高技术通讯 2024 年 第 34 卷 第 7 期:734-743

doi:10.3772/j.issn.1002-0470.2024.07.007

受限密集环境下基于对比学习和强化学习的机器人导航方法①

禹鑫燚② 胡加南 郑万财 欧林林③

(浙江工业大学信息工程学院 杭州 310023)

摘 要 动态环境下的机器人导航是一个重要且具有挑战性的任务。针对机器人在受限密集环境下的导航任务,本文提出了一种基于深度强化学习(DRL)和对比学习结合的机器人导航方法。首先通过轨迹向量化方法来获取机器人和行人的历史信息,并设计了一个子图网络对其进行聚合,从而提高机器人对未来场景的预测能力。其次通过图神经网络(GNN)提取智能体(机器人、行人)之间的交互信息,赋予机器人预测行人意图的能力。最后在强化学习的基础上融入对比学习,并基于随机性策略强化学习算法性质提出了一种正样本增强方法,从而赋予机器人判断场景中其余位置安全性的能力以及找到更多可行路径的能力,提高其在复杂环境中的导航成功率。仿真实验验证了本文方法在受限密集环境中比现有的方法具有更好的性能。

关键词 深度强化学习(DRL);对比学习;机器人导航;人机交互

随着人工智能和机器人学的发展,机器人逐渐从一个孤立的工作空间走向与人类共享的工作空间,如商场中的导航机器人[1]、校园中的快递运输机器人等。尽管许多研究者针对机器人在动态环境下的导航任务提出了各种解决方法[2-3],但依然存在困难和挑战。一方面,人类的行为由于受到自身意图、外在环境和社会规则的影响而具有复杂性[4];另一方面,机器人和人类之间没有明确的通信方式。

早期在机器人导航领域的研究将障碍物建模成静态模型[5]或者简单动态模型[6-7],这与人类具有复杂的行为模式相互矛盾。传统的机器人导航方法分为基于反应的导航方法和基于轨迹的导航方法[8]。基于反应的方法使用几何或者物理法则规避碰撞,例如互利速度障碍算法[9]。这类方法可以根据当前场景中的状态快速做出反应,但是缺乏对未来的预测。基于轨迹的方法在路径规划之前进行一个长时间的轨迹预测从而解决了上述问题[10-12]。然而轨迹预测和路径规划的解耦可能会导致机器人出现冻结现象[13]。因为行人的轨迹预测可能会分布在整个空间中,从而导致机器人没有可行路径[14]。

为了同时解决上述 2 种方法存在的问题,机器人需要实时做出决策的同时考虑各个智能体(机器人和行人)之间的交互信息。针对交互建模的研究,文献[15-17]采用了手工建模的方式对交互进行建模。这类方法能够很好地模拟出当前场景下各个智能体之间交互信息,但是其移植性较差且非常耗时。文献[18,19]基于模仿学习从人类数据集中学习各个智能体的交互关系。这类方法所获得交互模型具有很好的通用性,但是其性能取决于数据集的质量好坏和规模大小。

除此之外,基于模仿学习的机器人导航方法直接从原始二维雷达数据[20]或者原始深度图像[21]中学习交互模型或者导航策略,尽管其能够直接处理场景中数量不确定的行人,但是受环境因素较大。

近年来,深度强化学习(deep reinforcement learn-

① 国家重点研发计划(2018YFB1308400)和浙江省自然科学基金(LY21F030018)资助项目。
② 男,1979 年生,博士;研究方向:机器人控制与规划;E-mail:yuxy@zjut.edu.com。
③ 通信作者,E-mail:linlinou@zjut.edu.com。
(收稿日期:2023-02-14)

ing,DRL)被成功用于很多不同的领域。在机器人导航领域,DRL 已经使机器人成功学习到包含各个智能体之间交互和合作的有效策略[22-23]。基于 DRL 的机器人导航方法可以直接从原始数据[24-25]或者从多个传感器中提取出的状态表征中[26-27]学习导航策略。尽管原始数据方法可以直接处理场景中数量不定的智能体,但是无法获得包含行人意图和交互更高级别的状态表征。

为了处理状态表征方法中数量不定的智能体,文献[28]引入了长短期记忆(long short-term memory,LSTM)递归神经网络模块并且将行人的状态按照和机器人距离的远近依次输入。然而这种输入方式并不适用于所有场景,比如机器人和行人相距很近但相反方向的运动。相比于 LSTM 只能获得单向交互信息,文献[29]提出的单智能体强化学习(single agent reinforeement learning,SARL)算法使用了局部地图和自注意力模块对行人进行编码和捕捉机器人、行人之间的双向交互。文献[30]引入动态局部目标设置机制和基于地图的安全行动空间改进了原有的 SARL 算法,并解决了局部地图范围限制问题。在此之后,文献[23]提出的关系图学习(relational graph learning,RGL)算法通过图神经网络(graph neural networks,GNN)提取所有智能体之间(包括行人和行人之间)的交互信息,并引入一个神经网络对行人运动模型进行建模。

然而上述方法中只有少部分考虑了和交互信息同样重要的历史信息。历史信息能够提高机器人对未来场景的预测能力从而提高导航任务的成功率。并且尽管这些方法在密集环境下也取得了良好的结果,但是其机器人并没有受到环境范围限制,即机器人在导航过程可以出现越界状态,这和实际场景存在一定差距。

针对机器人在受限密集环境中的导航问题,本文提出了一种基于深度强化学习和对比学习结合的机器人导航方法。为了获取历史信息,本文通过轨迹向量化方法并设计子图神经网络对其进行聚合,然后通过图神经网络对交互进行建模。除此之外,在深度强化学习的基础上融入对比学习。由于对比学习中正样本和负样本的存在,使机器人知道场景中哪些位置是可行点以及哪些位置是不可行点,从而降低了机器人导航任务中的碰撞率。本文基于随机性策略强化学习算法性质,提出了一个正样本增强方法,其目的是让机器人能够学到更多可行方案,从而在受限密集环境中能够找到一条成功路径。

1 问题描述

在一个二维平面中,机器人从起点导航到终点,同时需要避开 N 个互不通信的行人,如图 1 所示,该问题符合马尔可夫性质。第 i 个智能体在 t 时刻的可观状态包括位置 $\boldsymbol{P}_t^i = [p_{x,t}^i, p_{y,t}^i]$、速度 $\boldsymbol{V}_t^i = [v_{x,t}^i, v_{y,t}^i]$ 以及半径 r_i,不可观状态 $\boldsymbol{S}^{hid,i}$ 包括终点位置 $\boldsymbol{P}_g^i$ 和最大速度 $\boldsymbol{V}_{pref}^i$。$\boldsymbol{S}_{i,t}^{his}$ 表示第 i 个智能体以时间 t 为结尾的历史状态,其由可观状态构成。$\boldsymbol{S}_{H,t}^{his} = [\boldsymbol{S}_{1,t}^{his}, \boldsymbol{S}_{2,t}^{his}, \cdots, \boldsymbol{S}_{n,t}^{his}]$ 表示 n 个行人在 t 时刻的历史状态集合。$i = 0$ 代表机器人,为了便捷性,下文通常省略。

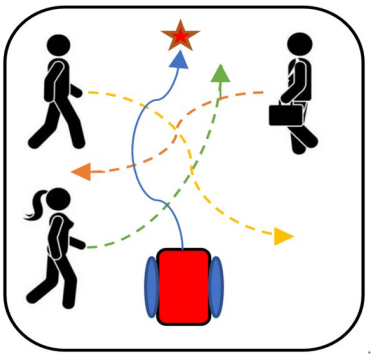


图 1 机器人导航示意图

整个决策过程为:在每个时刻 t,机器人首先观察到状态 $\boldsymbol{S}_t^{jn} = [\boldsymbol{S}_t^{his}, \boldsymbol{S}_{H,t}^{his}, \boldsymbol{S}^{hid}]$;然后根据该状态以及当前策略 π 计算出动作 $\boldsymbol{a}_t = [v_x, v_y]$ 并执行;最后从环境中得到奖励 $R(\boldsymbol{S}_t^{jn}, \boldsymbol{a}_t)$,状态变为下一时刻状态 $\boldsymbol{S}_{t+1}^{jn}$。

本文中强化学习算法采用软演员-评论家(soft actor-critic,SAC)算法[31]。强化学习奖励[29]设置如式(1)所示。

$$R(\boldsymbol{S}_t^{jn}, \boldsymbol{a}_t) = \begin{cases} 2 & 若\ \boldsymbol{P}_t = \boldsymbol{P}_g \\ -0.4 & 若\ d_{\min} < R\ 或出界 \\ d_{\min} - 0.2 & 若\ d_{\min} < 0.2 + R \\ i_t & 其他 \end{cases} \tag{1}$$

其中，$d_{\min}$ 表示机器人和所有行人之间的最短距离；$R = r + r^i$ 表示机器人和行人半径和；0.2 m 表示行人不舒适距离；i_t 是鼓励机器人朝终点移动的步进奖励，其定义如式(2)所示。

$$i_t = 1.6 \times (\| \boldsymbol{P}_t - \boldsymbol{P}_g \| - \| \boldsymbol{P}_{t+1} - \boldsymbol{P}_g \|) \tag{2}$$

2　方　法

本文所提方法的目标是提高机器人在受限密集环境下的导航性能。本方法包括 3 部分，即历史信息和交互信息的建模与聚合、对比学习和强化学习的结合以及正负样本的增强。本方法所采用的神经网络框架如图 2 所示，子图网络用于每个智能体历史信息的聚合，图神经网络用于智能体之间交互信息的提取，映射器和编码器用于将样本和正负样本映射到表征空间从而进行对比学习，软演员-评价网络包含了输出动作的策略网络以及用于网络更新的评价网络。

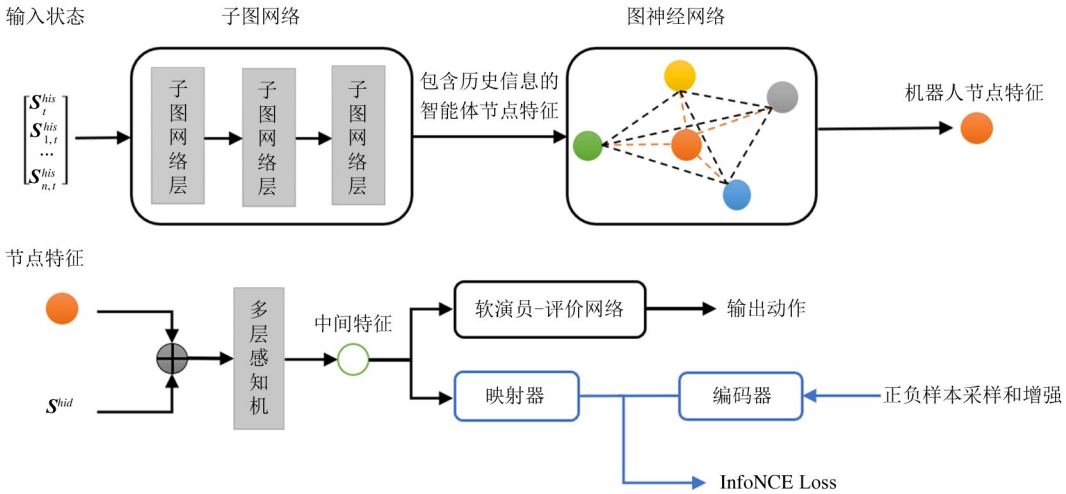


图 2　神经网络结构图

2.1　历史信息和交互信息的建模与聚合

在机器人导航任务中，智能体的轨迹是一条关于时间 t 的曲线。当以足够小且固定的时间 Δt 对轨迹采样时，连续的曲线可离散化成向量，如图 3 所示。对于真实场景中的静态障碍物，如道路边界、花坛等可以采用等距离采样的方式使其向量化。基于上述两点，本文提出了轨迹向量化方法用于提取障碍物的历史信息并设计了子图网络对其聚合。

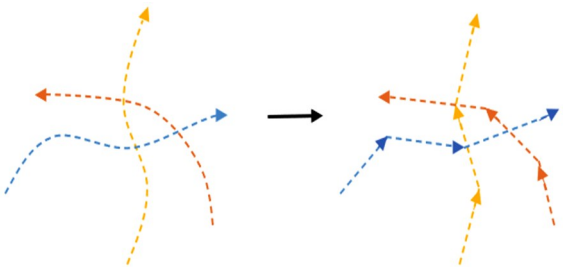


图 3　轨迹向量化

第 i 个智能体在 t 时刻的向量 $\boldsymbol{vec}_{i,t}$ 定义如下：

$$\boldsymbol{vec}_{i,t} = [\boldsymbol{P}^i_{t-1}, \boldsymbol{V}^i_t, r_i, \boldsymbol{P}^i_t] \tag{3}$$

对每个智能体 i，选取时间 t 为结尾的 k 个向量当作其历史状态 $\boldsymbol{S}^{his}_{i,t}$，其定义为

$$\boldsymbol{S}^{his}_{i,t} = [\boldsymbol{vec}_{i,t-k+1}, \boldsymbol{vec}_{i,t-k+2}, \cdots, \boldsymbol{vec}_{i,t}] \tag{4}$$

然后通过包含线性层、归一化层和 ReLU(rectified linear unit)激活函数的多层感知机(multilayer perceptron，MLP)对每个历史状态 $\boldsymbol{S}^{his}_{i,t}$ 进行映射转换：

$$\boldsymbol{A}^i_t = f_\alpha(\boldsymbol{S}^{his}_{i,t}; \theta_\alpha) \tag{5}$$

其中，θ_α 是所有智能体共享的神经网络权重；$f_\alpha(\cdot)$ 表示 MLP；$\boldsymbol{A}^i_t = [\boldsymbol{\alpha}^i_{t-k+1}, \boldsymbol{\alpha}^i_{t-k+2}, \cdots, \boldsymbol{\alpha}^i_t]$，其表示历史状态 $\boldsymbol{S}^{his}_{i,t}$ 的投影。

为了能够从历史信息推理出智能体的意图，首先将 $\boldsymbol{A}^i_t$ 聚合成 $\boldsymbol{b}^i_t$ 以提取历史相关性，然后复制 k 次得到和 $\boldsymbol{A}^i_t$ 相同维度的 $\boldsymbol{B}^i_t = [\boldsymbol{b}^i_t, \boldsymbol{b}^i_t, \cdots, \boldsymbol{b}^i_t]$，最后将 $\boldsymbol{A}^i_t$ 和 $\boldsymbol{B}^i_t$ 在特征维度进行拼接：

$$\boldsymbol{B}^i_t = \varphi(\boldsymbol{A}^t_i) \tag{6}$$

$$\boldsymbol{C}_t^i = \boldsymbol{A}_t^i \oplus \boldsymbol{B}_t^i \tag{7}$$

其中,$\varphi(\cdot)$ 表示复制操作和聚合操作;⊕表示拼接操作,聚合操作在实际算法中使用了最大池化技术(max pooling);$\boldsymbol{C}_t^i = [\boldsymbol{c}_{t-k+1}^i, \boldsymbol{c}_{t-k+2}^i, \cdots, \boldsymbol{c}_t^i]$ 表示特征集合,$\boldsymbol{c}_t^i$ 表示在时刻 t 下每个向量所对应的特征。为了获取智能体级别的特征,$\boldsymbol{C}_t^i$ 被聚合成 $\boldsymbol{z}_t^i$:

$$\boldsymbol{z}_t^i = \phi(\boldsymbol{C}_t^i) \tag{8}$$

其中,$\phi(\cdot)$ 表示聚合操作,式(3)~(8)的操作对应图 2 中的一层子图神经网络层。$\boldsymbol{z}_t^i$ 是每一层子图神经网络层的输出特征。

针对各个智能体之间交互,本文引入基于自注意力机制的图神经网络进行建模[23]。每个智能体的状态 $\boldsymbol{S}_{i,t}^{his}$ 经过子图神经网络得到 $\boldsymbol{z}_t^i$ 后同时输入到图神经网络。

从图神经网络提取出机器人节点特征 $\boldsymbol{h}_t$ 与机器人自身不可观状态先进行拼接,此时的机器人节点特征包含了各个智能体之间的交互信息和历史信息;然后再送入 MLP 中得到中间特征作为强化学习和对比学习的输入项:

$$\boldsymbol{h}_t = f_\beta(\boldsymbol{h}_t \oplus \boldsymbol{S}^{hid}; \theta_\beta) \tag{9}$$

其中,MLP 由线性层和 ReLU 激活函数构成,θ_β 表示神经网络权重。

2.2 对比学习和强化学习结合

在基于强化学习的机器人导航方法中,机器人通过利用和探索能够对当前策略进行更新从而学习到最优策略,但是利用和探索两者相互对立。在训练过程中,机器人若选择了利用原有的经验则会放弃探索,若选择探索则放弃利用原有的经验。因此,本文提出了一种强化学习和对比学习结合的学习方式,其目的是让机器人无论处于利用还是探索都能知道场景中其他位置的安全性,从而赋予机器人更强的导航能力。

首先在 2.1 节所提出方法的基础上,将中间特征 $\boldsymbol{h}_t$ 输入到映射器 $f_p(\cdot)$,完成样本到表征空间的映射:

$$\boldsymbol{q}_t = f_p(\boldsymbol{h}_t; \theta_p) \tag{10}$$

其中,映射器由线性层、ReLU 激活函数和归一化层构成,θ_p 表示神经网络权重。

为了对正负样本进行统一增强,本文选取当前 t 时刻及后续 $g-1$ 个时刻的正负样本作为 t 时刻的正负样本。然后将正负样本通过样本编码器 $f_e(\cdot)$ 映射到表征空间:

$$\boldsymbol{k}_{t+\delta t} = f_e(\boldsymbol{O}_{t+\delta t}, \delta t; \theta_e) \tag{11}$$

其中,样本编码器由线性层、ReLU 激活函数和归一化层构成,θ_e 表示神经网络权重,$\boldsymbol{O}_{t+\delta t}$ 表示 t 时刻下包含 g 个时刻的正负样本,后文直接用 t 时刻下的正负样本简述,$\delta t > 0$ 表示采样时刻和当前时刻的差值。

最后将样本和正负样本在表征空间所得到的 $\boldsymbol{q}_t$、$\boldsymbol{k}_{t+\delta t}$ 进行对比。基于 InfoNce Loss 的损失函数[32]定义如下:

$$L_{NCE,u} = -\log \frac{\exp(\boldsymbol{q}_t \cdot \boldsymbol{k}_{t+\delta t}^{u,+}/\tau)}{\sum_{n=0}^{W}\sum_{\delta t=0}^{g}\exp(\boldsymbol{q}_t \cdot \boldsymbol{k}_{t+\delta t}^{n}/\tau)} \tag{12}$$

其中,τ 表示温度参数,$\boldsymbol{k}_{t+\delta t}^{u,+}$ 表示 t 时刻下第 u 个正样本对应的特征,W 表示 t 时刻下所有正负样本数量。对 t 时刻下所有关于 u 的 $L_{NCE,u}$ 求平均得到对比学习更新函数 L_{cl}。

整个算法的训练目标分为:强化学习训练目标和对比学习训练目标 2 个部分,其损失函数 L_{all} 定义如下:

$$L_{all} = L_{rl} + \lambda L_{cl} \tag{13}$$

其中,λ 表示两者之间的权重,L_{rl} 表示 SAC 更新公式。

2.3 正负样本采样及增强

正样本和负样本的采样方式是对比学习能否成功应用于导航任务的关键因素之一。在成功使用对比学习的任务中,负样本通常是从整个数据集中随机采样得到[33]。然而这种通用的采样方式不适用于机器人导航任务,其原因有 2 点:(1)随机采样得到的负样本位置点可能是机器人未来轨迹中的一部分;(2)随机采样和导航路径与多模态性质不符。

为了利用机器人导航任务的先验知识,本文采用将行人位置及其附近位置作为负样本的方法[32],如图 4 所示,其定义如下:

$$\boldsymbol{O}_t^{i,-} = \boldsymbol{P}_t^i + \Delta\boldsymbol{S} \tag{14}$$

其中,$\Delta\boldsymbol{S} = \rho(\cos\theta, \sin\theta)$,表示以当前行人位置为原点的局部位移,$\rho = r + r^i$,$r^i$ 表示第 i 个行人,$\theta \in$

$\{0,0.25,0.5,0.75,1,1.25,1.5,1.75\}\cdot\pi$。

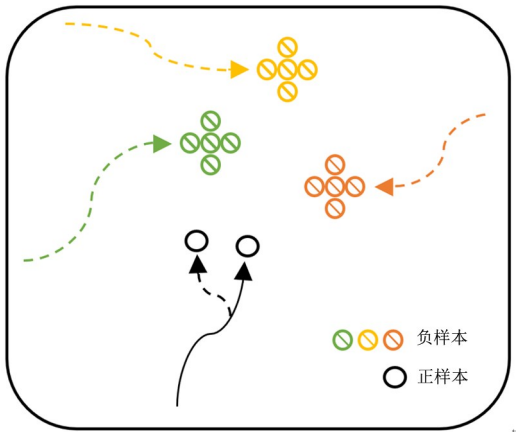


图 4　正负样本及增强

在强化学习和对比学习结合的方法中,正负样本等训练数据是从经验池中采样获得,其可能来自于导航失败的轨迹,这与正样本性质相矛盾,因此需要对其进行筛选。筛选过程如下:如果机器人当前时刻位置来自于成功轨迹则把其直接当作正样本;反之,则需要判断该位置和各个行人是否会发生碰撞,若没有碰撞则当作正样本。若在 t 时刻没有符合要求的正样本,则相应的负样本采样增强和正样本增强也同样无需执行。

由于导航任务中机器人路径具有多模态的性质,因此除了当前轨迹外场景中存在其余可行路径。此外,SAC 算法的输出动作是从一个策略分布中采样得到,因此具有随机性。基于上述两点原因,本文提出了一种正样本增强方式:首先根据策略神经网络输出的均值和标准差生成正态策略分布;接着从分布中采样得到 L 个速度样本;然后根据机器人上一时刻位置和采样得到的速度生成 L 个位置点;最后筛选出和所有行人最短距离最远且不发生碰撞的 $m(m<L)$ 个正样本并和机器人当前时刻的位置构成总正样本数,其定义如下:

$$\boldsymbol{O}_t^{+}=\{\boldsymbol{P}_t,f_s(mean,std,\boldsymbol{P}_{t-1})\} \tag{15}$$

其中,$f_s(\cdot)$ 表示采样筛选函数,$mean$ 表示均值,std 表示标准差。

为了增加正负样本数,在上述基础上对正负样本进行了统一增强。对于时刻 t,选取 t 到 $t+g-1$ 时间段内的所有正负样本,正负样本定义如式(16)所示。

$$\boldsymbol{O}_{t+\delta t}=(\boldsymbol{O}_{t+\delta t}^{+},\boldsymbol{O}_{t+\delta t}^{1,-},\cdots,\boldsymbol{O}_{t+\delta t}^{N,-}) \tag{16}$$

其中,t 时刻下正样本最大个数为 $(m+1)g$,负样本个数恒定为 $8gN$;若无正样本个数,则负样本个数则为 0。

3　仿真实验

3.1　仿真和参数设置

本文基于 Gym[29] 强化学习仿真平台设计了一个适用于机器人导航任务的仿真环境。在该环境中,设置一个边长为 8.0 m 的正方形场景,其中心为坐标原点,使用半径 r 为 0.3 m 的圆表示智能体。机器人起点位置设置为(0,-4),终点位置设置为(0,4),行人的起点和终点随机设置在正方形区域内。为了保证受限密集环境,行人数量设置为 10,并且机器人超出边界视为碰撞。为了能体现出机器人导航性能,将机器人在导航过程中设置为对行人不可见,行人无需主动避让机器人。为了提高训练效率,将机器人导航时间上限设置为 25 s,超过 25 s 则视为超时。在整个场景中,机器人由本文提出的方法控制,行人由最优相互避障算法(optima reciprocal collision avoidance,ORCA)[6] 控制,并且两者最大速度 $\boldsymbol{V}_{pref}$ 设置为 1 m·s^{-1}。实验参数如表 1 所示。

将本文方法和 ORCA[6]、RGL[23] 进行对比,并通过消融实验对各个模块(历史信息、对比学习和正样本增强)的有效性进行验证。训练得到的模型在仿真环境中的 500 个随机场景中进行测试。为了后续的便捷性,将本文所提出的方法定义为 HR-SACCLNav,其中 H 表示历史模块、R 表示正样本增强模块、CL 表示带负样本增强的对比学习模块。

3.2　定量分析

表 2 列出了 ORCA、RGL 和 HR-SACCLNav 这 3 种方法在测试环境中的成功率、碰撞率、超时率和平均导航时间。其中,ORCA 方法由于在不可见设置中违背了互利假设原则,成功率仅达到 62%。基于强化学习的 RGL 方法在受限密集环境下的成功率也只达到 76%。相比前两者而言,本文所提出的 HR-SACCLNav方法的成功率最高,能够达到87%。

表 1　实验参数设置

参数	值	参数	值
向量个数 k	3	采样个数 L	20
正样本增强个数 m	2	优化器	Adam
时间段长度 g	3	子图神经网络层数	3
权重 λ	1	子图神经网络尺寸	[7,64],[128,64],[128,64]
温度参数 τ	0.1	图神经网络层数	1
SAC 温度参数	0.05	图神经网络尺寸	[128,128]
学习率	0.000 3	多层感知机神经网络尺寸	[131,256,256]
SAC 折扣因子	0.99	SAC 评价神经网络尺寸	[256,256,1]
经验池大小	400 000	SAC 策略神经网络尺寸	[256,2]
批量大小	256	编码器神经网络尺寸	[1,64],[2,64],[128,64,16]
任务执行幕数	20 000	映射器神经网络尺寸	[256,128,16]

表 2　500 次测试对比实验结果

方法	成功率/%	碰撞率/%	超时率/%	导航时间/s
ORCA	62	16	22	13.9
RGL	76	18	6	**11.0**
HR-SACCLNav	**87**	**13**	**0**	12.6

除成功率以外,本文方法在碰撞率和超时率指标上都取得了最好的结果。由此可见本文方法在受限密集环境下拥有更好的导航性能,包括较高的成功率、低碰撞率和零超时率。

为了对历史模块、带负样本增强的对比学习模块和正样本增强模块进行评估,本文对 SACNav、H-SACNav、H-SACCLNav 和 HR-SACCLNav 进行了对比。表 3 列出了各个方法在测试环境中的成功率、碰撞率、超时率和平均导航时间。对比 SACNav,H-SACNav 的成功率提升了 2%,碰撞率降低了 1%,超时率降低了 1%。这说明历史信息的增加使得机器人具有了对未来场景的预测能力从而提高了其在复杂场景中的导航性能。对比 H-SACNav,H-SAC-CLNav 的成功率提高了 5%,碰撞率下降了 4%。这说明对比学习使得机器人能够识别出场景中的危险位置和安全位置从而有效减少碰撞。对比 H-SAC-CLNav,HR-SACCLNav 的成功率提高了 7%,碰撞率下降了 3%,超时率下降了 4%。这说明基于 SAC 随机性策略性质的正样本增强方法使得机器人能够找到更多的可行路径,不仅能降低机器人碰撞率也避免冻结现象的出现。

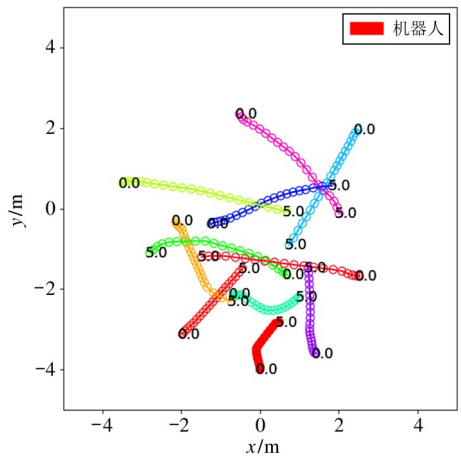
表 3　500 次测试消融实验结果

方法	成功率/%	碰撞率/%	超时率/%	导航时间/s
SACNav	73	21	6	13.4
H-SACNav	75	20	5	13.5
H-SACCLNav	80	16	4	**12.6**
HR-SACCLNav	**87**	**13**	**0**	**12.6**

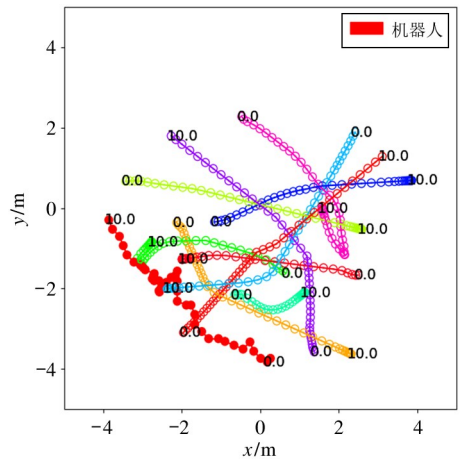
3.3　定性分析

图 5 描述了 3 种方法在同一个测试用例中的轨迹图,其中实心圆表示机器人,其余空心圆表示行人,数字表示各个智能体的导航时间。从图 5 中可以看出,ORCA 方法因为和行人发生碰撞而导致导航失败。RGL 方法尽管能够成功识别出行人从而

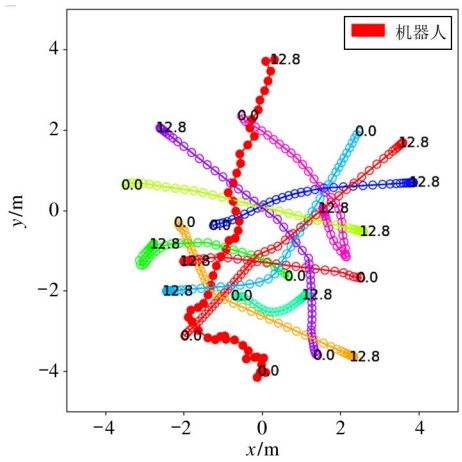
避免发生碰撞，但是由于区域中心行人较为密集其选择了绕行从而和环境边界发生了碰撞，同样也导致了导航失败。与前两者相反，本文方法能够成功地完成导航任务。



（a）ORCA



（b）RGL



（c）HR-SACCLNav

图 5　3 种方法对比轨迹图

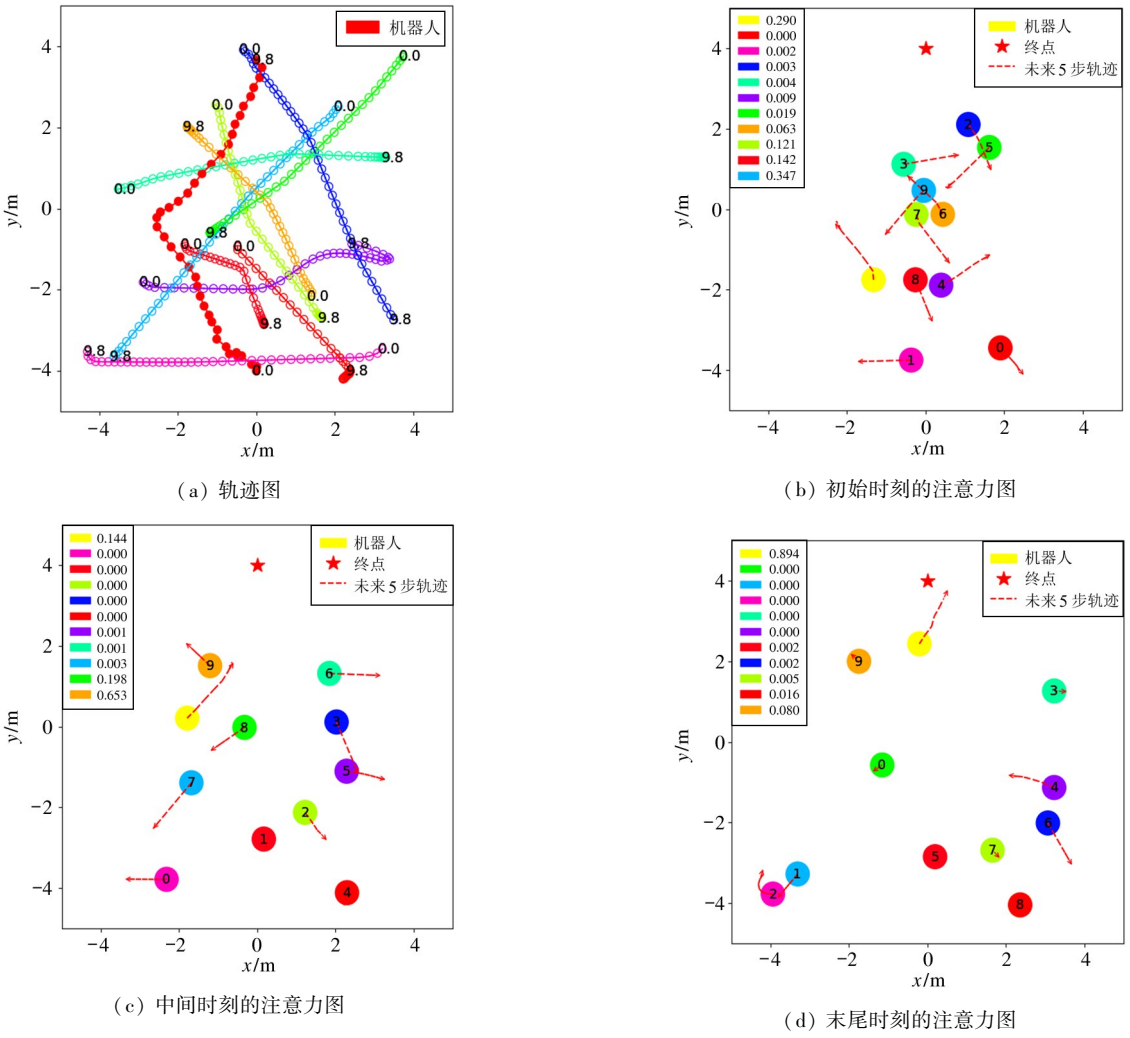
图 6 表示一个成功测试用例的轨迹图以及对应的注意力图。从图 6 中可以看出，机器人在这种受限密集的环境中不仅能够成功到达终点，还能够有效地避开行人。并且从导航时间上来看，机器人处于行人舒适范围之外且没有出现即将碰撞的现象。

图 6(b) ~ (d)描述了成功用例中的注意力，其左上角是机器人对各个智能体(包括自己)的注意力权重。从图 6(b)看出，机器人对最近的 8 号行人分配了第 2 大注意力权重，将最大注意力权重分配给了 9 号。因为从机器人当前朝向与 9 号行人未来的 5 步位置来看，两者很大可能发生碰撞，也因此在机器人未来 5 步轨迹中出现转弯现象。这证实了历史模块的加入使得机器人拥有了对未来场景的预测能力，从而提前进行转弯避免碰撞。

此外，在图 6(c)中，由于除 9 号和 8 号行人外其余行人未来轨迹和机器人轨迹都不会出现交叉，因此都分配了较低的注意力权重。而 9 号行人当前位置和未来 5 步位置都处于机器人未来轨迹的附近，因此分配了最大的注意力权重。在图 6(d)中，因为所有行人和机器人不可能发生碰撞，所以此时对于机器人来说首要任务是到达终点。因此，机器人分配给所有行人较低的注意力权重，而给了自己最高的注意力权重。

4　结　论

本文提出了一种基于强化学习和对比学习结合的导航方法，用于解决机器人在受限密集环境下的导航问题。首先通过轨迹向量化方法引入历史信息并设计子图神经网络对其进行聚合，提高了机器人对未来场景的预测能力，从而使其从时间维度上对各个行人分配更合理的注意力权重，提高机器人在复杂环境中的导航成功率，降低碰撞率和超时率。同时在强化学习的基础上，结合了对比学习并基于随机性策略强化学习算法性质提出了一个正样本增强方法，赋予了机器人判断场景中安全位置和找到更多可行路径的能力。最后，将本文方法和其他方法在受限密集环境下了进行了对比，并对各个模块(历史模块、带负样本增强的对比学习模块和正样



(a) 轨迹图　(b) 初始时刻的注意力图

(c) 中间时刻的注意力图　(d) 末尾时刻的注意力图

图6　轨迹图和注意力图

本增强模块)进行了消融实验。实验结果证明了各个模块的有效性,并且验证了本文方法在成功率、碰撞率和超时率方面都要优于其他方法。

参考文献

[1] KAYUKAWA S, HIGUCHI K, GUERREIRO J, et al. BBeep: a sonic collision avoidance system for blind travellers and nearby pedestrians[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, England: ACM, 2019:1-12.

[2] BRITO B, EVERETT M, HOW J P, et al. Where to go next: learning a subgoal recommendation policy for navigation in dynamic environments[J]. IEEE Robotics and Automation Letters, 2021,6(3):4616-4623.

[3] 路浩, 陈洋, 吴怀宇, 等. 受路网和测量约束的变电站巡检机器人路径规划[J]. 中国机械工程, 2021,32(16):1972-1982.

[4] RUDENKO A, PALMIERI L, HERMAN M, et al. Human motion trajectory prediction: a survey[J]. The International Journal of Robotics Research, 2020,39(8):895-935.

[5] FOX D, BURGARD W, THRUN S. The dynamic window approach to collision avoidance[J]. IEEE Robotics & Automation Magazine, 1997,4(1):23-33.

[6] VAN DEN BERG J, GUY S J, LIN M, et al. Reciprocal n-body collision avoidance[C]//The 14th International Symposium Robotics Research. Lucerne, Switzerland: ISRR, 2011:3-19.

[7] VAN DEN BERG J, SNAPE J, GUY S J, et al. Reciprocal collision avoidance with acceleration-velocity obstacles[C]//2011 IEEE International Conference on Robotics and Automation. Shanghai, China: IEEE, 2011:3475-

3482.

[8] EVERETT M, CHEN Y F, HOW J P. Motion planning among dynamic, decision-making agents with deep reinforcement learning[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018:3052-3059.

[9] VAN DEN BERG J, LIN M, MANOCHA D. Reciprocal velocity obstacles for real-time multi-agent navigation[C]//2008 IEEE International Conference on Robotics and Automation. Pasadena, USA: IEEE, 2008:1928-1935.

[10] AOUDE GS, LUDERS B D, JOSEPH J M, et al. Probabilistically safe motion planning to avoid dynamic obstacles with uncertain motion patterns[J]. Autonomous Robots, 2013,35(1):51-76.

[11] UNHELKAR V V, PÉREZ-D'ARPINO C, STIRLING L, et al. Human-robot co-navigation using anticipatory indicators of human walking motion[C]//2015 IEEE International Conference on Robotics and Automation (ICRA). Washington, USA: IEEE, 2015:6183-6190.

[12] KIM S, GUY S J, LIU W, et al. Brvo: predicting pedestrian trajectories using velocity-space reasoning[J]. The International Journal of Robotics Research, 2015, 34(2):201-217.

[13] TRAUTMAN P, KRAUSE A. Unfreezing the robot: navigation in dense, interacting crowds[C]//2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. Taibei, China: IEEE, 2010:797-803.

[14] CHEN Y, LIU C, SHI B E, et al. Robot navigation in crowds by graph convolutional networks with attention learned from human gaze[J]. IEEE Robotics and Automation Letters, 2020,5(2):2754-2761.

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

[16] FERRER G, GARRELL A, SANFELIU A. Robot companion: a social-force based approach with human awareness-navigation in crowded environments [C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE, 2013:1688-1694.

[17] MEHTA D, FERRER G, OLSON E. Autonomous navigation in dynamic social environments using multi-policy decision making[C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Korea: IEEE, 2016:1190-1197.

[18] ALAHI A, GOEL K, RAMANATHAN V, et al. Social LSTM: human trajectory prediction in crowded spaces [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nevada, USA: IEEE, 2016:961-971.

[19] GUPTA A, JOHNSON J, LI F F , et al. Social GAN: socially acceptable trajectories with generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:2255-2264.

[20] LONG P, LIU W, PAN J. Deep-learned collision avoidance policy for distributed multiagent navigation [J]. IEEE Robotics and Automation Letters, 2017,2(2):656-663.

[21] TAI L, ZHANG J, LIU M, et al. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE, 2018:1111-1117.

[22] EVERETT M, CHEN Y F, HOW J P. Collision avoidance in pedestrian-rich environments with deep reinforcement learning[J]. IEEE Access, 2021,9:10357-10377.

[23] CHEN C, HU S, NIKDEL P, et al. Relational graph learning for crowd navigation[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, USA: IEEE, 2020:10007-10013.

[24] TAI L, PAOLO G, LIU M. Virtual-to-real deep reinforcement learning: continuous control of mobile robots for mapless navigation[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017:31-36.

[25] LONG P, FAN T, LIAO X, et al. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). Prague, Czech Republic: IEEE, 2018:6252-6259.

[26] CHEN Y F, EVERETT M, LIU M, et al. Socially aware motion planning with deep reinforcement learning[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017:1343-1350.

[27] CHEN Y F, LIU M, EVERETT M, et al. Decentralized

non-communicating multiagent collision avoidance with deep reinforcement learning[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE, 2017:285-292.

[28] KATYAL K D, HAGER G D, HUANG C M. Intent-aware pedestrian prediction for adaptive crowd navigation [C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020: 3277-3283.

[29] CHEN C, LIU Y, KREISS S, et al. Crowd-robot interaction: crowd-aware robot navigation with attention-based deep reinforcement learning[C]//2019 International Conference on Robotics and Automation. Montreal, Canada: ICRA, 2019:6015-6022.

[30] LI K, XU Y, WANG J, et al. SARL*: deep reinforcement learning based human-aware navigation for mobile robot in indoor environments[C]//2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). Dali, China: IEEE, 2019:688-694.

[31] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning. Stockholm, Sweden: IEEE, 2018:1861-1870.

[32] LIU Y, YAN Q, ALAHI A. Social NCE: contrastive learning of socially-aware motion representations [C]// Proceedings of the IEEE/ CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 15118-15129.

[33] PAGLIARDINI M, GUPTA P, JAGGI M. Unsupervised learning of sentence embeddings using compositional n-gram features[EB/OL]. (2018-12-28)[2023-01-04]. http://arxiv.org/pdf/1703.02507.

Robot navigation method based on contrastive learning and reinforcement learning in restricted and dense environments

YU Xinyi, HU Jianan, ZHENG Wancai, OU Linlin

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

Abstract

Robot navigation in dynamic environment is an important but challenging task. For the robot navigation in restricted and dense environment, this paper proposes a robot navigation method based on the combination of deep reinforcement learning (DRL) and contrastive learning. Firstly, the trajectory vectorization is used to obtain the history information of robot and humans, and a subgraph network is designed to aggregate it, so that the ability of robot to predict future scenes is improved. Secondly, the interaction information between agents (robot and humans) is extracted by the graph neural network (GNN), which gives the robot the ability to predict the intention of humans. Finally, on the basis of reinforcement learning, contrastive learning is integrated, and a positive sample enhancement method is proposed based on the nature of stochastic policy reinforcement learning algorithm, so as to give the robot the ability to judge the security of other position in the scene and to find more feasible paths, improving navigation success rate in complex environment. Simulation results show that the proposed method has better performance than the existing method in restricted and dense environment.

Key words: deep reinforcement learning (DRL), contrastive learning, robot navigation, human-robot interaction