence of emergency signage on building evacuation behavior: an experimental study [J]. Fire and Materials, 2019, 43(1): 22

[20] YUAN Z, JIA H, ZHANG L, et al. A social force evacuation model considering the effect of emergency signs[J]. Simulation, 2018, 94(8): 723

[21] GALEAE R, XIE H, DEERE S, et al. Evaluating the effectiveness of an improved active dynamic signage system using full scale evacuation trials[J]. Fire Safety Journal, 2017: 908

[22] CHO J, LEE G, LEE S, et al. An automated direction setting algorithm for a smart exit sign[J]. Automation in Construction, 2015: 139

[23] YAO Z, ZHANG G, LU D, et al. Data-driven crowd evacuation: a reinforcement learningmethod[J]. Neurocomputing, 2019, 366: 314

[24] HENRY P, VOLLMER C, FERRIS B, et al. Learning to navigate through crowded environments [C] //2010 IEEE International Conference on Robotics and Automation. [S. l.]: IEEE, 2010: 981

[25] 周婉, 胡学敏, 史晨寅, 等. 基于深度 Q 网络的人群疏散机器人运动规划算法[J]. 计算机应用, 2019, 39(10): 2876
ZHOU Wan, HU Xuemin, SHI Chenyin, et al. Motion planning algorithm of robot for crowd evacuation based on deep Q-network [J]. Journal of Computer Application, 2019, 39(10): 2876

[26] 谭峒, 刘士豪, 周婉, 等. 基于深度时空 Q 网络的机器人疏散人群的算法[J/OL]. 计算机工程, 2020
TAN Mei, LIU Shihao, ZHOU Wan, et al. Crowd evacuation algorithm by a robot based on deep spatio-temporal Q-network [J/OL]. Computer Engineering, 2020 DOI: 10.19678/j.issn.1000-3428.0057878

[27] WAN Z, JIANG C, FAHAD M, et al. Robot-assisted pedestrian regulation based on deep reinforcement learning [J]. IEEE Transactions on Cybernetics, 2018

[28] SUTTONR S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT press, 2018

[29] SILVERD, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484

[30] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350

[31] HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]// 30th AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2016: 2094

[32] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[C]// 4th International Conference on Learning Representations (ICLR). New York: Springer, 2016: 256

(**编辑** 苗秀芝)

封面图片说明

封面图片来自本期论文“块核范数的 RPCA 分解与熵权类稀疏的壁画修复”,是兰州交通大学电子与信息工程学院陈永教授课题组提出的块核范数的 RPCA 分解与熵权类稀疏的壁画修复字典构造过程示意图。首先,采用提出的基于块核范数的 RPCA 图像分解算法得到待修复壁画的结构层图像。然后,对结构层图像进行类稀疏修复时,采用提出的熵加权的 k-means 聚类算法,得到结构相似的各子类图像并构造生成相应的类字典。最后,通过奇异值分解和分裂 Bregman 迭代优化的类稀疏修复方法,完成结构层图像的重构。通过对真实敦煌壁画数字化修复的实验结果表明,该算法能够有效地保护壁画图像的边缘和纹理等重要特征信息,为敦煌壁画的人工修复提供了参考依据。

(图文提供:陈永,陶美风,陈锦. 兰州交通大学电子与信息工程学院)