文章编号:1005-0523(2023)01-0068-08

基于深度强化学习的机器人导航算法研究

熊李艳,舒垚淞,曾 辉,黄晓辉

(华东交通大学信息工程学院,江西 南昌 330013)

摘要:移动机器人穿越动态密集人群时,由于对环境信息理解不充分,导致机器人导航效率低且泛化能力弱。针对这一问题,提出了一种双重注意深度强化学习算法。首先,对稀疏的奖励函数进行优化,引入距离惩罚项和舒适性距离,保证机器人趋近目标的同时兼顾导航的安全性;其次,设计了一种基于双重注意力的状态价值网络处理环境信息,保证机器人导航系统兼具环境理解能力与实时决策能力;最后,在仿真环境中对算法进行验证。实验结果表明,提出的算法不仅提高了机器人导航效率还提升了导航系统的鲁棒性,主要表现为:在 500 个随机的测试场景中,碰撞次数和超时次数均为 0,导航成功率优于对比算法,且平均导航时间比最好的算法缩短了 2%;当环境中行人数量、导航距离发生变化时算法依然有效,且导航时间短于对比算法。

关键词:深度强化学习;奖励函数;状态价值网络;双重注意力

中图分类号:U495;TP242 **文献标志码**:A

本文引用格式:熊李艳,舒垚淞,曾辉,等. 基于深度强化学习的机器人导航算法研究[J]. 华东交通大学学报,2023,40(1):67-74.

Research on Robot Navigation Algorithm Based on Deep Reinforcement Learning

Xiong Liyan, Shu Yaosong, Zeng Hui, Huang Xiaohui

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: When the mobile robot passes through the dynamic dense crowd, due to the insufficient understanding of environmental information, the robot navigation efficiency is low and the generalization ability is weak. To solve this problem, a double-attention deep reinforcement learning algorithm is proposed. Firstly, the sparse reward function was optimized, and the distance penalty term and comfort distance were introduced to ensure that the robot approached the target while taking into account the safety of navigation. Secondly, a state value network based on double attention was designed to process environmental information to ensure that the robot navigation system has both environmental understanding ability and real-time decision-making ability. Finally, the algorithm was verified in the simulation environment. Experimental results show that the proposed algorithm not only improves the navigation efficiency, but also improves the robustness of the robot navigation system; The main performance is that in 500 random test scenarios, the collision times and timeout times are 0, the navigation success rate is better than the comparison algorithm, and the average navigation time is 2% shorter than the best algorithm; When the number of pedestrians and navigation distance in the environment change, the algorithm is still effective, and the navigation time is shorter than the comparison algorithm.

Key words: deep reinforcement learning; reward function; state value network; double attention

收稿日期:2022-03-21

基金项目:国家自然科学基金项目(62067002,61967006,62062033);江西省自然科学基金项目(20212BAB202008);江西省交通厅科技项目(2022X0040)

Citation format: XIONG L Y, SHU Y S, ZENG H, et al. Research on robot navigation algorithm based on deep reinforcement learning[J]. Journal of East China Jiaotong University, 2023, 40(1): 67–74.

研究对环境模型依赖程度低、能通过自主学习适应复杂环境的导航方法是移动机器人导航研究的必然趋势[1]。动态密集的人群是一种典型的动态避障导航场景[2],机器人通过感知实时变化的环境信息,选择合适的动作,最终安全无碰撞穿越运动人群,在保证安全的前提下尽快到达目标位置。运动的行人相比于道路中行驶的车辆,行为更加灵活与不可预测,因此理解并推理行人意图,对于移动机器人在人群环境中顺利导航至关重要[3]。

传统的导航算法主要针对环境基本可知且固定、机器人定位准确且运动方式简单的情况,利用经典的搜索算法或规划算法,计算出一条安全可靠的路径[4]。常用的有蚁群算法、A* 算法、人工势场法以及动态窗口法等方法[5-9]。以上传统的方法在复杂的环境中无法处理复杂的高维环境信息,容易陷入局部最优,并且在动态障碍物较多的场景中效果不佳[10]。

考虑到深度学习对环境的感知能力,以及强化学习优秀的决策能力,将深度学习与强化学习相结合提出的深度强化学习[11-13](deep reinforcement learning, DRL),能够实现移动机器人在复杂的环境中能够不依赖地图信息进行自主导航[14]。Chen 等[15]将运动的行人视为不合作的机器人, 融合多智能体导航[16]与 DRL,提出了一种多机器人在无通信场景下的无碰撞算法,实现多个机器人在同一个环境中到达各自目标位置不发生碰撞。后续的工作中[17-18] 增加了社交意识(socially aware)模块,并将该算法扩展至人群社交性导航场景中。

然而行人的运动具有一定的随机性,并不完全和机器人一样, 为了编码机器人与人群的交互过程,Chen 等[19]将注意力机制引入 DRL 提出了 SARL (socially attentive with deep reinforcement learning, SARL)算法,使得机器人的导航过程更符合人群的社交行为,后续的工作中[20]使用图卷积编码交互过程。Li 等[21]引入动态局部目标设定机制,使得 SARL 更适应于长距离导航。

为了更好理解人群运动,本文提出一种双重注意深度强化学习算法 (double attention deep reinforcement learning algorithm, DADRL)。

1 问题建模

本文将移动机器人导航环境简化为二维平面,如图 1 所示。环境中存在个向着各自终点运动的行人,各自半径为 $r_i(i=1,2,\cdots,n)$,机器人为平面中一个半径为的圆,运动方向由红色箭头标出。机器人通过传感器获取人群的实时状态,并且由运动策略控制执行离散的动作。

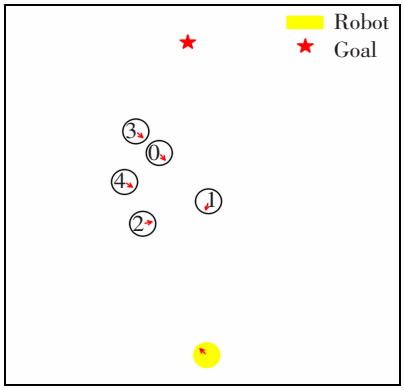

图 1 机器人导航环境

Fig.1 Robot navigation environment

在这样一个部分可观测的环境中,机器人根据实时获取的环境状态进行运动,是一个顺序决策过程。将人群导航问题建模为马尔科夫决策过程,用元组 $M\equiv<S,A,P,R,\gamma>$表示[22]。其中 S 为状态空间,A 为动作空间,P 为状态转移概率,R 为奖励函数,γ 为折扣因子。机器人通过与环境的交互学习控制策略,目标是得到策略函数,从而机器人可以根据接收的联合状态选择最佳的动作,安全无碰撞到达目标位置。

1.1 状态空间与动作空间

环境将联合状态 $S_t^{jn}=[S_t,O_t^i]$反馈给机器人,其中 S_t 表示 t 时刻机器人自身的状态信息,O_t^i 表示 t 时刻第 i 个人被机器人观测到的信息,S_t 和 O_t^i 都是状态空间 S 的子集。为了更好描述机器人的局部信息,对全局坐标系重建,以机器人所在的位置为原点,机器人与目标点的连线为 X 轴,建立以机器人为中心的坐标系。转换后的 S_t 和 O_t^i 为

$$\left.\begin{aligned} S_t&=[v_x,v_y,v_{\text{pref}},r,d_g]\\ O_t^i&=[p_{ix},p_{iy},v_{ix},v_{iy},r_i,r_i,+r,d_i]\\ S_t^{jn}&=[S_t,O_t^1,O_t^2,\cdots O_t^n]\end{aligned}\right\} \tag{1}$$

式中：v_x，v_y 为机器人的速度信息；v_{pref} 为首选速度；r 为机器人半径；d_g 表示机器人到目标位置的距离。p_{ix}，p_{iy}，v_{ix}，v_{iy} 分别为第 i 个人的位置信息和速度信息；r_i 为第 i 个人的半径大小；d_i 为第 i 个人与机器人的距离；S_t^{jn} 为整个环境的状态信息。

动作空间 $A=[v,\omega]$ 由 80 个离散的动作构成，其中 v 表示线速度，在区间 $[0,v_{\text{pref}}]$ 内以指数间隔取 5 个值；ω 表示角速度，在区间 $[0,2\pi]$ 内均匀取 16 个值。

1.2　奖励函数

奖励函数 R 的表达式为

$$R_t(S_t^{jn},a_t)=-\eta d_g+\begin{cases}2, d_g=0\\ -1, d_{\min}\leqslant 0\\ 0.3d_\Delta, 0<d_{\min}\leqslant d_c\\ 0, \text{其他}\end{cases}\tag{2}$$

式中：R_t 为联合状态为 S_t^{jn} 时，机器人采取动作 a_t 所得到的奖励值，机器人执行的操作 $a_t\in A$。

本文设计了一个大于 0 的距离惩罚系数 η，当机器人与终点的距离 d_g 越远则奖励值越小；d_g 为 0 则意味着机器人成功到达了终点，此时导航结束；$d_{\min}$ 表示人群中离机器人最近的单位与机器人的距离，小于 0 意味着发生了碰撞，此时导航终止；$d_\Delta=d_{\min}-d_c$，其中 d_c 是人为指定的舒适性距离，$0<d_{\min}\leqslant d_c$ 意味着虽然还没有碰撞，但人与机器人的距离太近产生了不舒适感，给一个负奖励。

奖励函数的设计，目的是使得机器人尽快向目标位置靠近，在这个过程中尽可能避免碰撞，同时在前进的过程中兼顾与人群的舒适性。

1.3　策略函数

策略函数 $\pi(S_t^{jn}):S_t^{jn}\rightarrow a_t$，表明在联合状态 S_t^{jn} 下采取的最优动作为 a_t。遍历动作空间 A，考虑当前状态下执行动作 a_t 的奖励值，以及执行动作后下一状态的价值，综合衡量后选取最佳动作，表达式为

$$\pi(S_t^{jn})=\arg\max_{a_t\in A}\left[R_t(S_t^{jn},a_t)+\gamma^{\Delta t v_{\text{pref}}}\sum_{s'\in S}P_{ss'}^{a}V\left(S_{t+\Delta t}^{jn}\right)\right]\tag{3}$$

式中：$P_{ss'}^{a}=P\left[s'=S_{t+\Delta t}^{jn}\mid s=S_t^{jn},a=a_t\right]$，表示在联合状态 S_t^{jn} 下机器人执行动作 a_t 后，下一个联合状态为 $S_{t+\Delta t}^{jn}$ 的概率，机器人执行两个相邻动作的时间间隔为 Δt。由于联合状态由 S_t 和 O_t^i 组成，机器人执行动作 S_t 后人群的状态 $O_{t+\Delta t}^{i}$ 是不确定的，$S_{t+\Delta t}^{jn}$ 也是不确定的，状态转移矩阵 $P_{ss'}^{a}$ 描述了这种不确定性。

1.4　状态价值函数的贝尔曼方程

本文将时间间隔 Δt 设置为 0.25 s，并假定机器人得到动作后能立即执行，将得到确定的机器人的下一状态 $S_{t+\Delta t}$。假定在 $[t,t+\Delta t]$ 内每个人继续按照之前的速度和方向进行运动，得到人群的下一状态 $O_{t+\Delta t}^{i}$ 作为预测值。通过这种简化，机器人采取某一动作后，能够得到确定的下一联合状态 $S_{t+\Delta t}^{jn}$，不再需要状态转移矩阵。状态价值函数的贝尔曼方程为

$$V(S_t^{jn})=R_t\left(S_t^{jn},\arg\max_{a_t\in A}\left[R_t(S_t^{jn},a_t)+\gamma^{\Delta t v_{\text{pref}}}V\left(S_{t+\Delta t}^{jn}\right)\right]\right)+\gamma^{v_{\text{pref}}}V\left(S_{t+\Delta t}^{jn}\right)\tag{4}$$

式中：V 为某一联合状态的长期累积奖励，其中折扣因子 $\gamma\in[0,1]$，越大表明越看重长远奖励。

2　状态价值网络的构建与训练

状态价值函数 $V(S_t^{jn})$ 表示从 t 时刻到导航终止这一过程中的累积总奖励，综合衡量了联合状态 S_t^{jn} 的长期价值，不仅涉及对环境的感知，更是支撑智能体进行决策的关键依据；因此能够准确衡量不同状态对应的长期累积价值显得尤为重要。

本文用神经网络拟合状态价值函数，提出了一种基于双重注意力的状态价值网络用以整合状态信息，提取机器人与人群的交互特征，计算输入状态的累积折损奖励。

状态价值网络的结构如图 2 所示，由 3 部分组成：状态预处理模块，特征融合模块，决策模块。下文将详细介绍这 3 个模块的具体结构和功能，以及状态价值网络的训练。

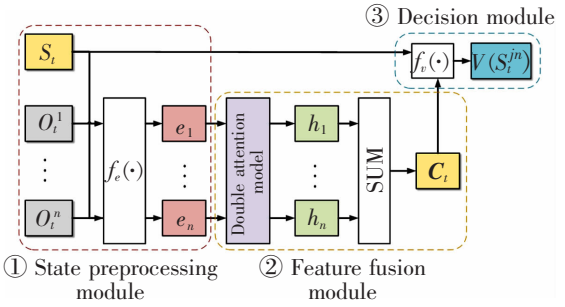

图 2　状态价值网络

Fig.2　State value network

2.1 状态预处理模块

状态预处理模块主要的功能是得到机器人与第 i 个人的融合特征 $\boldsymbol{e}_i$，表达式为

$$\boldsymbol{e}_i=f_e(S_t,O_t^i;w_e) \tag{5}$$

式中 S_t 为 t 时刻机器人自身的状态信息，包含有5个维度信息；O_t^i 为 t 时刻第 i 个人被机器人观测到的信息，包含有7个维度信息；$f_e(\cdot)$为一个多层感知机；w_e 为神经网络的权值，激活函数为ReLU。

融合后的向量 $\boldsymbol{e}_i$ 将机器人信息与第 i 个行人的信息进行整合，得到 n 个维度相同的向量，为后续的特征提取做准备。

2.2 特征融合模块

特征融合模块主要功能，是将两两交互特征 e_i 通过双重注意力模型，转化为机器人与人群的交互特征 C_t，其中双重注意力模型如图3所示。

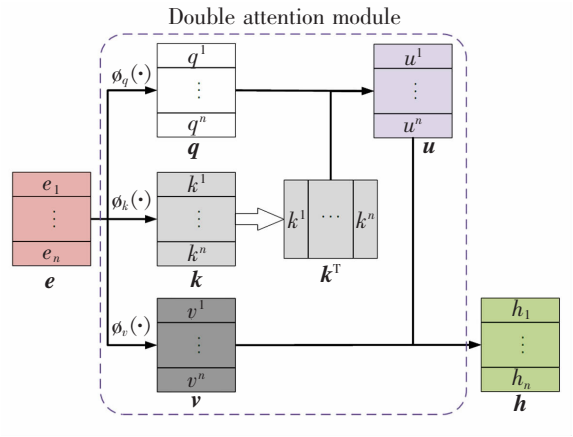

图3 双重注意力模型

Fig.3 Double attention model

预处理后的 n 个融合特征 $\boldsymbol{e}_i$ 拼接成矩阵输入双重注意力模型。矩阵 $\boldsymbol{e}$ 通过结构相同但权值不同的三个多层感知机，得到3个新的矩阵 $\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$，表达式为

$$\boldsymbol{q}^i=\phi_q(e_i;w_q);\ \boldsymbol{k}^i=\phi_k(e_i;w_k);\ \boldsymbol{v}^i=\phi_v(e_i;w_v) \tag{6}$$

式中：$\phi_q(\cdot)$，$\phi_k(\cdot)$，$\phi_v(\cdot)$为不同的多层感知机，w_q，w_k，w_v 为对应神经网络的权值，激活函数为ReLU。降维后的 $\boldsymbol{q}^i$、$\boldsymbol{k}^i$、$\boldsymbol{v}^i$ 依然包含了第 i 个人与机器人的融合特征。

注意力向量 $\boldsymbol{u}^i$ 的含义为第 i 个人对整个环境的注意力系数，具体表达式为

$$\boldsymbol{u}^i=\mathrm{softmax}(\boldsymbol{q}^i,\boldsymbol{k}^{\mathrm{T}}) \tag{7}$$

式中：向量 $\boldsymbol{q}^i$ 与矩阵 $\boldsymbol{k}$ 按行做内积运算，得到的向量经过softmax之后转化为 $\boldsymbol{u}^i$。在这个过程中第 i 个人综合考察了环境中每一个人的信息，其中也包括了机器人的信息，可视为第 i 个人对整个环境的感知，即第一重注意。

$\boldsymbol{u}^i$ 为第 i 个人从第一个维度对环境的考察，双重注意力向量 $\boldsymbol{h}^i$ 则是第 i 个人从第二个维度考察环境，即双重注意，具体的表达式为

$$\boldsymbol{h}_i=\boldsymbol{u}^i\boldsymbol{v} \tag{8}$$

式中：向量 $\boldsymbol{u}^i$ 与矩阵 $\boldsymbol{v}$ 按列做内积运算，得到第 i 个人与环境深度融合后的特征向量 $\boldsymbol{h}^i$。

机器人与人群的交互特征 $\boldsymbol{C}_t$，表达式为

$$\boldsymbol{C}_t=\sum_{i=1}^{n}\boldsymbol{h}_i \tag{9}$$

式中：人群中有 n 个行人，$\boldsymbol{h}_i$ 是两两交互特征 $\boldsymbol{e}_i$ 通过双重注意力模型得到的双重注意力向量。

2.3 决策模块

决策模块的功能是得到联合状态 S_t^{jn} 的累积折损奖励，即状态的长期价值，表达式为

$$V(S_t^{jn})=f_v(S_t,C_t;w_s) \tag{10}$$

式中：$f_v(\cdot)$为一个多层感知机；w_s 为对应的神经网络的权值，激活函数为ReLU。

决策模块将机器人自身的状态信息、机器人与人群的交互特征整合为一个具体的值，代表了该联合状态的价值大小，根据式(3)选择该状态下的最佳动作。

2.4 状态价值网络的训练

为加速模型的训练以及选择更合适的初始化参数，引入模仿学习，用最佳相互避免碰撞算法[23](optimal reciprocal collision avoidance，ORCA）驱动机器人在人群中运动，进行3 000回合(episode)的探索，生成轨迹数据并构造专家经验池，使用模仿学习对价值网络进行预训练。

状态价值网络的迭代更新使用时序差分法，训练过程则采用DQN[24]的双网络结构和经验回放池。探索过程采用 ε-greedy 策略，前5 000回合 ε 从0.5线性减小至0.1，后5 000回合 ε=0.1，导致探索终止的条件包括：机器人到达目标位置、机器人发生碰撞、导航时间超过上限，探索过程中将获得的信息同步存储于经验池中。机器人每走一步，从经验池随机采样一批经验更新价值网络。

3　实验结果与分析

3.1　仿真环境与实验参数

本文使用的仿真环境是 CrowdNav[19]。为了构造机器人穿越人群的场景，且行人的运动距离有效，将环境设置为一个圆形，行人的初始位置随机分布,其各自的终点与起点关于圆心对称,再对终点加上随机的扰动,机器人的起点和终点同样大致关于圆心对称。

为了评估本文提出的方法的有效性,设置机器人对人群不可见,即行人不会刻意避免不与机器人发生碰撞,机器人则需要理解人群的运动并做出合适的动作。模型的训练参数如表 1 所示。

表 1　实验参数

Tab.1　Experimental parameters

Parameter	Value
Distance penalty factor (η)	0.3
Comfort distance (d_c)	0.2
Learning rate (α)	0.001
Discount factor (γ)	0.3

3.2　定量分析

为了衡量算法的有效性,将本文提出的方法与 CADRL[15]、GA3C-CADRL[18]、SARL[19] 3 种成熟的导航算法进行对比,在 500 个随机的测试环境中进行验证,环境的半径为 4 m,行人数量为 5。

评价指标包括:导航成功率、碰撞次数、超时次数、平均导航时间、不舒适频率。导航成功率即 500 次测试中,机器人安全无碰撞到达目标位置的次数所占的比例,是最重要的指标;碰撞次数指的是导航过程中机器人与行人发生碰撞的次数，碰撞即意味着导航失败;超时次数指的是导航时间超过 25 s 但没有发生碰撞的次数,意味着机器人发生“冻结”[25];平均导航时间是成功进行导航的平均耗时;本文定义机器人与人的距离小于 0.2 m 时会让行人产生不舒适感,不舒适次数的占比即不舒适频率。对比实验的结果如表 2 所示。

在 4 种对比算法中,CADRL 的导航成功率最低，碰撞率达到了 4%,GA3C-CADRL 的导航时间最长,且成功率也仅优于 CADRL,在提升导航成功率的过程中牺牲了导航时间。这是由于这两种算法只考虑了单一的“机器人—行人”交互过程,对整个环境的理解具有局限性,这也证明了编码整个人群运动的必要性。

表 2　实验结果

Tab.2　Experimental result

Method	Navigation success rate	Number of collisions	Timeout times	Average navigation time	Discomfort frequency
CADRL[15]	0.950	21	4	11.37	0.064
GA3C-CADRL[18]	0.978	9	2	11.90	0.067
SARL[19]	0.998	1	0	10.73	0.007
DADRL (ours)	1.000	0	0	10.50	0.011

SARL 和本文提出的 DADRL 都对机器人与人群的交互进行编码,在 500 个测试例子中,SARL 的方法发生了一次碰撞,本文提出的方法则全部安全到达,且平均导航时间相比 SARL 缩短了 2 个百分点，代价是牺牲了很小的舒适性。与先进的 SARL 算法相比,DADRL 在奖励函数中增加了距离惩罚项,价值网络也进行了优化。在保证导航成功率的情况下,降低了导航时间,不舒适频率也相差不大,这证明了本文提出的方法的有效性。

3.3　鲁棒性分析

为了进一步衡量算法的鲁棒性,将训练好的算法应用于不同的环境中。

表 3 记录在环境半径保持 4 m 不变的情况下,环境中行人数量 p 发生变化时,不同方法的导航成功率。结果表明 SARL 和 DADRL 在行人数量变化时成功率保持在 99%以上,且效果优于对比算法。

表 3　人数变化时的导航成功率

Tab.3　Navigation success rate as pedestrian number changes

Method	p=3	p=4	p=5	p=6
CADRL[15]	0.99	0.98	0.95	0.93
GA3C-CADRL[18]	0.99	0.99	0.98	0.96
SARL[19]	0.99	1.00	1.00	0.99
DADRL(ours)	0.99	1.00	1.00	1.00

图 4 则表示在环境半径保持 4 m 不变的情况下,机器人平均导航时间与行人数量之间的变化关系。环境中行人越多意味着导航环境越复杂,相应的导航时间也越长。本文提出的 DADRL 在 4 种算

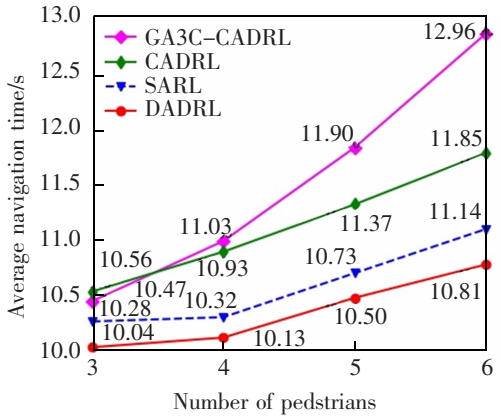

图4 行人数量与导航时间的关系

Fig.4 Relationship between pedestrian number and navigation time

法中始终保持导航时间最短,比最优秀的SARL还缩短了2%,表明本文的方法在整合人群运动特征方面优于对比算法,能更好处理行人数量变化的人群导航问题。

表4为环境中仅有5个行人的情况下,导航距离发生变化时不同方法的导航成功率,表明SARL和DADRL能够一定程度上适应导航距离的变化。

图5则表示环境中存在5个行人的情况下,机器人平均导航时间与环境半径之间的变化关系。随着环境半径增大即导航距离变长,导航时间大致呈线性增长的趋势,本文提出的DADRL相比于其他三种算法,始终具有更短的平均导航时间。当环境半径为3 m时,DADRL与SARL的导航时间差距达到最大为10%。实验结果表明,本文提出的方法能够更好应对导航距离变化的问题。

表4 导航距离变化时的导航成功率

Tab.3 Navigation success rate as navigation distance changes

Method	3 m	3.5 m	4 m	4.5 m	5 m
CADRL[15]	0.83	0.91	0.95	0.93	0.85
GA3C-CADRL[18]	0.94	0.96	0.98	0.98	0.98
SARL[19]	0.99	0.99	1.00	1.00	1.00
DADRL(ours)	0.99	1.00	1.00	1.00	0.99

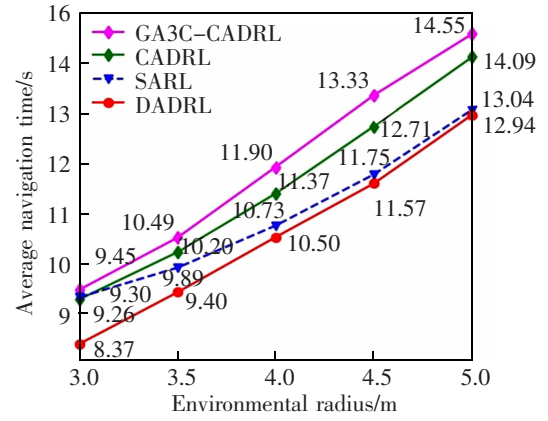

图5 导航距离与导航时间的关系

Fig.5 Relationship between navigation distance and navigation time

3.4 定性分析

为了更直观展示导航效果,本文对表2的对比实验进行定性分析,绘制四种算法控制下的机器人在人群中的导航轨迹。图中圆圈的大小代表行人与机器人的大小,实心圆代表的是机器人,空心圆则是运动的行人。圆圈中的数字代表某一时刻行人/机器人所处位置,例如:0.0标注起始位置,4.0标注了第4 s时5个行人与机器人分别所处位置,到达目标位置时再次记录各自的运动时长。将运动的行人编号1~5,每隔1 s记录一次所处位置,将位置连起来得到行人的运动轨迹,同时记录机器人的运动轨迹,其中5条虚线为行人的运动轨迹,实线为机器人的运动轨迹。

图6表示由CADRL驱动的机器人穿越人群的过程,可以观察到该算法较为“鲁莽”,在第4 s与5号行人交错而过时,并没有注意到不久将会与1号行人相遇,既没有加速前进也没有减速避让,导致后续的3 s时间内试图超过前进的1号行人,到第7 s才意识到可以减速让行,最终耗时12 s到达目的地。该运动过程表明机器人能够完成对单个行人的避障,但缺乏对人群运动的整体认识。

图7表示由GA3C-CADRL驱动的机器人穿越

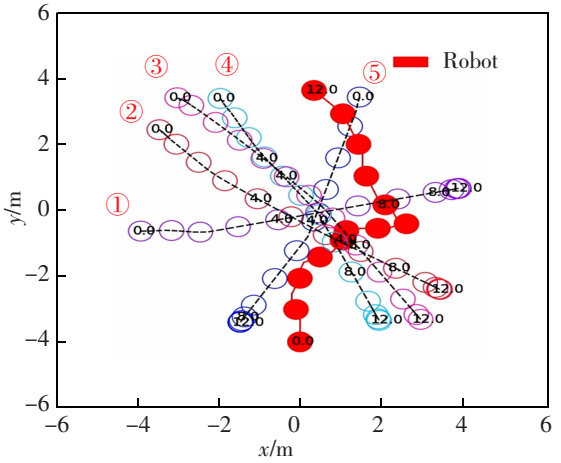

图6 CADRL运动轨迹

Fig.6 Movement trajectory of CADRL

人群的过程，在前 4 s 的时间内都在起点附近打转，直到行人的路程过半才出发，最终耗时 12.5 s 到达终点。

机器人的运动过程表明，CADRL 控制的机器人有些“鲁莽”，GA3C-CADRL 控制的机器人又有些“保守”，都使得导航时间被拉长，是没有充分理解环境的表现。

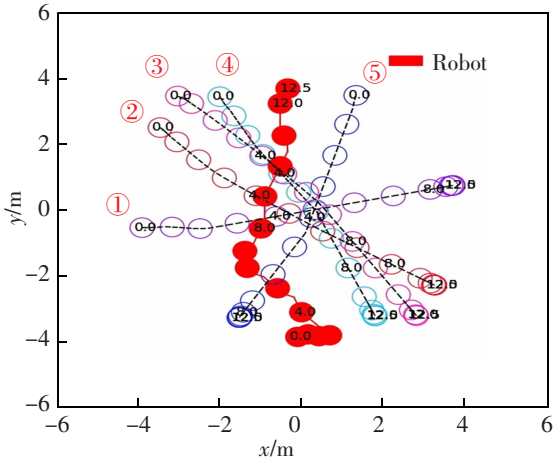

图 7　GA3C-CADRL 运动轨迹

Fig.7　Movement trajectory of GA3C-CADRL

图 8 表示由 SARL 驱动的机器人穿越人群的过程，该算法显然对环境有一定的认识，一开始就注意到行人都是向右侧运动的，因此很早就开始向左侧绕行，到第 6 s 时转头直奔目的地，最终耗时 9.8 s 到达终点。

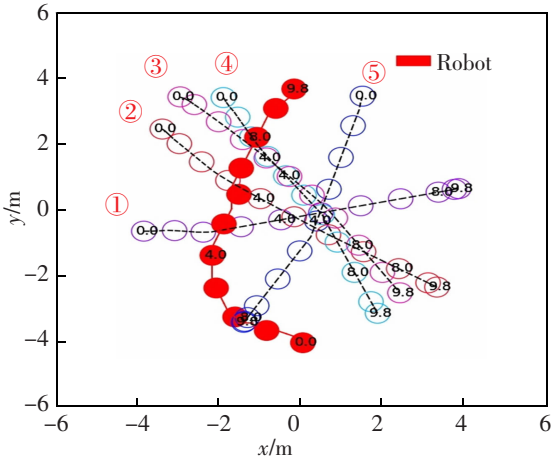

图 8　SARL 运动轨迹

Fig.8　Movement trajectory of SARL

图 9 描述了由 DADRL 算法控制的机器人穿越人群的过程。实验结果表明 DADRL 算法控制的机器人兼具对环境的理解能力和实时场景的应对能力，前 4 s 靠右侧快速前进从而避开密集人群，随后转向目标位置果断前进，耗时 8.2 s 到达终点，运动路径平滑且耗时最短，是 4 种策略中最优的方案。

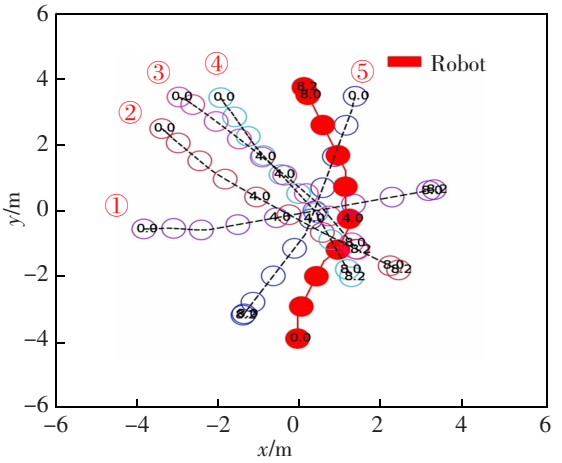

图 9　DADRL 运动轨迹

Fig.9　Movement trajectory of DADRL

4　结论

1）DADRL 具有比对比算法更高的导航效率，体现为导航成功率更高，导航时间更短，不舒适频率与最优算法相差不大；

2）在导航距离变长、环境中行人数量增长的情况下，DADRL 的导航效率优于对比算法；

3）通过分析导航轨迹，DADRL 的运动路径更加平滑，到达终点耗时更短。

参考文献：

[1] KONTOUDIS G P，VAMVOUDAKIS K G. Kinodynamic motion planning with continuous-time q-learning：an online，model-free，and safe navigation framework[J]. IEEE Transactions on Neural Networks and Learning Systems，2019，30(12)：3803-3817.

[2] 魏伟和. 动态密集人群环境下基于深度强化学习的移动机器人导航[D]. 哈尔滨：哈尔滨工业大学，2021.
WEI W H. Mobile robot navigation based on deep reinforcement learning in dynamic dense crowd environment [D]. Harbin：Harbin Institute of Technology，2021.

[3] SUN L，ZHAI J，QIN W. Crowd navigation in an unknown and dynamic environment based on deep reinforcement learning[J]. IEEE Access，2019，7：109544.

[4] 林韩熙，向丹，欧阳剑，等. 移动机器人路径规划算法的研究综述[J]. 计算机工程与应用，2021，57(18)：38-48.
LIN H X，XIANG D，OUYANG J，et al. Research review of path planning algorithms for mobile robots [J]. Computer En-

gineering and Applications,2021,57(18):38-48.

[5] 刘二根,谭茹涵,陈艺琳,等. 基于改进人工蚁群的智能巡线机器人路径规划[J]. 华东交通大学学报,2020,37(6):103-107.
LIU E G,TAN R H,CHEN Y L,et al. Path planning of intelligent line patrol robot based on improved artificial ant colony[J]. Journal of East China Jiaotong University,2020,37(6):103-107.

[6] HE Z ,LIU C,CHU X,et al. Dynamic anti-collision A-star algorithm for multi-ship encounter situations[J]. Applied Ocean Research,2022,118:102995.

[7] KATHIB O. Real-time obstacle avoidance for manipulators and mobile robots [J]. The International Journal of Robotics Research,1986,5(1):90-98.

[8] FOX D,BURGARD W,THRUN S. The dynamic window approach to collision avoidance[J]. IEEE Robotics & Automation Magazine,2002,4(1):23-33.

[9] 王洪斌,尹鹏衡,郑维,等. 基于改进的A* 算法与动态窗口法的移动机器人路径规划[J]. 机器人,2020,42(3):346-353.
WANG H B,YIN P H,ZHENG W,et al. Path planning of mobile robots based on improved a algorithm and dynamic window method[J]. Robotics,2020,42(3):346-353.

[10] 刘林韬. 基于深度强化学习的动态环境运动规划的研究[D]. 哈尔滨:哈尔滨工业大学,2021.
LIU L T. Research on dynamic environment motion planning based on deep reinforcement learning[D]. Harbin: Harbin Institute of Technology,2021.

[11] SILVER D,SCHRITTWIESER J,SIMONYAN K,et al. Mastering the game of go without human knowledge[J]. Nature,2017,550(7676):354-359.

[12] SILVER D,HUBERT T,SCHRITTWIESER J,et al. A general reinforcement learning algorithm that masters chess shogi and Go through self-play[J]. Science,2018,362(6419):1140-1144.

[13] VINYALS O,BABUSCHKIN I,CZARNECKI W M,et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature,2019,575(7782):350-354.

[14] ZHU K,ZHANG T. Deep reinforcement learning based mobile robot navigation:a review[J].Tsinghua Science and Technology,2021,26(5):674-691.

[15] CHEN Y F,LIU M,EVERETT M,et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]//Singapor:2017 IEEE international conference on robotics and automation (ICRA) IEEE,2017.

[16] 孙彧,曹雷,陈希亮,等. 多智能体深度强化学习研究综述[J]. 计算机工程与应用,2020,56(5):13-24.
SUN Y,CAO L,CHEN X L,et al. Overview of multi-agent deep reinforcement learning[J]. Computer Engineering and Applications,2020,56(5):13-24.

[17] CHEN Y F,EVERETT M,LIU M,et al. Socially aware motion planning with deep reinforcement learning[C]//Vancouver,British Columbia:2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) IEEE,2017.

[18] EVERETT M,CHEN Y F,HOW J P. Motion planning among dynamic,decision-making agents with deep reinforcement learning[C]//Madrid:2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE,2018.

[19] CHEN C,LIU Y,KREISS S,et al. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning[C]//Montreal:2019 International Conference on Robotics and Automation (ICRA) IEEE,2019.

[20] CHEN C,HU S,NIKDEL P,et al. Relational graph learning for crowd navigation[C]//Las Vegas:2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),2020.

[21] LI K,XU Y,WANG J,et al. SARL deep reinforcement learning based human-aware navigation for mobile robot in indoor environments[C]//Dali:2019 IEEE International Conference on Robotics and Biomimetics(ROBIO) IEEE,2019.

[22] LIU L, DUGAS D, CESARI G,et al. Robot navigation in crowded environments using deep reinforcement learning [C]//Las Vegas:2020 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE,2020.

[23] BERG J,GUY S J,LIN M,et al. Reciprocal n-body collision avoidance[J]. Robotics Research,2011,70:3-19.

[24] MNIH V,KAVUKCUOGLU K,SILVER D,et al. Human-level control through deep reinforcement learning[J]. Nature,2015,518(7540):529-533.

[25] FAN T,CHENG X,PAN J,et al. Getting robots unfrozen and unlost in dense pedestrian crowds[J]. IEEE Robotics and Automation Letters,2019,4(2):1178-1185.

第一作者:熊李艳(1968—),女,教授,硕士,硕士生导师,研究方向为交通大数据。E-mail:276477130@qq.com。

(责任编辑:姜红贵)