信息与控制　2022 年　第 51 卷　第 1 期：107 ~ 118

DOI：10.13976/j.cnki.xk.2022.0099　　文章编号：1002 - 0411(2022) - 01 - 0107 - 12

人群环境中基于深度强化学习的移动机器人避障算法

孙立香[1]，孙晓娴[2]，刘成菊[3]，靖　文[1]

1. 盐城工业职业技术学院智能制造学院，江苏 盐城　224005；2. 同济人工智能(苏州)研究院，江苏 苏州　215131；
3. 同济大学电子与信息工程学院，上海　201804

基金项目：国家重点研究开发计划(2016YFD0700905)；2020 年江苏省产学研合作项目(BY2020338)；2020 年江苏省大学生创新创业训练计划项目(202013752028Y)

通信作者：靖文，jwn967@163.com　收稿/录用/修回：2021 - 04 - 06/2021 - 05 - 31/2021 - 09 - 28

摘要

为了控制移动机器人在人群密集的复杂环境中高效友好地完成避障任务，本文提出了一种人群环境中基于深度强化学习的移动机器人避障算法。首先，针对深度强化学习算法中值函数网络学习能力不足的情况，基于行人交互(crowd interaction)对值函数网络做了改进，通过行人角度网格(angel pedestrian grid)对行人之间的交互信息进行提取，并通过注意力机制(attention mechanism)提取单个行人的时序特征，学习得到当前状态与历史轨迹状态的相对重要性以及对机器人避障策略的联合影响，为之后多层感知机的学习提供先验知识；其次，依据行人空间行为(human spatial behavior)设计强化学习的奖励函数，并对机器人角度变化过大的状态进行惩罚，实现了舒适避障的要求；最后，通过仿真实验验证了人群环境中基于深度强化学习的移动机器人避障算法在人群密集的复杂环境中的可行性与有效性。

关键词

深度强化学习
人机共融
行人空间行为
移动机器人避障

中图法分类号：TP242.6

文献标识码：A

Obstacle Avoidance Algorithm for Mobile Robot Based on Deep Reinforcement Learning in Crowd Environment

SUN Lixiang[1], SUN Xiaoxian[2], LIU Chengju[3], JING Wen[1]

1. *Institute of Intelligent Manufacturing, Yancheng Polytechnic College, Yancheng* 224005, *China*;
2. *Tongji Artifical Intelligence Research Institute, Suzhou* 215131, *China*;
3. *School of Electronics and Information Engineering, Tongji University, Shanghai* 201804, *China*

Abstract

To control mobile robots to efficiently perform obstacle avoidance in crowded and complex environments, a mobile robot obstacle avoidance algorithm based on deep reinforcement learning in the human-robot integration environment is proposed. First, in response to the lack of learning capability of the value network of deep reinforcement learning algorithms, the value function network is improved based on crowd interaction. The crowd information is exchanged through the angel pedestrian grid. The temporal characteristics of a single pedestrian are then extracted through an attention mechanism, which learns the relative importance of historical trajectory state and joint impact on the obstacle avoidance strategy of the robot, providing a first step for the subsequent learning of the multilayer perceptron. Next, a reward function was developed for reinforcement learning

Keywords

deep reinforcement learning;
human-robot integration;
human spatial behavior;
obstacle avoidance for mobile robot

based on human spatial behavior. The state where the robot angle changes significantly is punished to achieve the requirements of comfortable obstacle avoidance. Finally, the feasibility and effectiveness of the proposed algorithm in crowded and complex environments are verified through simulation experiments.

0 引言

移动机器人的工作场景逐渐从静态环境转向人群密集的复杂动态环境，例如机场、商场等。传统的避障算法利用当前时刻周围障碍物的信息，依靠几何构型[1-3]或采样[4-6]等方法得到最终的避障策略，其未考虑周围行人未来状态的变化，在行人密集的环境中，由于行人运动的随机性，该算法容易产生震荡等不自然的行为[7]，使得机器人的控制信号频繁跳动，不能满足人群密集环境中安全性与舒适性的要求。因此，为了在人群密集的复杂环境中高效地完成避障任务，移动机器人的避障算法需要对行人的行为进行分析。同时，随着人机交互理念的发展[8]，机器人的避障任务不仅仅需要满足高效与实时性，更加需要在与人交互的过程中考虑行人的舒适度要求[9-13]。

针对行人行为，可以通过隐马尔科夫模型(hidden Markov model, HMM)、高斯混合模型(Gaussian mixture model, GMM)[14-15]等数学模型对行人的历史轨迹进行拟合，进而分析其运动轨迹，但是该类方法适应性差，未考虑物理环境约束、周围智能体的影响以及社会规范等智能体行进时的具体因素，不能很好地拟合以及预测行人轨迹。因而许多学者通过研究行人运动行为及其特征，建立行人流模型[16-18]来对行人行为做更好的分析。其中，Helbing 与 Molnar 提出的社会力模型(social force model, SFM)[17]是传统行人运动模型中应用最为广泛的模型之一，该模型将行人与周围障碍物以及自身目标点之间的关系用力的形式进行描述，假定行人之间的潜在斥力是关于两者距离以及行人占有半径的单调递减函数，具有一定的实际意义。缺点是该类方法较为粗糙，且需要根据环境进行参数调整，存在一定的局限性。随着时序网络的发展，基于大规模行人运动数据集训练得到的时序网络模型可以更加精确地预测行人运动的轨迹[19-21]。其中，Social-LSTM(Social Long Short Term Memory)[19]与 Social-GAN(Social Generative Adversarial Network)[20]首先对周围环境进行栅格化处理，得到行人局部地图信息，利用网络对行人之间的潜在规则进行学习，预测得到了精确的轨迹信息，但栅格化的离散处理损失了一定的环境信息，且由于预测模型自身计算量大以及预测误差的存在导致的重规划问题，基于行人轨迹预测的避障算法难以满足实时性的要求。除此之外，基于轨迹预测的避障算法在人群密集的情况下容易导致机器人的冻结现象，从而找不到机器人的可行路径。

Google 在 2015 年提出了深度强化学习算法[22]，将深度学习的感知能力和强化学习的决策能力相结合，可以更好地解决复杂系统的感知决策问题，因此许多学者将其应用在机器人避障策略的学习中[23-26]。在人群密集环境中，基于深度强化学习的避障算法将行人行为预测与机器人运动规划进行了整合，直接通过环境状态得到机器人的控制输出信号，保证了避障算法的实时性与可靠性，成为了新的研究热点。Everett 等[27-29]使用多层感知机拟合值函数，通过智能体与环境的不断交互，值函数网络可以在一定程度上学习到行人运动的不确定性，提升了机器人在行人密集环境中的避障效率。但由于行人行为具有较强的随机性以及值函数网络学习能力有限，基于深度强化学习算法的避障算法依旧存在不自然的避障行为。

舒适度要求主要考虑到机器人需要与人保持一定的距离，该距离不仅可以避免碰撞，而且还可以防止行人由于与机器人距离太近而引发的情绪不适[30-33]。其中，Hall[30]提出的个人空间理论(human spatial behavior)被广泛应用于人机交互中，该理论将人与人之间的距离分为亲密距离(intimate distance)、私人距离(personal distance)、社会距离(social distance)与公共距离(public distance)四种，为了满足人机交互舒适度的要求，机器人应该避免出现在亲密或者私人距离内。Pandey[31]依据行人舒适度等社会要求对路径规划算法进行了修改，虽然使避障策略具备了舒适性，但是引入了大量的参数以及复杂的场景设计，在环境发生变化后需要调整参数，并有可能针对新的环境需要重新设计避障算法，因而适应性较差。

为了更加准确地对行人行为进行学习，并且在不引入额外计算参数的情况下使避障算法满足舒适

度的要求，本文提出了人群环境中基于深度强化学习的移动机器人避障算法。该算法受行人轨迹预测方法的启发，通过分析环境中行人的行为，提高了深度强化学习算法中值函数的学习能力；通过奖励函数对智能体进行合适的引导，在不进行精确建模的情况下，便可以得到符合人机交互要求的避障策略，达到机器人避障效率与行人舒适度的平衡。本文主要的创新点在于：

1）使用角度行人网格对行人周围动态障碍物进行编码，并利用注意力机制对行人的时序特征进行了提取，得到了更加丰富的行人动态信息，提高了值函数网络的学习能力。

2）依据行人空间行为，通过对进入行人私人距离内以及角速度变化过大的状态进行惩罚，重新设计了强化学习的奖励函数，使得机器人在避障的过程中可以实现行人舒适度要求与避障效率的平衡，从而更好地与人进行交互。

1　基于深度强化学习的避障算法

利用强化学习算法解决避障问题的目标为得到机器人联合状态 $\boldsymbol{s}_t^{\mathrm{jn}}$ 到控制输出 $\boldsymbol{a}_t$ 的最优策略 $\boldsymbol{\pi}^*$[24]，如图 1 所示，在基于强化学习的避障算法中，智能体通过与环境的交互得到当前时刻机器人以及周围所有行人的状态 $\boldsymbol{s}_t^{\mathrm{jn}}$；然后将机器人的控制信号离散化为一定大小的动作空间，通过一步预测模型（one-step look ahead）对动作空间中每个控制信号执行之后的状态进行预测，得到预测后的状态 $\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}}$；之后将预测状态 $\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}}$ 输入到值函数网络中，结合奖励函数得到该状态的价值，选取最大价值所对应动作空间中的动作即最优动作（optimal action）作为机器人最终的输出动作 $\boldsymbol{a}_t$，如式（1）所示。在智能体与环境的交互过程中，值函数网络 V^* 不断进行优化直至收敛。

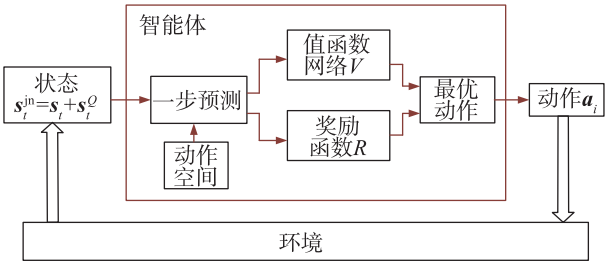


图 1　基于深度强化学习的避障算法

Fig.1　Obstacle avoidance algorithm based on deep reinforcement learning

$$\boldsymbol{a}_t=\underset{\boldsymbol{a}_t\in A}{\operatorname{argmax}}R(\boldsymbol{s}_t^{\mathrm{jn}},\boldsymbol{a}_t)+\gamma^{\Delta t\cdot v_{\mathrm{pref}}}V^*(\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}})\quad(1)$$

其中，$\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}}=\mathrm{prop}(\boldsymbol{s}_t^{\mathrm{jn}},\Delta t,\boldsymbol{a}_t)$，表示预测模型通过简单的线性模型对 Δt 时刻后机器人及其周围行人的运动进行预测，近似得到 Δt 时刻后机器人的联合状态。以下为强化学习各个元素的具体选择方法：

状态空间：考虑机器人自身的状态和环境中其他行人的状态都将对最终的决策造成影响，使用联合状态 $\boldsymbol{s}_t^{\mathrm{jn}}=[\boldsymbol{s}_t,\boldsymbol{s}_t^{o}]$ 作为 t 时刻强化学习的输入状态。其中，$\boldsymbol{s}_t=[p_{\mathrm{x}},p_{\mathrm{y}},v_{\mathrm{x}},v_{\mathrm{y}},r,g_{\mathrm{x}},g_{\mathrm{y}},v_{\mathrm{pref}},\varphi]$ 为机器人的状态，具体包括机器人的位置、运动速度、大小、目标点、最佳运行速度与转向角度；$\boldsymbol{s}_t^{o}=[\boldsymbol{s}_t^{1\mathrm{o}},\boldsymbol{s}_t^{2\mathrm{o}},\cdots,\boldsymbol{s}_t^{no}]$ 为环境中所有行人的状态，不同于 $\boldsymbol{s}_t$，由于在人群环境中，机器人无法获取行人的目标点，最佳运行速度与转向角度的信息，所以第 i 个行人的状态 $\boldsymbol{s}_t^{io}$ 仅包括行人的位置、运动速度、大小以及周围行人的编码信息，即 $\boldsymbol{s}_t^{io}=[p_{\mathrm{x}},p_{\mathrm{y}},v_{\mathrm{x}},v_{\mathrm{y}},r,r_i]$。

动作空间：动作空间取为一系列可行的机器人速度控制信号，即 t 时刻机器人的动作 $\boldsymbol{a}_t=[v,w]$，其中，v，ω 分别代表机器人的线速度与角速度。

值函数：强化学习算法的目标是获取最优值函数 V^*，且最优策略 $\boldsymbol{\pi}^*$ 可以通过值函数网络获取，使机器人在与环境的交互中得到最佳的期望奖励：

$$V^*(\boldsymbol{s}_t^{\mathrm{jn}})=\sum_{t'=t}^{T}\gamma^{t'\cdot v_{\mathrm{pref}}}R_{t'}(\boldsymbol{s}_{t'}^{\mathrm{jn}},\boldsymbol{\pi}^*(\boldsymbol{s}_{t'}^{\mathrm{jn}}))$$

$$\boldsymbol{\pi}^*(\boldsymbol{s}_t^{\mathrm{jn}})=\underset{a_t}{\operatorname{argmax}}R(\boldsymbol{s}_t^{\mathrm{jn}},\boldsymbol{a}_t)+\gamma^{\Delta t\cdot v_{\mathrm{pref}}}\int_{\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}}}P(\boldsymbol{s}_t^{\mathrm{jn}},\boldsymbol{a}_t,\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}})V^*(\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}})\mathrm{d}\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}}\quad(2)$$

其中，$R(\boldsymbol{s}_t^{\mathrm{jn}},\boldsymbol{a}_t)$ 是 t 时刻的奖励值；$\gamma\in(0,1)$ 是衰减系数；$P(\boldsymbol{s}_t^{\mathrm{jn}},\boldsymbol{a}_t,\boldsymbol{s}_{t+\Delta t}^{\mathrm{jn}})$ 为状态转移概率；v_{pref} 表示机器人在无障碍物时运行的速度，在之后的衰减系数中作为归一化项。

值函数通过机器人与环境的不断交互可以学习到一定的环境信息，但是仅仅使用浅层的多层感知机不能很好地对复杂的行人交互规则进行学习，本文对值函数进行修改，以提高值函数在人群密集的复杂环境下的学习能力，并针对行人舒适度的要求对奖励函数做了进一步的修改。除此之外，本文针对行人数量可变的情况，使用一个 LSTM 网络[25]进行处理。

2　具体算法

本节通过在值函数网络中引入行人交互信息以

及修改强化学习的奖励函数，基于深度强化学习得到了符合人机交互要求的避障算法。

2.1 基于行人交互的值函数网络改进

如图 2 所示，改进的值函数网络 μ 由行人交互信息模块（crowd interaction module）、LSTM 网络 $\boldsymbol{\Phi}_{\mathrm{LSTM}}(\cdot)$、多层感知机（MLP）三个部分组成：行人交互模块先对原始智能体状态 $\boldsymbol{s}_t^{\mathrm{jn}}$ 进行行人特征的提取；再通过 LSTM 网络 $\boldsymbol{\Phi}_{\mathrm{LSTM}}(\cdot)$ 进行不定数量行人特征的组合，得到所有行人的联合隐藏状态 $\boldsymbol{h}^{\mathrm{o}}$；最后将 $\boldsymbol{h}^{\mathrm{o}}$ 与机器人自身状态联合输入到多层感知机网络 $\psi_{\mathrm{M}}(\cdot)$ 中，得到对应的价值。本节主要介绍改进的值函数网络中的行人交互行为分析的具体方法。

行人之间的决策是互相影响的，如果将每对行人之间的关系进行准确的描述，将导致 $O(N^2)$ 的时间复杂度，随着环境中行人数量的增加，计算负担将会变得非常繁重。除此之外，一些距离较远的行人之间几乎不会对彼此的运动产生影响，将两者之间的关系进行描述不仅会加大计算复杂度，也会加重网络的学习负担。据此，本文采取一个特殊的混合网格—角度行人网格（APG）[32] 对行人的局部环境进行编码，以剔除对避障策略无用的信息。

为了更好地捕获行人的动态特征，角度行人网格（APG）引入一个高分辨率的网格，增加了信息的广泛性。图 3 表示了对第 i 个行人 P_{H}^i 进行编码的过程。设行人的总数为 N，角度行人网格（APG）仅考虑以行人 P_{H}^i 为中心、以 $r_{\max}$ 为半径的圆内的行人。将该圆均分为 K 个扇形区域，选择每一个扇形区域内与行人 P_{H}^i 最近的距离作为该扇形区域的编码值。行人 P_{H}^i 经过角度行人网格（APG）编码后输出为一维向量 $\boldsymbol{r}_i=[r_{i,1},\cdots,r_{i,k},\cdots,r_{i,K}]$，其中 $r_{i,k}$（第 k 个扇形的编码值）的数学表示方法如式（3）所示：

$$\begin{aligned} &r_{i,k}:=\min(r_{\max},\min(\{\rho_{i,j}\mid j\in\mathcal{N}_{i,k}\}))\\ &\mathcal{N}_{i,k}:=\{j\in\{1,\cdots,N\}\setminus\{i\}\mid\varphi_{i,j}\in[\gamma_k,\gamma_{k+1}]\} \end{aligned} \tag{3}$$

其中，$k\in\{1,\cdots,K\}$，每个网格所对应的扇形角度 $\gamma_k:=(k-1)\dfrac{2k}{K}$，$(\rho_{i,j},\varphi_{i,j})$ 代表了行人 P_{H}^j 在以行人 P_{H}^i 为中心的极坐标系下的坐标。因此，$\boldsymbol{r}_i$ 仅仅与行人位置有关。

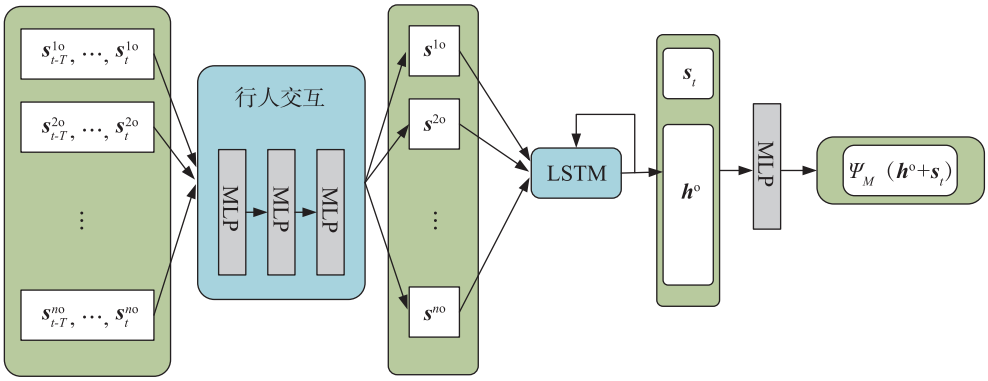


图 2 改进的值函数网络结构

Fig.2 Improved network structure of value function

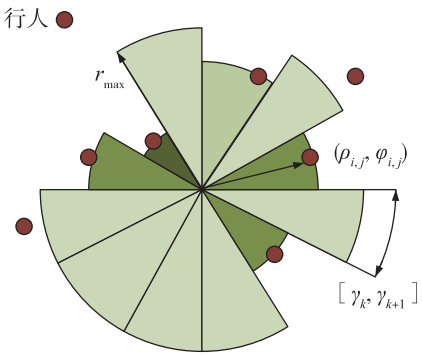


图 3 角度行人网格

Fig.3 Angular pedestrian grid

与标准 2D 网格相比，APG 仅仅线性地影响输入的维数，而仍然能够以连续的分辨率而不是离散的网格单元捕获径向距离的变化。此外，行人 P_{H}^j 距离行人 P_{H}^i 的距离越近，行人 P_{H}^j 角度位置变化的观察性变得越精确。对于给定参数的 APG 网格，可以得到行人 P_{H}^i 在 t 时刻的状态 $\boldsymbol{s}_t^{oi}=[p_{\mathrm{x}}^i,p_{\mathrm{y}}^i,v_{\mathrm{x}}^i,v_{\mathrm{y}}^i,r,r_{\mathrm{i}}]$，该状态变量将作为改进的值函数网络的输入。

机器人的运动不仅仅受到当前时刻行人信息的影响，与之前时刻行人的信息也有很大的关系，而且依据多个时刻的共同影响，可以克服避障策略的短时性（short-sighted），减少了由于仅仅依据当前时刻的即时环境信息带来的控制信号跳变的问题，可

以利用历史数据进行避障，达到更好的避障效果。

在时序预测中，注意力机制取得了较好的应用效果。本文通过自注意力机制(self-attention)[33]提取单个行人的时序特征，学习得到包括当前时刻在内的 T 个时刻数据的相对重要性以及对机器人避障策略的联合影响。

T 时刻内行人 P_{H}^{i} 的连续轨迹序列[$\boldsymbol{s}_{t-T}^{oi}$，…，$\boldsymbol{s}_{t-k}^{oi}$…，$\boldsymbol{s}_{t}^{oi}$]已知。如图 4 所示，首先对每一个时刻的行人特征进行编码，通过多层感知机 $\psi_e(\cdot)$(MLP)得到定长的输出 $\boldsymbol{e}_m$：

$$\boldsymbol{e}_m=\psi_{\mathrm{e}}(\boldsymbol{s}_{t-m}^{io}\mid\boldsymbol{W}_{\mathrm{e}}) \tag{4}$$

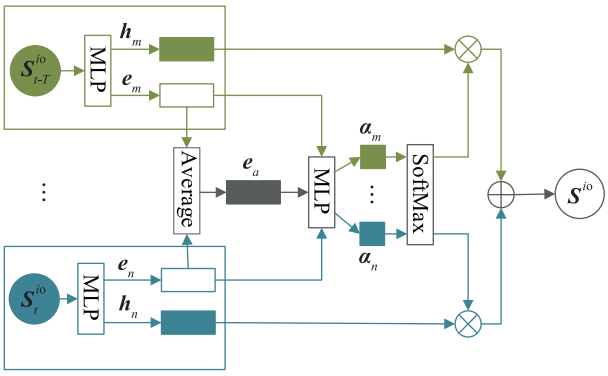


图 4　注意力机制模型

Fig.4　Attention mechanism model

然后，将向量 $\boldsymbol{e}_m$ 输入感知机 $\psi_h(\cdot)$中，从中提取得到行人 P_{H}^{i} 不同时刻状态之间的交互特性向量 $\boldsymbol{h}_m$：

$$\boldsymbol{e}_k=\psi_h(\boldsymbol{e}_m\mid\boldsymbol{W}_h) \tag{5}$$

接着通过多层感知机 $\psi_\alpha(\cdot)$计算不同时刻行人状态的相对重要性，得到每个交互特征向量的注意力得分 α_m。

$$\boldsymbol{e}_{\mathrm{a}}=\frac{1}{T}\sum_{k=1}^{T}\boldsymbol{e}_k \tag{6}$$

$$\boldsymbol{a}_m=\psi_\alpha(\boldsymbol{e}_m,\ \boldsymbol{e}_{\mathrm{a}}\mid\boldsymbol{W}_\alpha) \tag{7}$$

$\boldsymbol{e}_{\mathrm{a}}$ 为所有时刻的平均池化向量，$\psi_e(\cdot)$，$\psi_h(\cdot)$，$\psi_\alpha(\cdot)$为全连接网络，其非线性激活函数均为 ReLU，$\boldsymbol{W}_e$，$\boldsymbol{W}_h$，$\boldsymbol{W}_\alpha$ 分别为全连接网络的权重值，具体的网络参数将在实验部分给出。

利用每一个时刻的交互特性向量 $\boldsymbol{e}_m$ 与相应的注意力得分 α_m，可以通过所有时刻交互特性向量的线性组合求出行人 P_{H}^{i} 最终的状态输出 $\boldsymbol{e}^{\mathrm{o}}$：

$$\boldsymbol{e}^{\mathrm{o}}=\sum_{k=1}^{T}\mathrm{softmax}(\boldsymbol{a}_k)\boldsymbol{e}_m \tag{8}$$

最后，通过 LSTM 网络处理环境中不定数量的障碍物，最后得到定长的输出向量 $\boldsymbol{h}^{\mathrm{o}}$。

2.2　舒适奖励函数

机器人在与人交互的场景中，如果行人没有向机器人发起服务需求，为了保证行人舒适度的要求，机器人应该尽量避免出现在行人的舒适范围内，即机器人不应该进入行人的私人距离内。除此之外，由文[13]可知，机器人从后向前超越行人的横向舒适距离是 0.7 m，如果小于该距离同样会引发行人的不舒适感。因此，本文依据上述行人舒适度的要求修改了奖励函数，对机器人进入行人舒适范围内的状态进行了惩罚。

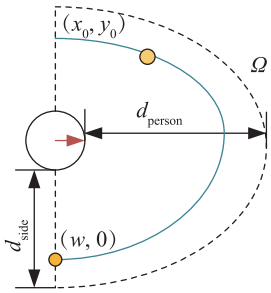


图 5　行人私人距离区域

Fig.5　Personal distance zones for people

依据行人舒适度的知识，定义行人舒适区域 Ω：以行人为坐标原点，前进方向为 x 轴，建立行人坐标系。行人舒适区域通过以行人为原点，私人距离最大值 d_{person}为长轴，侧边舒适距离 d_{side}为短轴的半椭圆 Ω 表示(只考虑出现在行人视野范围内的情况，d_{side}为 0.7 m，d_{person}为 1.2 m)。

假设机器人在行人坐标系中的坐标为(x_0，y_0)，通过指示函数 $I(x_0, y_0)$即可判断机器人是否行驶到了行人的舒适区域 Ω 内，$I(x_0, y_0)$为 1 表示机器人进入了行人舒适区域，$I(x_0, y_0)$为 0 表示机器人未进入行人舒适区域。

$$I(x_0, y_0)=\begin{cases}0, & \dfrac{x_0^2}{d_{\mathrm{person}}^2}+\dfrac{y_0^2}{d_{\mathrm{side}}^2}>1\\[2ex] 1, & \dfrac{x_0^2}{d_{\mathrm{person}}^2}+\dfrac{y_0^2}{d_{\mathrm{side}}^2}\leqslant 1\end{cases} \tag{9}$$

如果机器人到达了行人的行人舒适区域内，即 $I(x_0, y_0)=1$，会引发行人的不舒适感，需要给予机器人相应的惩罚，惩罚的大小取决于机器人在行人舒适区域内的位置，距离行人越近惩罚越大。如图 5 所示，机器人为黄色标志，如果机器人(黄色圆圈)到达了行人舒适区域内，可以假设在行人舒适区域 Ω 内机器人坐标所在的与行人舒适区域具有相同长短轴比的椭圆上具有相同的惩罚，即图 5 所示蓝色椭圆上的所有点具有相同的惩罚项，所以

(x_0, y_0)与$(w, 0)$处具有相同的惩罚。由于与人的距离越近越容易引发人的不适，且不适感会越来越大，因而采用指数函数表征该惩罚值，惩罚的最大值为r_p，所以得到的与w有关的惩罚函数如下：

$$R_t^{\text{social}}(w) = -r_p(\mathrm{e}^{-(w-d_{\text{side}})} - 1) \tag{10}$$

其中，w为蓝色椭圆的短轴：

$$w = \sqrt{\frac{(x_0^2 \cdot d_{\text{side}}^2)}{d_{\text{person}}^2 + y_0^2}} \tag{11}$$

原始的奖励函数$R_t^{\text{o}}(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t)$为文[17]中所使用的奖励函数，其中去除了当机器人与障碍物相距较近时的惩罚项：

$$R_t^{\text{o}}(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t) = \begin{cases} -0.25, & \text{if} \quad d_t < 0 \\ 2, & \text{else if} \quad p_t = p_{\text{g}} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

同时，为了避免机器人角速度变化过大的情况，对角速度变化过大的情况进行了惩罚：

$$R_t^{\text{A}}(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t) = \begin{cases} -r_{\text{Angle}}, & \text{if} \quad \varphi_t - \varphi_{t-1} > \pi \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

最终的奖励函数为

$$R_t(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t) = R_t^{\text{o}}(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t) + R_t^{\text{social}}(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t) + R_t^{\text{A}}(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t) \tag{14}$$

通过该奖励函数，可以使得机器人的避障策略更加符合人的舒适度准则。

2.3 网络训练

整个强化学习算法的训练方法采取了时序差分法，为了增加样本的利用率，将每次强化学习与环境交互的信息存储在缓存回放区(replay buffer)中，在训练网络时通过均匀采样的方法从缓存回放区中进行采样，得到训练数据，并通过固定目标网络的方法来加速算法的收敛并减小算法的方差。同时，为了进一步加快收敛，训练前利用传统方法 ORCA[3] 生成的数据集进行网络的预训练，详细的训练算法如算法 1 所示。

算法 1：网络训练

1：使用传统算法 ORCA 生成预训练数据 D
2：使用预训练数据 D 进行模仿学习，初始化表现模型(behaviour network)$\boldsymbol{\mu}(\cdot)$，$\boldsymbol{\mu}(\cdot) = [\psi_e(\cdot), \psi_h(\cdot), \psi_\alpha(\cdot), \boldsymbol{\Phi}_{\text{LSTM}}(\cdot), \psi_M(\cdot)]$
3：初始化目标网络(target network)的参数：$\tilde{\boldsymbol{\mu}}(\cdot) \leftarrow \boldsymbol{\mu}(\cdot)$
4：初始化经验回放池：$\boldsymbol{E} \leftarrow \boldsymbol{D}$
5：for episode = 1, M do
6：　随机初始化动态障碍物的状态 $\boldsymbol{s}_0^{\text{jn}}$
7：　repeat：
8：　$\boldsymbol{a}_t \leftarrow \text{plan}(\boldsymbol{s}_{t+\Delta t}^{\text{jn}})$
9：　将样本数据$(\boldsymbol{s}_t^{\text{jn}}, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+\Delta t}^{\text{jn}})$存入经验回放池
10：　从经验回放池中随机选取 batch 个样本
11：　通过目标网络(target network)计算样本输出：$y_i = r_i + \gamma^{\Delta t \cdot v_{\text{pref}}} \tilde{\mu}(\boldsymbol{s}_{i+\Delta t}^{\text{jn}})$
12：　通过梯度下降法更新表现模型
13：　until 终止状态或者 $t \geqslant t_{\max}$
14：　更新目标网络(target network)$\tilde{\boldsymbol{\mu}}(\cdot) \leftarrow \boldsymbol{\mu}(\cdot)$
15：end for
16：return $\boldsymbol{\mu}(\cdot)$

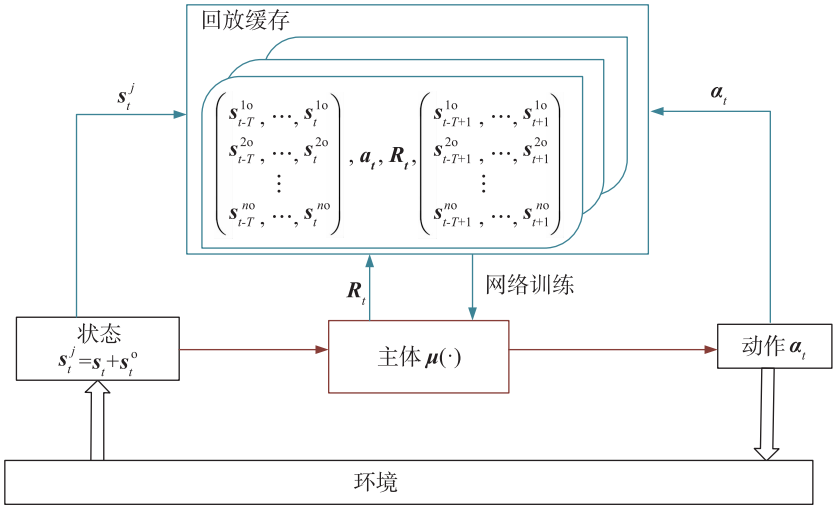


图 6　网络训练

Fig.6　Network training

3　实验验证

本章将通过比较不同编码方式的避障算法的性能，以及修改奖励函数前后避障算法的性能，分别验证改进的值函数网络的有效性与泛化性以及舒适奖励函数的可行性。

硬件平台：实验的硬件平台为 Intel Core i7-7700 3.60 GHz，NVIDIA Geforce GTX 1050 Ti 的台式机。

仿真环境：所有的实验均在基于 Gym 搭建的仿真环境中进行，在仿真环境中，使用半径为0.3 m 的圆形代替行人与机器人，行人与机器人的任务为从起始点运行到目标点。其中，行人的运动使用传统避障算法 ORCA[3] 策略进行控制。起始点与目标点的选取采用了交叉相遇的方式，即行人和机器人的初始位置随机定位在同一个半径为 4 m 的圆圈上，目标点为其初始位置关于圆心中心对称的点。

机器人动作空间设置：仿真中的作为机器人的角速度与线速度，实验中选取机器人离散的动作空间为机器人线速度空间与角速度空间的组合，大小为 80，其中线速度空间在[0，v_{pref}]中均匀采样，角速度空间在[0，2π]中均匀采样，采样大小分别为 5 和 16。

网络设置：本文首先通过 ORCA[3] 这一传统方法收集得到了 3 000 个训练回合的数据完成了网络权重的初始化。预训练结束后，为了平衡对未知状态的探索与已有结果的利用，采用 e-greedy 的方法进行训练。在训练开始时，为了能够更佳有效地探索环境中到的未知状态，设置探索率为 0.5，之后随着训练次数的增加，不断减小探索率，在 5 000 回合时减小为 0.1，之后的训练将不再改变探索率。其余的超参数和实验过程中值函数的网络参数分别如表 1、表 2 所示。

表 1　超参数设置

Tab.1　Hyper parameters setting

参数	预训练	训练
N	6	6
v_{pref}	1	1
K	12	12
$r_{\max}$	3	3
T	5	5
学习率	0.01	0.001
batch-size	100	100
Δt	0.25	0.25
γ	0.9	0.9

表 2　网络参数取值

Tab.2　Parameters of network

网络	网络参数设置
$\psi_e(\cdot)$	[150，100]
$\psi_h(\cdot)$	[150，50]
$\psi_\alpha(\cdot)$	[100，100，1]
$\Phi_{\text{LSTM}}(\cdot)$	[50，50]
$\psi_M(\cdot)$	[150，100，100，1]

3.1　改进的值函数网络实验验证

3.1.1　可行性实验验证

本小节通过与 ORCA[3]，CADRL[27] 以及基于社会力(social force)与局部地图(local map)两种不同编码方法的避障算法进行比较，验证了本文改进的值函数网络的有效性。为了避免舒适奖励函数对该模块的影响，此处的奖励函数采取论文[24]中的奖励函数，因此，基于强化学习的避障算法的不同之处仅仅在于对行人交互信息的处理方式。为了表述方便，将基于社会力编码的算法称为 SF_RL，基于局部地图编码的方式称为 LM_RL，本文提出的基于 APG 编码的算法称为 APG_RL。

图 7 为 CADRL、SF_RL、LM_RL 以及 APG_RL 四种基于强化学习的避障算法在行人数量为 6 的仿真环境中，经过 10 000 个回合的训练，累积回报函数的变化曲线。其中，CADRL、LM_RL 以及 APG_RL 三种算法的收敛性能明显高于 SF_RL，这是由于使用社会力模型对行人之间的交互信息的描述可能存在误差，因而降低了网络的收敛性。

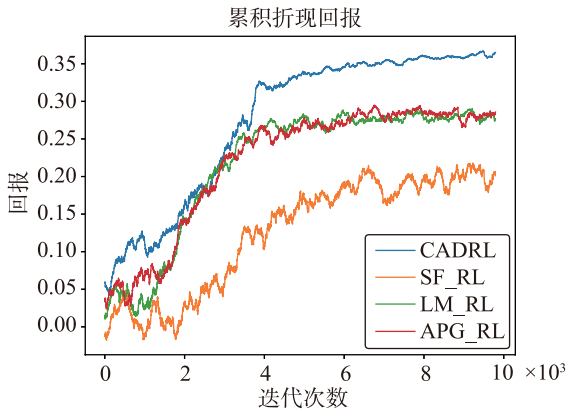


图 7　不同算法的学习回报曲线

Fig.7　Learning return curve of different algorithms

训练结束后，在行人数量为 6 的仿真环境中对上述 5 种算法分别进行了 500 次测试，测试的结果如表 3 所示。其中，基于强化学习的算法使用多层感知机拟合值函数，并通过机器人与环境的不断交

互，在一定程度上学习得到了行人之间的交互特性，提升了机器人在行人密集环境中的避障性能，在测试过程中的成功率高于传统的 ORCA 算法。但由于提取行人交互信息方式的不同，其算法的性能也存在差别。CADRL 由于未对行人交互特征进行处理，在测试过程中的避障成功率较低，表明仅仅使用一个浅层的网络不能很好地提取复杂动态环境中的信息。而 SF_RL、LM_RL 以及 APG_RL 由于增加了对行人交互特征的提取以及行人的时序特征，具备了一定的先验知识，大大提升了避障的成功率。相较于 SF_RL，使用局部信息编码的避障算法 LM_RL 与 APG_RL 在网络训练过程中具有更好的收敛性，且避障用时更短，效率更高。

表 3 不同算法性能比较

Tab.3 Quantitative results comparison among different methods

算法	成功率	失败率	平均时间 /s	回报
ORCA	0.33	0.66	11.04	−0.065 2
CADRL	0.59	0.40	10.69	0.073 8
SF_CADRL	0.96	0.03	11.86	0.260 8
LM_RL	0.99	0.01	10.97	0.319 0
APG_RL	0.99	0.01	10.82	0.319 6

图 8 为不同的避障算法在同一测试场景的避障结果，其中黄色的实心圆形代表机器人，其余颜色的圆圈代表不同行人，圆内数字表示机器人当前的运行时间。图 8(a)中机器人控制算法采用 ORCA，机器人在 3.5 s 时与行人发生了碰撞，导致导航失败，

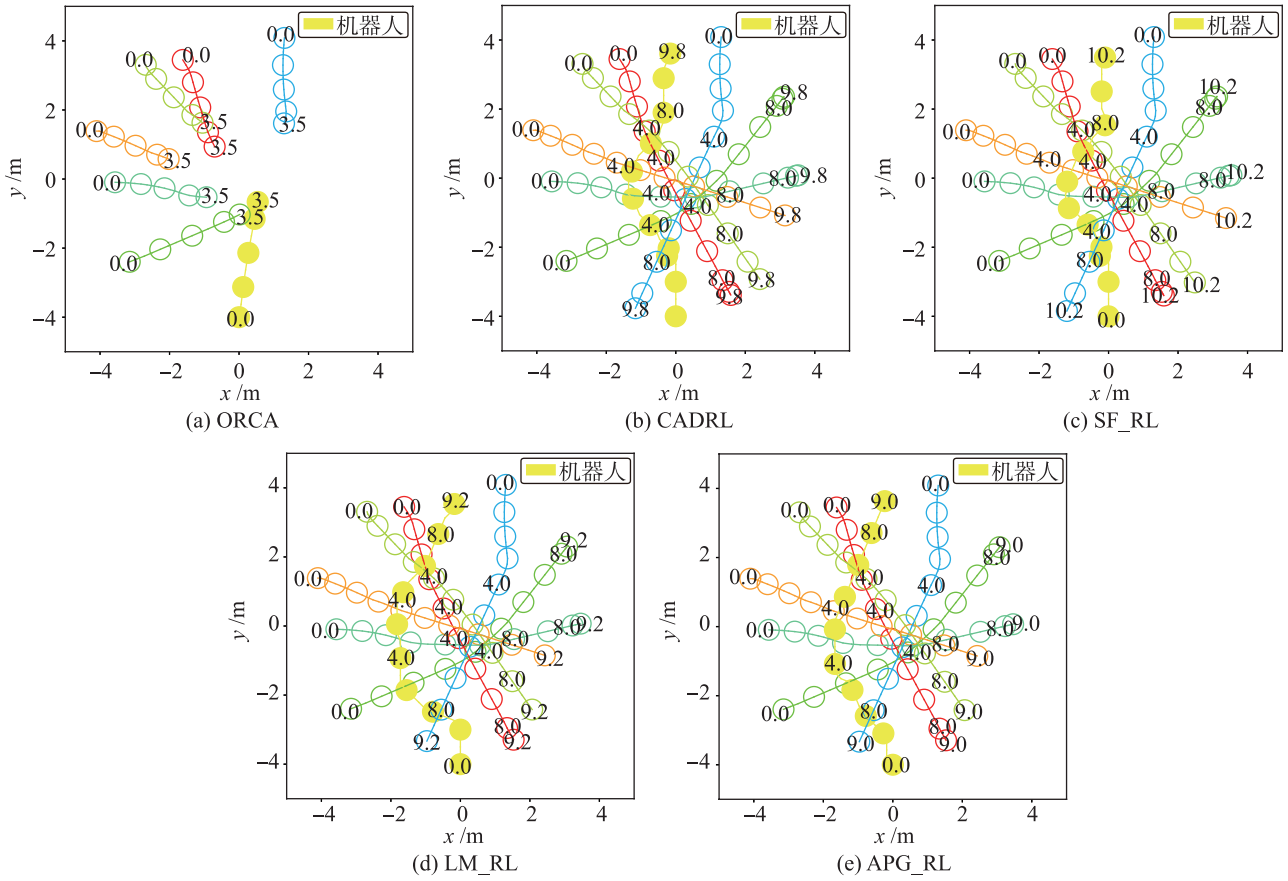


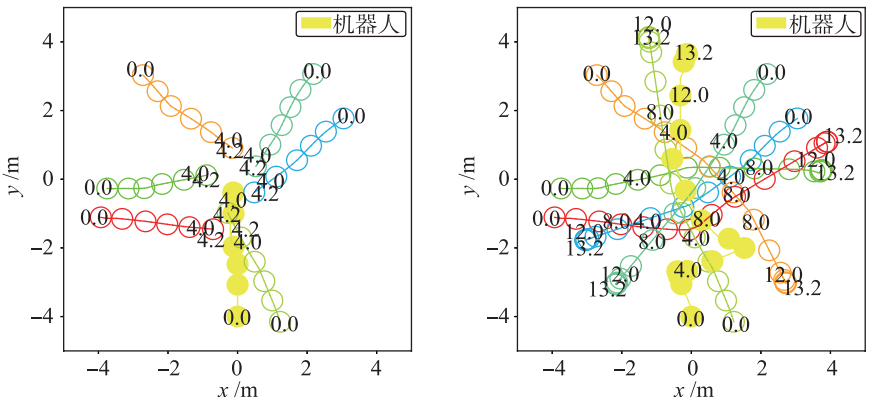
图 8 不同算法同一仿真条件机器人轨迹比较

Fig.8 Comparison of robot trajectories with different algorithms under the same simulation condition

其余方法分别在 9.8 s、10.2 s、9.2 s、9.0 s 到达了目标点。其中，APG_RL 在最短的时间内到达了目标点，结合表 3 可知，APG_RL 经过 500 次测试的平均导航时间仅为 10.82 s，极大地提升了避障的效率，很好地验证了本文改进的值函数网络的有效性。

除了更高的导航效率，经过实验发现，APG_RL 可以适应某些更加复杂的环境，在 LM_RL 避障失败的情况下仍然可以完成导航任务，具体的轨迹图如图 9 所示。如图 9(a)所示，6 个行人从不同的方向圆心行驶，在第 4 s 的时候，LM_RL 算法控制的机器人由于行驶到了行人中间，与行人发生了碰撞。而 APG_RL 由于采取了角度行人网格对周围的行人进行编码，获取得到的是周围行人连续的值，相比于局部地图的栅格化处理，可以更加精确

地捕获行人的变化，因而在图 9 所示复杂环境中，机器人在距离行人较近时便改变了运动方向，避免了进入人群的情况，如图 9(b)所示，机器人在 4 s 时向右转弯，最终成功地到达了目标点。



(a) LM_RL　　(b) APG_RL

图 9　LM_RL 与 APG_RL 同一仿真条件机器人轨迹比较

Fig.9　Trajectory comparison in a test case between LM_RL and APG_RL

3.1.2　泛化性实验验证

APG_RL 算法在训练时仿真环境中行人的数量为 6，为了进一步测试模型的泛化性能，在不重新训练网络的基础上，改变仿真环境行人的数量 N，表 4 给出了在不同行人个数时进行 500 次避障测试的结果。实验结果显示，在行人数量发生变化的情况下，机器人的避障成功率均保持在 0.91 以上。该实验充分证明了 APG_RL 算法可以很好地应对环境中行人变化的情况，尤其在行人数量为 8 的仿真环境中依旧可以取得较高的成功率，因而可以更好地应对复杂的行人环境。

表 4　APG_RL 在不同数量行人时仿真验证结果

Tab.4　Simulation results of different numbers of pedestrians using APG_RL

行人个数	3	4	5	6	7	8
成功率	1.00	1.00	0.99	0.99	0.98	0.91
平均时间	10.20	10.18	10.81	10.82	12.30	12.75

3.2　行人舒适度奖励函数实验验证

本实验对本文提出的奖励函数的可行性进行了实验验证，r_{p} 与 r_{Angle} 的取值均为 0.02。在不改变 APG_RL 避障算法的基础上，依照式(13)设计的舒适奖励函数进行训练，分别得到了舒适避障策略 APG_CRL 与 APG_CARL，其中 APG_CRL 的奖励函数仅添加了 $R_t^{\mathrm{social}}(w)$ 项，APG_CARL 的奖励函数添加了 $R_t^{\mathrm{social}}(w)$ 与 $R_t^{\mathrm{A}}(\boldsymbol{s}_t^{\mathrm{jn}}, \boldsymbol{a}_t)$ 两项内容。表 5 为上述 3 种算法在行人数量为 6 的仿真环境中测试 500 次得到的结果。

表 5　APG_RL 使用不同奖励函数仿真测试结果

Tab.5　Simulation test results of different reward function using APG_RL

算法	成功率	失败率	平均时间 /s	距离 /m
APG_RL	0.99	0.01	10.82	0.14
APG_CRL	0.99	0.01	10.96	0.64
APG_CARL	0.99	0.01	10.77	0.50

3.2.1　行人舒适度奖励函数实验验证

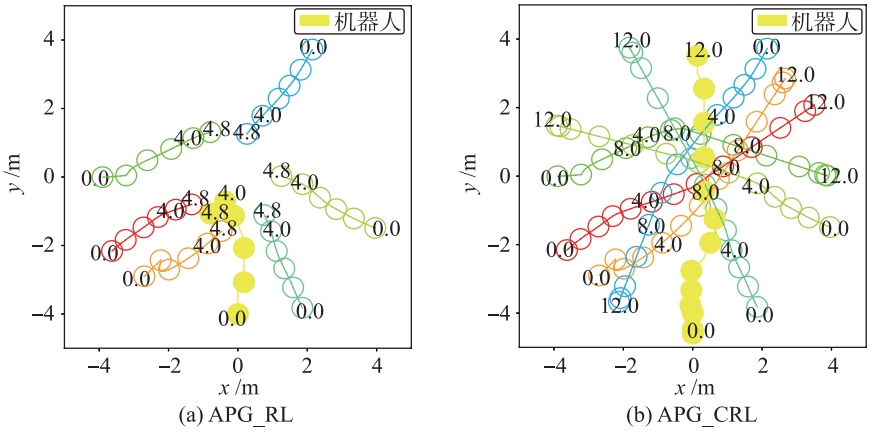
通过表 5 可知，APG_RL 与 APG_CRL 均拥有较高的导航成功率，而机器人在平均到达目标点的时间仅仅增加了 0.14 s 的情况下，距离行人的最近距离从 0.14 m 增大到 0.64 m，很好地平衡了导航效率与行人舒适度的要求。

除此之外，行人舒适度奖励函数不仅提高了导航时与行人的最大距离，与 APG_RL 避障算法相比，APG_CRL 由于惩罚进入行人私人距离内的状态，可以很好地适用于某些环境更加复杂的情况。在图 10(a)所示的动态复杂环境下，APG_RL 由于仅仅在机器人距离行人较近的情况下才对机器人进行惩罚，机器人会向人群聚集的地方行驶，而 APG-CRL 通过对机器人进入行人舒适范围的状态进行惩罚，使得机器人倾向于选择与人保持较远距离的行为，如图 10(b)所示，机器人会在原地等待行人离开后，再向目标点行驶。

3.2.2　角速度奖励函数实验验证

角度变化太过剧烈不仅不利于实际机器人的控制、传感器数据的获取等，而且对于周围环境中的行

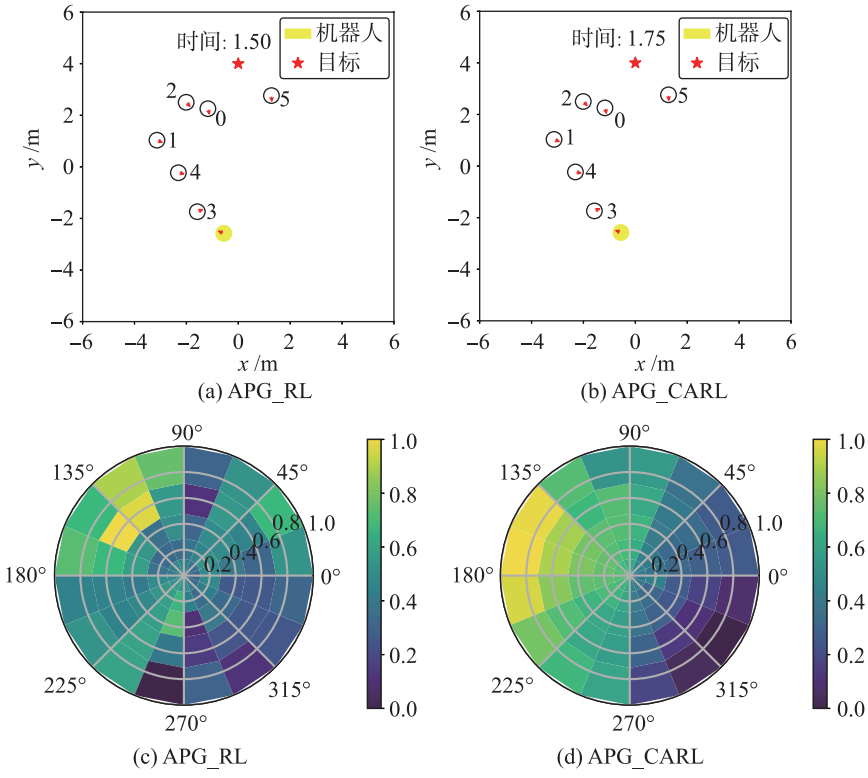
人也会产生影响，因而 APG_CARL 对机器人角度变化过大的情况进行惩罚。由表 5 可知，与 APG_CRL 相比，在对角度进行惩罚后，其平均到达目标点的时间减少了 0.19 s，APG_CARL 由于避免了大幅度的角度变化，可以减少机器人得震荡行为，一定程度上提升了避障得效率。其某次避障任务中机器人的动作选择如图 11 所示，图 11(a)、11(b)分别表示了 APG_CRL 与 APG_CARL 所处的仿真环境，其中红色箭头代表了机器人的运动方向。图 11(c)、11(d)分别表示机器人在执行动作空间中的动作后所获得的价值，颜色越浅代表执行该动作后获得的价值越高。APG_RL 由于仅仅考虑了行人的舒适距离，因而在图 11 所示的情况下，机器人倾向于选择角度变化更大但距离行人更远的角速度，而 APG_CARL 对角速度变化较大的情况进行了惩罚，在相同情况下，角速度的最优选择范围集中在当前角速度附近，不会造成控制信号大幅度的跳变，更加便于今后实体机器人的控制。



(a) APG_RL (b) APG_CRL

图 10 不同奖励函数同一仿真条件测试轨迹比较

Fig.10 Trajectory comparison in a test case among different reward functions



(a) APG_RL (b) APG_CARL

(c) APG_RL (d) APG_CARL

图 11 机器人执行每一个可行动作的值函数结果

Fig.11 The value function result of each feasible action performed by the robot

4　结论

本文提出的移动机器人舒适避障算法首先针对浅层的值函数网络难以拟合复杂行人环境的不足，对值函数进行了改进，设计了行人交互信息模块，通过角度行人网格提取行人之间的交互信息，并且利用注意力机制提取行人行走轨迹的时序特征。其次，通过修改奖励函数的方法将行人舒适度要求引入了避障策略中，使得机器人的避障策略更加符合人机交互的需求。通过仿真实验比较分析，本文提出的算法在人群密集的复杂动态环境中不仅拥有良好的避障成功率与适应性，而且可以满足行人舒适度的要求。

本文今后将继续研究不同行人舒适区域及动态障碍物提取方法对算法性能的影响。除此之外，本文提出的算法仅仅在仿真取得了较好的效果，后续研究将准备将该理论方法应用到实际的移动机器人，如 NAO、Pepper 机器人上，通过实际应用进一步改进算法，使其可以适应真实环境中传感器信息不准确等环境因素。

参考文献

[1] Mellinger D, Kushleyev A, Kumar V. Mixed-integer quadratic program trajectory generation for heterogeneous quadrotor teams [C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2012: 477 – 483.

[2] Fiorini P, Shiller Z. Motion planning in dynamic environments using velocity obstacles[J]. The International Journal of Robotics Research, 1998, 17(7): 760 – 772.

[3] Peng H, Zhen L. Crowd simulation research based on reciprocal velocity obstacle collision avoidance[J]. Computer Simulation, 2012. DOI: 1006-9348(2012)11-0034-04.

[4] 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. Mobile robot path planning based on improved A* algorithm and dynamic window method [J]. Robot, 2020, 42(3): 346 – 353.

[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] Burgard W. Controlling synchrodrive robots with the dynamic window approach to collision avoidance[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 1996: 1280 – 1287.

[7] Trautman P, Krause A. Unfreezing the robot: Navigation in dense, interacting crowds[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2010: 797 – 803.

[8] Ferrer G, Garrell A, Sanfeliu A. Social-aware robot navigation in urban environments[C]//European Conference on Mobile Robots. Piscataway, USA: IEEE, 2013: 331 – 336.

[9] Kruse T, Pandey A K, Alami R, et al. Human-aware robot navigation: A Survey[J]. Robotics and Autonomous Systems, 2013, 61(12): 1726 – 1743.

[10] Pacchierotti E, Christensen H I, Jensfelt P. Human-robot embodied interaction in hallway settings: A pilot user study[C]//International Workshop on Robot and Human Interactive Communication. Piscataway, USA: IEEE, 2005: 164 – 171.

[11] 张大鹏，肖峰. 基于个人空间理论的人群疏散机理研究[J]. 交通运输工程与信息学报，2016，14(2)：144 – 152.
Zhang D P, Xiao F. Research on the mechanism of crowd evacuation based on the theory of personal space[J]. Journal of Transportation Engineering and Information, 2016, 14(2): 144 – 152.

[12] 孙月，刘景泰. 基于 RGB-D 传感器的室内服务机器人舒适跟随方法[J]. 机器人，2019，41(6)：823 – 833.
Sun Y, Liu J T. Comfortable follow method of indoor service robot based on RGB-D sensor[J]. Robot, 2019, 41(6): 823 – 833.

[13] Pacchierotti E, Christensen H I, Jensfelt P. Evaluation of passing distance for social robots[C]//International Symposium on Robot and Human Interactive Communication. Piscataway, USA: IEEE, 2006: 315 – 320.

[14] Zhao Y, Tian G H, Yin J Q, et al. Human trajectory analysis method based on hidden markov model in home intelligent space [J]. Pattern Recognition and Artificial Intelligence, 2015, 28(6): 542 – 549.

[15] Wiest J, Hoffken M, Kresel U, et al. Probabilistic trajectory prediction with Gaussian mixture models[C]//Intelligent Vehicles Symposium. Piscataway, USA: IEEE, 2012: 141 – 146.

[16] Burstedde C, Klauck K, Schadschneider A, et al. Simulation of pedestrian dynamics using a two-dimensional cellular automaton

[J]. Physica A, 2001, 295(3/4): 507 – 525.
[17] Helbing D, Molnar P. Social force model for pedestrian dynamics[J]. Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 1998, 51(5): 4282 – 4286.
[18] Hoogendoorn S, Bovy P. Gas-kinetic modeling and simulation of pedestrian flows[J]. Transportation Research Record: Journal of the Transportation Research Board, 2000, 1710: 28 – 36.
[19] Alahi A, Goel K, Ramanathan V, et al. Social LSTM: Human trajectory prediction in crowded spaces[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 961 – 971.
[20] Gupta A, Johnson J, Li F F, et al. Social GAN: Socially acceptable trajectories with generative adversarial networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 2255 – 2264.
[21] Vemula A, Muelling K, Oh J. Social attention: Modeling attention in human crowds[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 1 – 7.
[22] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529 – 533.
[23] Faust A, Oslund K, Ramirez O, et al. PRM-RL: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 5113 – 5120.
[24] Chiang H T L, Hsu J, Fiser M, et al. RL-RRT: Kinodynamic motion planning via learning reachability estimators from RL policies[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4298 – 4305.
[25] Tai L, Li S, Liu M. A deep-network solution towards model-less obstacle avoidance[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2016: 2759 – 2764.
[26] Chiang H T L, Faust A, Fiser M, et al. Learning navigation behaviors end-to-end with AutoRL[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 2007 – 2014.
[27] Chen Y F, Liu M, Everett M, et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 285 – 292.
[28] Everett M, Chen Y F, How J P. Motion planning among dynamic, decision-making agents with deep reinforcement learning [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 3052 – 3059.
[29] Chen Y F, Everett M, Liu M, et al. Socially aware motion planning with deep reinforcement learning[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 1343 – 1350.
[30] Hall E T. The hidden dimension[M]. 1st ed. New York, USA: Anchor Books-Doubleday, 1966.
[31] Pandey A K, Alami R. A framework for adapting social conventions in a mobile robot motion in human-centered environment [C]//IEEE International Conference on Advanced Robotics. Piscataway, USA: IEEE, 2009: 1 – 8.
[32] Pfeiffer M, Paolo G, Sommer H, et al. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 5921 – 5928.
[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[M]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2017: 5998 – 6008.

作者简介

孙立香(1981 –), 女, 硕士, 高级技师, 讲师。研究领域为智能控制。
孙晓娴(1995 –), 女, 硕士生。研究领域为机器人导航与运动控制。
刘成菊(1980 –), 女, 博士, 教授。研究领域为机器人运动控制。