智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

动态环境下分布式异构多机器人避障方法研究

欧阳勇平, 魏长赟, 蔡帛良

引用本文:
欧阳勇平,魏长,蔡帛良. 动态环境下分布式异构多机器人避障方法研究[J]. 智能系统学报, 2022, 17(4): 752–763.
OUYANG Yongping,WEI Changyun,CAI Boliang. Collision avoidance approach for distributed heterogeneous multirobot systems in dynamic environments[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(4): 752–763.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202106044

DOI: 10.11992/tis.202106044
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220505.1636.004.html

动态环境下分布式异构多机器人避障方法研究

欧阳勇平[1], 魏长赟[1], 蔡帛良[1,2]
（1. 河海大学 机电工程学院, 江苏 常州 213022; 2. 英国卡迪夫大学 工学院, 威尔士 卡迪夫 CF10 3A）

摘　要: 多机器人系统在联合搜救、智慧车间、智能交通等领域得到了日益广泛的应用。目前, 多个机器人之间、机器人与动态环境之间的路径规划和导航避障仍需依赖精确的环境地图, 给多机器人系统在非结构环境下的协调与协作带来了挑战。针对上述问题, 本文提出了不依赖精确地图的分布式异构多机器人导航避障方法, 建立了基于深度强化学习的多特征策略梯度优化算法, 并考虑了人机协同环境下的社会范式, 使分布式机器人能够通过与环境的试错交互, 学习最优的导航避障策略; 并在 Gazebo 仿真环境下进行了最优策略的训练学习, 同时将模型移植到多个异构实体机器人上, 将机器人控制信号解码, 进行真实环境测试。实验结果表明: 本文提出的多特征策略梯度优化算法能够通过自学习获得最优的导航避障策略, 为分布式异构多机器人在动态环境下的应用提供了一种技术参考。

关键词: 异构多机器人; 深度强化学习; 非结构环境; 多特征策略梯度; 动态避障; 自学习; 分布式控制; 控制策略
中图分类号: TP273+.2　　**文献标志码**: A　　**文章编号**: 1673−4785(2022)04−0752−12

中文引用格式: 欧阳勇平, 魏长赟, 蔡帛良. 动态环境下分布式异构多机器人避障方法研究 [J]. 智能系统学报, 2022, 17(4): 752–763.
英文引用格式: OUYANG Yongping, WEI Changyun, CAI Boliang. Collision avoidance approach for distributed heterogeneous multirobot systems in dynamic environments[J]. CAAI transactions on intelligent systems, 2022, 17(4): 752–763.

Collision avoidance approach for distributed heterogeneous multirobot systems in dynamic environments

OUYANG Yongping[1], WEI Changyun[1], CAI Boliang[1,2]
(1. College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213022, China; 2. School of Engineering, Cardiff University, Cardiff CF10 3AT, UK)

Abstract: Multirobot systems have been widely used in cooperative search and rescue missions, intelligent warehouses, intelligent transportation, and other fields. At present, the path planning and collision avoidance problems between multiple robots and the dynamic environment still rely on accurate maps, which brings challenges to the coordination and cooperation of multirobot systems in unstructured environments. To address the above problem, this paper presents a navigation and collision avoidance approach that does not require accurate maps and is based on the deep reinforcement learning framework. A multifeatured policy gradients algorithm is proposed in this work, and social norms are also integrated so that the learning agent can obtain the optimal control policy via trial-and-error interactions with the environment. The optimal policy is trained and obtained in the Gazebo environment, and afterward, the optimal policy is transferred to several heterogeneous real robots by decoding the control signals. The experimental results show that the multifeature policy gradients algorithm proposed can obtain the optimal navigation collision avoidance policy through self-learning, and it provides a technical reference for the application of distributed heterogeneous multirobot systems in dynamic environments.

Keywords: heterogeneous multi-robot systems; deep reinforcement learning; non-structural environment; multi-feature policy gradients; dynamic collision avoidance; self-learning; distributed control; control policy

收稿日期: 2021−06−25.　**网络出版日期**: 2022−05−06.
基金项目: 国家自然科学基金项目(61703138); 中央高校基本科研业务费项目(B200202224).
通信作者: 魏长赟. E-mail: c.wei@hhu.edu.cn.

随着多机器人系统 (multi-robot system, MRS) 的广泛应用, 其路径规划和导航避障领域一直是

学者们关注的热点话题。传统的机器人避障算法主要有粒子群寻优算法[1]、基于障碍物的几何构型得到避障策略[2]、Khatib[3] 提出了最优避碰策略 (optimal reciprocal collision avoidance, ORCA) 及其衍生的其他避障算法等, 但这些导航模型在环境复杂的情况下调整效果不佳, 不适用于动态环境。近年来, 在基于强化学习的多机器人导航避障算法中, 相关学者们提出了构建状态空间到动作空间的映射的控制逻辑, 也即策略映射[4-8], 其中 Zhang 等[4] 提出了一种以深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 为基础的机器人控制模型, 最终构建了基于激光雷达和位置信息的策略映射, 但算法的收敛速度慢, 训练效率较低。Chen[5] 则提出了一种异步 DDPG 算法 (asynchronous DDPG, ADDPG), 使用多个机器人在同一个实验环境中进行实验, 提高了经验的搜集效率, 缩短了算法的训练时间, 但没有考虑移动机器人的导航避障规则。因此设计一种受客观条件限制较低, 且可以实现人机协同的机器人避障算法对于提高异构多机器人的工作效率和安全性具有重要意义。

故本文在此提出了基于深度强化学习的多特征策略梯度优化算法, 并引入人机协同环境下的社会范式以及提出经验优先采样机制, 不仅使多机器人移动按照一定规则避障, 而且提高了算法的训练速度以及控制精度, 同时搭建了分布式多机器人的控制模型, 在 Gazebo 仿真环境下进行算法的训练学习, 最后在现实环境下的多异构机器人平台上验证了导航避障方法的可行性。

1　问题描述

多机器人路径规划是在工作环境中为各个机器人都找到一条从起始点到目标点的最优无碰撞路径。其中, 单个机器人前往目标点不仅需要考虑与环境内的障碍物避免碰撞, 还需避免在移动时与其他机器人发生碰撞。因此, 如何为环境内的各个机器人在寻找路径时不发生碰撞是存在的难题。

本文以深度强化学习模型为基础, 为解决异构多机器人在动态环境下导航避障问题, 搭建了不依赖精确地图机器人导航避障模型, 机器人仅装有激光雷达传感器, 只需对原始的测量信号进行获取和处理, 即可实现异构多机器人的导航避障。具体模型如图 1 所示。

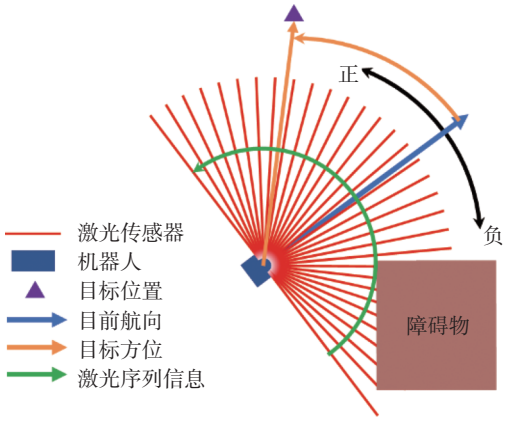


图 1　机器人导航避障模型

Fig. 1　Robot navigation and collision avoidance model

在图中机器人所获得的状态信息为目标点相对自身的极坐标位置(ρ,θ)(图中黄色箭头), 激光传感器所捕获的周围环境信息$\varsigma=[l_1,l_2,\cdots,l_{128}]$(红色线)且每条激光编号从 0 开始沿图中绿色箭头方向递增。

其中, 极坐标位置信息 (ρ,θ)的计算公式为

$$\begin{cases}\rho=\sqrt{(x_t-x_w)^2+(y_t-y_w)^2}\\ \theta=\text{sign}(\boldsymbol{v}_t\times\boldsymbol{v}_w)\arccos\left(\dfrac{\boldsymbol{v}_t\cdot\boldsymbol{v}_w}{|\boldsymbol{v}_t||\boldsymbol{v}_w|}\right)\end{cases}$$

式中: (x_t,y_t)和(x_w,y_w)分别是机器人和目标点在全局坐标系下的坐标, $\boldsymbol{v}_t$和$\boldsymbol{v}_w$分别是机器人的速度矢量和从机器人指向目标点的矢量。

最终上述两部分信息经过归一化处理后连同上一时刻的机器人动作信息a_{t-1}被组成一个状态信息s_t, 并传递给强化学习算法进行计算, 其中归一化公式为

$$\begin{cases}\bar{\rho}=\rho/\xi\\ \bar{\theta}=\theta/\pi\\ l_i=l_i/l_{\max}\end{cases}$$

式中: ξ是环境中的最大对角线长度, 最大感知距离$l_{\max}=3$ m。基于上述提出的不依赖精确地图导航的异构多机器人导航避障模型, 本文将在后续仿真与真实实验中进一步介绍。

2　深度强化学习控制模型

2.1　马尔可夫决策过程

为各个机器人寻找一条最优无碰撞路径问题可以简化为马尔可夫决策过程。马尔可夫决策过程 (Markov decision process, MDP) 作为强化学习理论的基础, 具有重要的理论价值。MDP 的数学要素可以表述为五元组, 即(S,A,P,R,γ), 其中:

S表示状态空间, 表示 MDP 所在环境下所有可能状态的集合;

A表示动作空间, 表示对应状态下所有可采取

的动作的集合；

P表示状态的条件转移概率，表示代理在t时刻s_t状态下采取动作a后，在$t+1$时刻的状态s_{t+1}的状态为s'的概率，其公式表述为

$$P(s'|s,a)=P(s_{t+1}=s'|s_t=s,a_t=a)$$

R为 MDP 的评价函数，是算法在s状态下执行动作a后变换为s'的过程对算法目标结果好坏的量化评价标准，其定义为

$$R(s,a,s')=\mathrm{E}[R_{t+1}|s_t=s,a_t=a,s_{t+1}=s'] \tag{1}$$

γ是折扣因子，表示 MDP 中每一个决策环节对相对于决策过程中未来的决策环节的重要性，$\gamma\leqslant 1$恒成立，表示算法更看重当前奖励而不是未来的奖励。

MDP 的实际流程可以表示为图 2，环境中受算法控制的代理对象在状态$s_0\in S$的条件下初始化，并由算法根据状态s_0选择建议动作$a_0\in S$并由代理对象完成该动作，环境根据所执行的动作，根据条件转移概率$P(s_1|s_0,a_0)$转移至下一状态s_1，同时，环境给出对应奖励$r_0(s_0,a_0,s_1)$。此后算法根据状态s_1选择新的建议动作，并重复执行上述步骤直至达到终止条件。

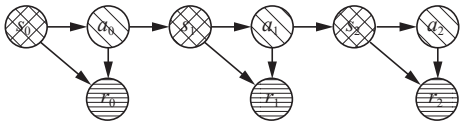


图 2　马尔可夫决策过程

Fig. 2　Markov decision process

2.2　深度强化学习

深度强化学习 (deep reinforcement learning) 算法模型是以一种通用的形式将深度学习的感知能力与强化学习的决策能力相结合，并能够通过端对端的学习方式实现从原始输入到输出的直接控制，在模拟环境中，从个人收集的所有数据都用于在中央服务器中进行训练[9-11]，例如深度 Q 学习网络 (deep Q-learning, DQN)[12] 解决了使用强化学习算法求解 Atari 游戏最优决策问题，此后，又有诸如深度确定策略网络梯度 (deep deterministic policy gradient, DDPG)[13]，信赖邻域策略梯度优化 (trust region policy optimization, TRPO)[14]，近似策略梯度优化 (proximal policy optimization, PPO)[15] 等算法都取得了较好的成果，实验证明，深度强化学习可以处理解决复杂的高纬度状态动作映射问题，从而实现更全面感知决策，具有较强的实用性[16-20]。在一些多智能体强化学习 (multi-agent reinforcement learning, MARL) 研究工作中，集中训练和分散执行方案用于训练多智能体系统，例如 counterfactual multi-agent (COMA) [21] 和 multi-agent DDPG(MADDPG) [22]。

其中，DQN 引入了两个重要策略实现了强化学习算法与深度神经网络的融合。第一个策略是目标网络的阶段性更新策略，保证了训练 Q 网络的 Q 值稳定性。另一个策略是经验回放机制，这个机制使得算法可以多次重复利用代理获得的经验，通过重复性采样，提高了经验的利用率，并有效降低了样本数据间的及关联参数，具体训练流程如图 3 所示。

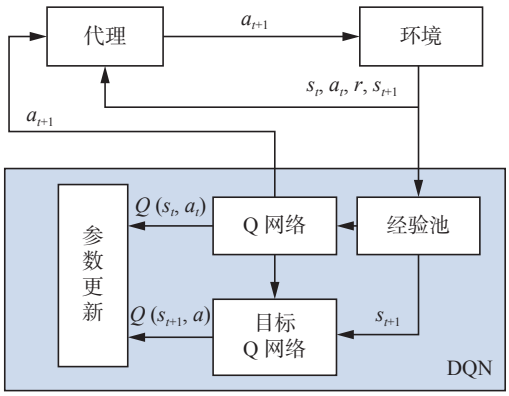


图 3　DQN 算法训练流程

Fig. 3　DQN algorithm training process

DQN 是应用于离散动作空间的算法，这导致无法应对控制精度高、动作空间复杂的问题。而 DDPG 作为解决连续控制型问题的算法适合本文所遇到的难题，其算法流程如图 4 所示。

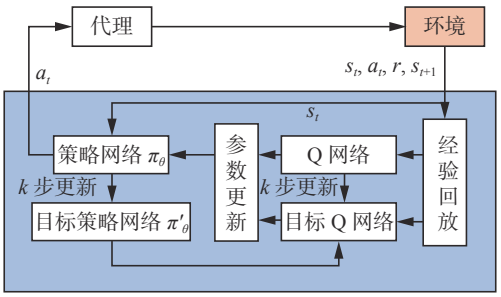


图 4　DDPG 算法训练流程

Fig. 4　DDPG algorithm training process

3　多特征策略梯度优化算法

在现有的深度确定性策略梯度算法的基础上通过对机器人导航任务的任务奖励进行拆分，并构建了各自的 Q 值网络用于优化策略网络，并在此基础上行构建了基于策略梯度优化算法的多特征策略梯度优化算法。

3.1　策略网络优化

针对前述算法中将机器人的导航奖励设计为统一奖励值的算法中存在的奖励显著性下降的问题，本文将机器人的导航问题分为避障任务和导

航任务，分别对 2 个任务进行量化评价并构建避障 Q 值网络和导航 Q 值网络，分别使用 2 个 Q 值网络计算 2 个 Q 值对策略网络参数的梯度，从而实现对策略网络的优化，本文将其称为多特征策略梯度优化算法（multi-featured policy gradients, MFPG）。

MFPG 将机器人的任务奖励分为两部分，分别称为导航奖励和避障奖励，因此在本算法中，算法在t时刻的经验则定义为

$$e_t = \{s_t, a_t, r_t^{\mathrm{Nav}}, r_t^{\mathrm{CA}}, s_{t+1}\}$$

式中：r_t^{Nav}表示t时刻的导航任务奖励，r_t^{CA}表示t时刻的避障任务奖励。由 2 个奖励构建的 Q 值网络分别为

$$\begin{cases} Q_\pi^{\mathrm{Nav}}(s,a) = \mathrm{E}_\pi(G_t(r^{\mathrm{Nav}})|s_t = s, a_t = a) \\ Q_\pi^{\mathrm{CA}}(s,a) = \mathrm{E}_\pi(G_t(r^{\mathrm{CA}})|s_t = s, a_t = a) \end{cases}$$

其中Q_π^{Nav}和Q_π^{CA}分别代表由导航任务奖励和避障任务奖励构建的 Q 值网络，根据式 (1) 计算出 2 个 Q 网络对状态s_t的策略梯度：

$$\begin{cases} \nabla_\varphi J^{\mathrm{Nav}}(e) = \dfrac{1}{m}\displaystyle\sum_{e\in B} \nabla_\varphi \log\pi(a_e|s_e) Q_\pi^{\mathrm{Nav}}(s_e, a_e) \\ \nabla_\varphi J^{\mathrm{CA}}(e) = \dfrac{1}{m}\displaystyle\sum_{e\in B} \nabla_\varphi \log\pi(a_e|s_e) Q_\pi^{\mathrm{CA}}(s_e, a_e) \end{cases}$$

因此在 MFPG 算法中策略网络π_φ的策略梯度是$\nabla_\varphi J = [\nabla_\varphi J^{\mathrm{Nav}}, \nabla_\varphi J^{\mathrm{CA}}]$，因此，最终的策略参数更新公式为

$$\varphi = \varphi + \alpha \cdot \boldsymbol{\Phi}^{\mathrm{T}} \nabla_\varphi J(e) \tag{2}$$

式中$\boldsymbol{\Phi}$是策略梯度权重，表示每个策略梯度分量的重要程度，其值与任务奖励、Q 网络损失值相关。

综上所述，本文所述多特征策略梯度算法的流程图如图 5 所示，从图中可以看出，所提出的多特征策略梯度优化方法通过将对奖励信息进行划分，并分别由划分的两个奖励构建 Q 网络，并在最终构建关于策略网络的优化梯度，实现了对策略网络的优化。

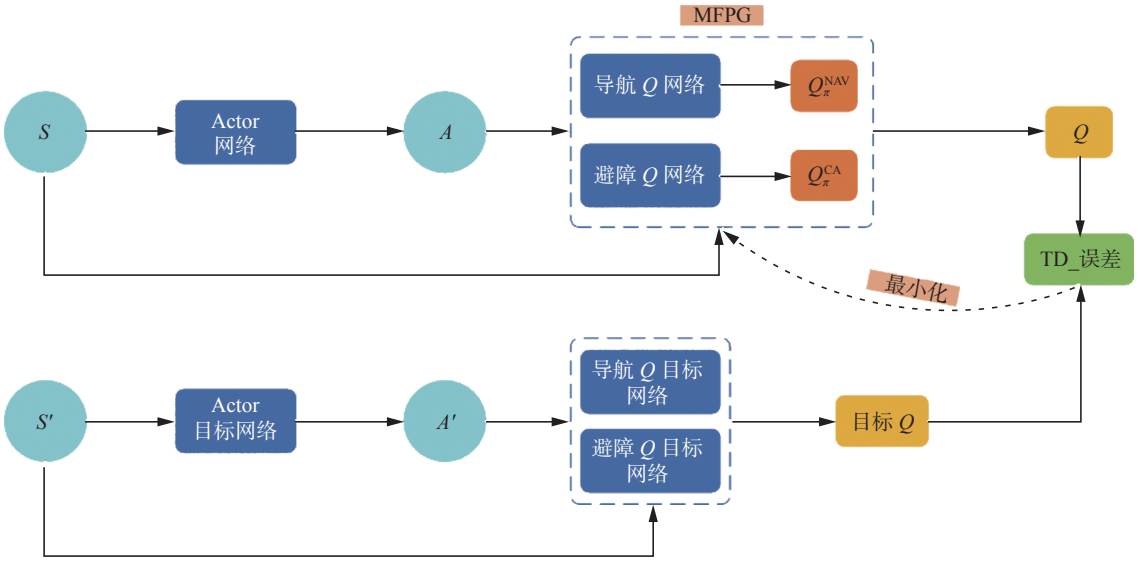


图 5　多特征策略梯度优化算法

Fig. 5　MFPG algorithm

3.2　社会范式的奖惩函数设计

借鉴人类社会发展过程中产生的行动规则（例如右侧通行等），引入了社会范式奖励，其具体方式如图 6 所示：当受控机器人（红）与其他机器人（黑）进行交互且产生图示的位置关系时候受控机器人会受到负奖励，从而降低出现图示位置关系的概率。然而，这种方法只是在图示状态下对机器人赋予了一个离散的负奖励信息，而且由于负奖励的判断范围较广（阴影所示区域），导致负奖励信息只能用于定性分析受控机器人状态，而不能用于提高算法的控制精度，且由于算法本身奖励稀疏，导致算法更无法学习在图示条件下的社会范式，因此将离散化的指标性奖励精确为基于实时状态的奖励可以有效提高算法的训练速度。

图 6　离散的社会范式奖励

Fig. 6　Discrete social paradigm rewards

综上，本研究在前文研究的基础上，针对现有导航算法中提出的离散式社会规范奖励存在的奖励稀疏、离散的社会负奖励信息只能定性分析机

器人的社会范式状态的问题提出了一种新的基于激光雷达信息的连续空间社会范式奖励计算方法，其计算公式为

$$r_l = [r_{\text{laser}}(l_{\min}) \cdot G(z)]_{-9.9}^{9.9}$$

上式表示$r_{\text{laser}}(l_{\min}) \cdot G(z)$的值最终在[−9.9,9.9]的边界区间内，其中$r_{\text{laser}}(l_{\min})$表示当前激光雷达探测区域最小值的计算奖励，$l_{\min} = l_{\min}/l_{\text{MAX}}$表示激光雷达的最小探测值的正则化值，$l_{\text{MAX}}$表示激光雷达的最大侦测范围，$G(z)$表示激光雷达最小值所在方位引起的奖励偏置因子，其中z表示激光雷达传感器探测到最短激光值所在位置的正则化序号值，其具体表述及序号关系详述于图 1。上述两值的计算公式为

$$\begin{cases} r_{\text{laser}}(l_{\min}) = -e^{k_l(l_{\min}-o_l)} \\ G(z) = U - D \times (1-z) \end{cases}$$

式中：k_l表示避障增益，o_l表示避障奖励的偏移量，在本文中两者分别为 20 和 0.5，该奖励值随$l_{\min}$的变化如图 7 所示。

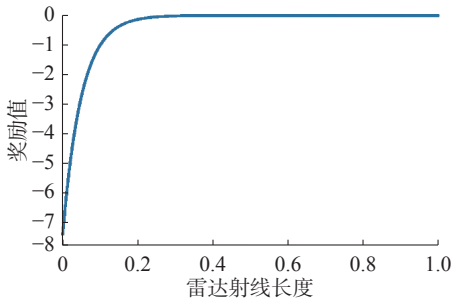


图 7　避障奖励值变化

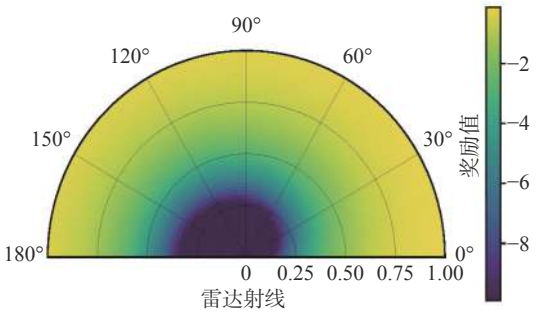
Fig. 7　Changes of obstacle avoidance bonus value

此外，U和D分别表示偏置上限和偏置零点的位置，在本文中U和D分别为 3 和 2.5。因而，避障奖励函数的最终计算公式为

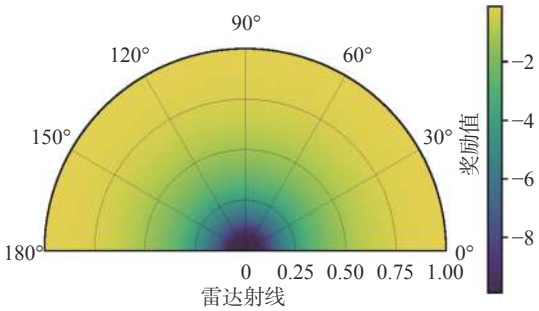
$$r_t^{\text{CA}} = \begin{cases} -10, & l_{\min} \leqslant 0.1 \\ r_1, & l_{\min} > 0.1 \end{cases}$$

将上式标绘在平面直角坐标系中得到图 8(a)，同时作为对比，图 8(b) 也标绘了$G(z) = 1$时的奖励分布。

从图 8(a) 的奖励状态分布可知，如果距离机器人最近的障碍物位于机器人两侧时，无偏置的奖励算法将输出同样的奖励结果，这导致了两辆车辆在相遇时，无法准确对对方的形为进行预测并进行有效规避，从而导致发生碰撞，而带有偏置的奖励计算方法可以对机器人左右两侧的信息进行有效区分，从而保证强化学习算法在训练过程中对于左侧和右侧的障碍物表现出明显的倾向性，因此可以保证车辆在相遇时会根据自身预设的策略倾向实现在无通讯信息条件下多机器人间的安全导航避障。



(a) 带有偏置奖励参数的奖励分布



(b) 无偏置奖励参数的奖励分布

图 8　奖励分布

Fig. 8　Bonus distribution

3.3　经验优先采样机制

在经验回放过程中，经验的选择会影响 Q 网络的收敛速度，进而影响策略网络的训练。而传统的均匀采样不能显著提高 Q 网络的训练速度，因此，本文采用了基于 Q 网络损失值的经验优先采样算法并进行改进，其核心在于根据 Q 值网络的损失值构建每一条经验的采样优先性，其主要流程为：对于每条经验e_i及其 Q 值网络的损失$L_\theta(e_i)$，定义其采样优先度为

$$P(e_i) = \frac{L_\theta(e_i)^\varepsilon}{\sum_e L_\theta(e)^\varepsilon}$$

式中ε表示采样优先度指数，当$\varepsilon = 0$时代表算法采用均匀采样方法。在经验采样时，算法按照概率$P(e_i)$随机选择经验组成训练经验组。因此，在实际训练时，每条经验e_i被采样的概率正比于其损失函数$L_\theta(e_i)$，因此可以显著提高 Q 网络的收敛速度。

此外，由于 Q 值网络的更新会改变 Q 值网络的分布，从而改变经验e_i的 Q 值期望，因此基于优先级的经验回放算法会引入偏差，需要对优先采样获得的经验添加重要性修正权重以降低偏差，其计算公式为

$$\omega(e_i) = \left(\frac{1}{|B| \cdot P(e_i)}\right)^\varsigma$$

式中：$|B|$表示经验样本集合的容量，ς表示算法的修正权重，因此，修正后的策略网络的参数更新公式为

$$\varphi \leftarrow \varphi + \alpha\omega(e)L_\theta(e)\nabla_\varphi J(e)$$

而在本研究中, 由于Q值损失函数$[L_\theta^{\mathrm{CA}}(e_i), L_\theta^{\mathrm{Nav}}(e_i)]$不为单一值, 因此在本文中, 需要对经验的采样优先级算法进行修改, 本文使用了线性加权法修改了经验采样优先级, 因此经验采样优先级的算法更新为

$$P(e_i) = \frac{\eta_1 L_\theta^{\mathrm{CA}}(e_i) + [\eta_2 L_\theta^{\mathrm{Nav}}(e_i)]^\varepsilon}{\sum_e \left[\eta_1 L_\theta^{\mathrm{CA}}(e_i) + \eta_2 L_\theta^{\mathrm{Nav}}(e_i)\right]}$$

当新的经验e_{new}被加入到经验池中时, 会替代当前经验池中采样优先度最小的经验, 其采样优先级会被设置为1。

综上, 多特征策略梯度算法的主要流程如下:

1) 初始化策略网络, 2个Q值网络$[Q_\pi^{\mathrm{Nav}}, Q_\pi^{\mathrm{CA}}]$;

2) 创建目标策略网络和目标Q值网络π_ϕ^*, $[Q_{\pi^*}^{\mathrm{Nav}}, Q_{\pi^*}^{\mathrm{CA}}]$, 其参数来自$\pi_\phi$和$[Q_\pi^{\mathrm{Nav}}, Q_\pi^{\mathrm{CA}}]$;

3) 初始化经验池N, 其最大容量为n;

4) 当不满足终止条件时, 获取当前代理状态s_t;

5) 根据当前状态选择动作$a_t = \pi_\phi(s_t)$, 并由代理执行;

6) 获取下一时刻状态s_{t+1}与奖励$r_t^{\mathrm{Nav}}, r_t^{\mathrm{CA}}$;

7) 将当前经验$e_t = \{s_t, a_t, r_t^{\mathrm{Nav}}, r_t^{\mathrm{CA}}, s_{t+1}\}$加入经验池;

8) 如果$|N| \geqslant n$, 根据Q值网络更新方式从$|N|$采样训练经验集合$|B|$, 更新Q值网络Q_π^{Nav}和Q_π^{CA}, 按照公式(2)更新策略网络参数;

9) 如果达到策略网络更新条件, $\phi* \leftarrow \phi$;

10) 如果达到Q值网络更新条件, $\theta^{\mathrm{Nav}*} \leftarrow \theta^{\mathrm{Nav}}$, $\theta^{\mathrm{CA}*} \leftarrow \theta^{\mathrm{CA}}$;

11) 返回更新后策略网络π_ϕ。

4 单机器人实验及结果分析

4.1 基于ROS的移动机器人控制架构

本文构建了基于ROS的多机器人控制系统以供强化学习算法进行机器人路径导航训练学习, 每个机器人与ROS主机的信息构图框架图如图9所示。

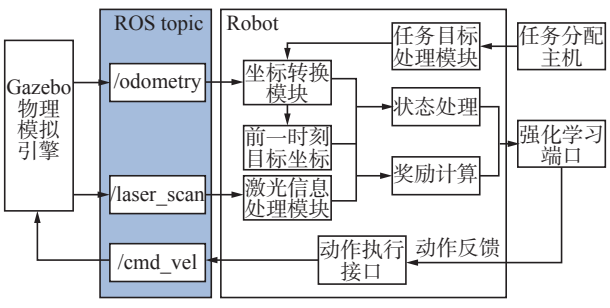


图 9　基于ROS的单机器人控制系统

Fig. 9　Single robot control system based on ROS

本研究使用Gazebo物理仿真环境作为机器人在实际环境中的模拟环境, 并使用了Turtlebot3作为虚拟实验机器人, 其装载有一个激光雷达扫描仪, 其探测距离为3.5 m, 激光雷达的采样率为128 Hz, 采样范围为180°。

4.2 实验环境

本文所述ROS系统基于Ubuntu18.04, 使用虚拟机器人为Turtlebot3 waffle, 在Gazebo中构建避障模拟环境如图10所示, 本实验中$\xi = 20$。图中, 绿色圆形区域为机器人的目标点区域, 红色区域则表示障碍物, 实际执行时, 机器人将在一个非障碍物且非目标点的区域随机初始化, 并在3个目标点中随机选择一个作为任务目标。

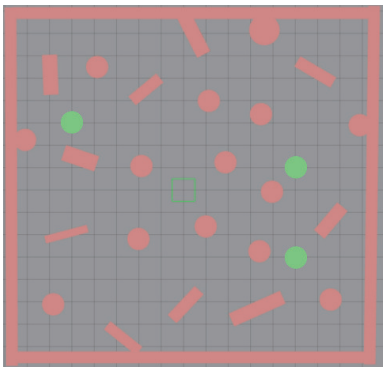


图 10　机器人导航实验环境

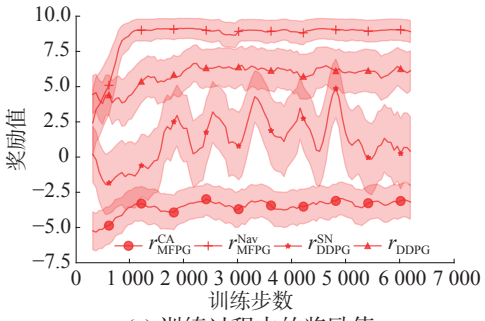
Fig. 10　Robot navigation experiment environment

4.3 实验结果

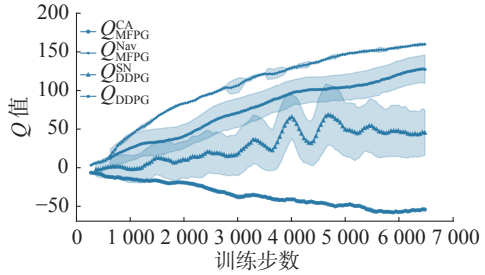
通过训练, 获得了基于MFPG算法的机器人导航模型, 在训练过程中, 与传统DDPG算法以及加入奖励偏执的DDPG算法相比, 机器人每步动作获得的奖励值及对应Q值随训练步长的变化如图11所示。

3种算法在训练过程中均具有收敛特征, 但相比较而言, 带有偏置奖励的DDPG算法稳定性较差, 训练过程中奖励值和Q值均出现大幅度波动, 且方差区间较高, 而标准DDPG算法则表现出相对稳定的特征, 并具有较为稳定的方差区间, Q值和奖励在训练过程中变化稳定, 表明算法可以较好的应对多机器人避障问题, 但仍存在训练过程中训练速度较慢等问题, 而在本文所述多目标策略梯度优化算法中, 从导航和避障奖励中可以看出, 算法在训练早期(1 000步)时已经可以稳定获得较高奖励且在后续训练过程中仍能保持稳定, 且方差较小, 同时Q值网络方差较小, 且数值变化稳定, 导航方面Q值的增长速度明显优于标准DDPG算法, 说明本文所述多目标策略梯度优化算法可以较好的完成多机器人避障任

务。经过多次测试，机器人导航避障成功率约为95.3%。



(a) 训练过程中的奖励值



(b) 训练过程中的 Q 值

图 11　模型训练

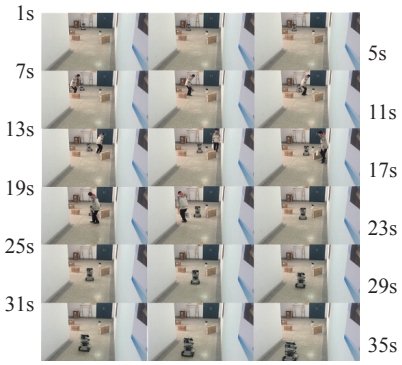
Fig. 11　Model training

4.4　现实环境实验测试

本节将前述小节中的机器人避障算法模型移植进入基于 ROS 的移动机器人平台中，并在实验场地分别进行静态寻路测试和动态障碍物避障测试，最终得到结果如图 12 所示，展示了 2 个导航过程的行进示意和轨迹图。

在图 12(a) 和 (b) 中可以观察到机器人从地图左侧中间出发向右侧的通道行进，其行进图如图所示。途中经过 3 个障碍物，并两次与动态障碍物相遇，从行进图中可以发现，第 1~5 s，机器人发现右前方障碍物，并向着第 1 个和第 2 个障碍物之间的空间前进，第 7 s 出现移动障碍物，机器人判断形势，在 11 s 向左侧规避，但在 13 s 发现移动障碍物已经快速达到左侧，因此在 13~15 s 恢复正常航向，向右前侧出发，并调整姿态，避让第 3 个障碍物，此时移动障碍物从右侧出现，并在 21 s 快速移动到机器人右侧，此时机器人已经经过第 3 个障碍物，向左前侧前进（21 s），并在 23~35 s 正常导航直至抵达任务规定目标点。

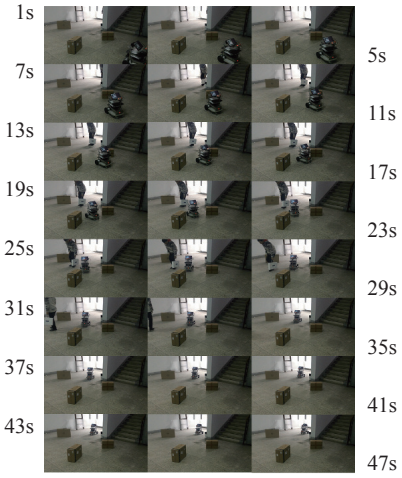
在图 12(c) 和 (d) 中机器人从地图右侧中间出发，向其左前侧的区域前进，途中机器人经过 3 个静态障碍物并与动态障碍物进行一次相遇，图中，1~9 s 机器人探测到前方的两个障碍物，并选择从二者中间的空间经过，11s 开始，动态障碍物开始出现，19~23 s 机器人探测到移动障碍物，并根据其位置调整其航向向右前方行驶 (25~29 s)，从而成功躲避了移动障碍物，31 s 至 47 s 机器人按照自身信息及位置导航至目标点。



(a) 单机器人动态障碍物避障测试 1

(b) 单机器人的静态寻路测试轨迹 1

(c) 单机器人的静态寻路测试轨迹 2



(d) 单机器人动态障碍物避障测试 2

图 12　单机器人静态测试结果

Fig. 12　Single robot static test results

从图中可以看出，移动机器可以准确完成机器人控制以及寻路任务，且对于移动障碍物具有明显的避障及寻路特征，说明训练所得算法在移

动机器人控制中可以执行有障碍物状态下的寻路和动态避障任务。

5　分布式异构多机器人实验

本节将在先前基础上建立多机器人系统的控制流程, 搭建用于多机器人导航的虚拟环境并进行训练, 最后使用多台移动机器人平台进行寻路实验。

5.1　分布式多机器人控制结构

由于训练过程中机器人需要在环境中进行初期探索并积累经验池的经验, 使用多机器人系统在训练早期可以有效提高经验搜集效率, 缩短训练模型学习时间, 从而提高算法训练效率, 因此本文在前述单机器人导航任务实验的基础上, 构建了多机器人训练的流程模型, 其流程图可见图 13。

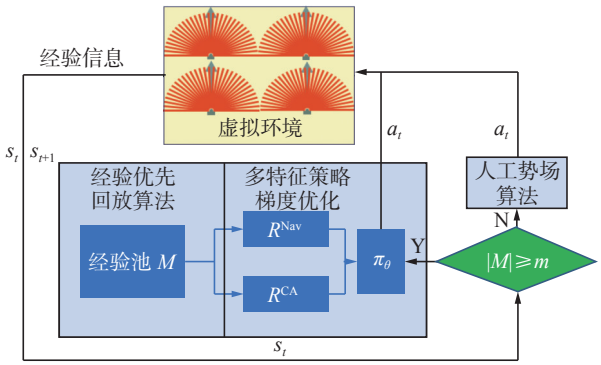


图 13　多机器人训练流程

Fig. 13　Multi-robot training process

1) 算法获取虚拟环境中机器人的信息(激光雷达传感器数值, 坐标位置信息), 并进行预处理;

2) 将上述预处理后得到的状态信息和动作信息和奖励信息一起添加进经验回放池, 留待训练;

3) 判断当前经验池的状态, 如果没有达到经验回放池上限, 则使用人工势场算法, 根据机器人的状态信息进行判断, 并提出决策信息。

4) 如果经验回放池达到上限, 则开始强化学习算法的网络训练, 神经网络会替代人工势场算法作为多机器人系统的控制算法, 同时搜集新的经验信息替换已有的经验;

5) 重复 3)~4), 直到达到终止条件。

其中为了提高早期经验搜集效率, 增加经验池中优质经验比例, 提高算法学习效率, 在经验池达到上限之前, 在环境中的各个机器人采用人工势场法进行导航, 当经验池达到上限时, MFPG 开始根据改进的人工势场算法对经验池进行采样, 并进行学习, 逐渐提高算法的决策能力。

5.2　控制信号解码

由于 Turtlebot3 与远程主机使用 WIFI 网络连接, 实际执行过程中容易受到无线信号影响, 因此, 需要对两个 Turtlebot3 的动作信息进行处理以保证动作信息在通信延迟较高的环境下仍然可以正常运行。

为了保证机器人的动作信息能够正常执行导航算法, 将应用在机器人上的策略数值进行了缩放, 保证机器人能正常执行动作策略, 其具体计算公式为

$$\hat{a}_t = \frac{2}{3\pi}\arctan(16\pi(s_t))$$

式中: $\hat{a}_t$表示缩放后的策略数值, 此后提出了实体机器人控制时的策略融合方法, 其公式为

$$\boldsymbol{a}_f = [\hat{a}_t\ \hat{a}_{t-1}\ \hat{a}_{t-2}]\cdot[0.625I_t\ 0.25I_{t-1}\ 0.125I_{t-2}]^{\mathrm{T}}$$

式中: I_t表示t时刻的状态补偿参数, 用来降低由于通信阻塞情况导致的策略误差, 其定义为

$$I_t=\begin{cases}1, & a_t\cdot\theta\geqslant 0\\ 0.7, & \text{其他}\end{cases}$$

执行任务时, 为了保证算法能同时开始执行导航动作, 在 Linux 系统中搭建了一个本地 NTP 服务器用于保证 Turtlebot3、远程主机和板载主机之间的时间统一, 实际执行时, 板载主机通过 WIFI 连接向远程主机申请进行时间校验, 并根据返回时间对自身信息进行校验, 此后两个主机将选择同一时间执行导航任务。

5.3　模拟厂区的导航及避障测试及分析

5.3.1　训练环境

本节在前述机器人控制架构的基础上构建了模拟厂区环境, 其环境如图 14 所示。

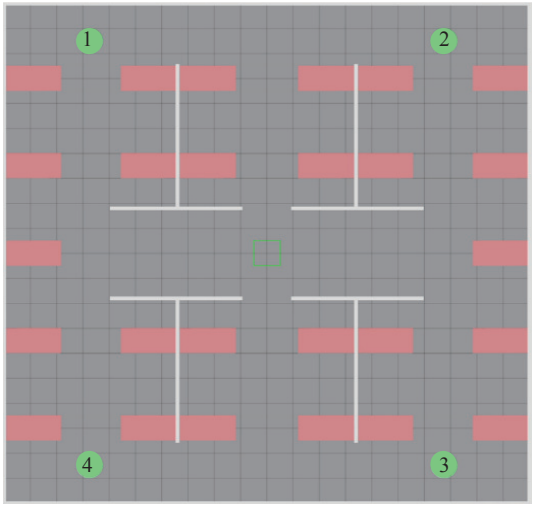


图 14　训练环境

Fig. 14　Training environment

我们选用 4 个移动机器人进行算法的训练以及测试。在训练过程中，每个机器人随机选择图中绿色圆形区域和绿色方框区作为算法的任务初始点，并随机选择除该机器人初始点以外的其他任务初始点作为导航目标。机器人通过算法输出的运动指令向目标区域进行移动，完成各自的导航任务。在算法的性能测试过程中，车辆被随机分配至四个圆形区域，而其导航目标位置为该车的起始点在模拟工厂环境的对角位置，即对于为 1 的机器人，其目标位置设置在 3。

5.3.2　训练结果

经过 200 轮训练后，对所得模型进行效果较好者进行性能测试对比，得到两算法在模拟厂区环境下的性能测试结果如表 1 所示。在本实验中，实验参数为总训练回合数：400，单回合最多执 250 步，ROS 系统运行频率 45 Hz，$\xi = 20$。

表 1　算法性能对比

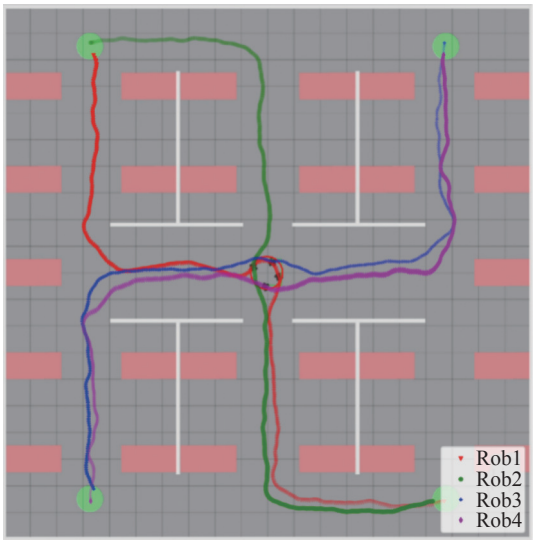
Table 1　Comparison of algorithm performance

算法	成功率	平均碰撞距离		最小距离圆面积	
		均值	方差	均值	方差
标准DDPG	73.3%	4.52	6.45	60.76	34.4
MFPG	80.5%	2.84	5.27	5.5	12.23

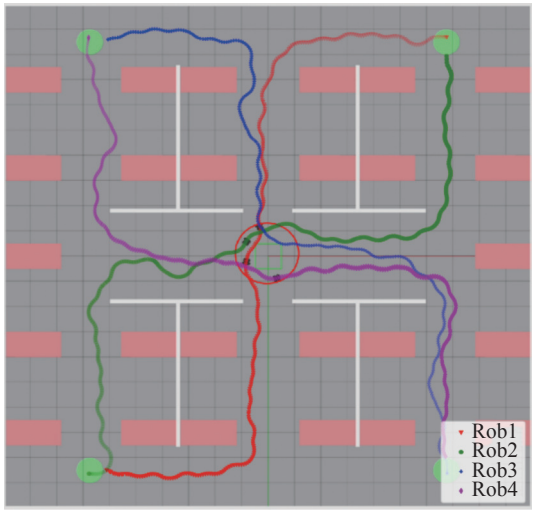
在测试中，采用了平均碰撞距离作为算法导航性能的衡量标准，其定义为：机器人发生碰撞时，距离目标点的直线距离，较小的平均碰撞距离反映算法具有较好的导航能力。从表中可以看出，相比较标准 DDPG 算法，本文所述多特征策略优化算法具有更高的成功率，并且平均碰撞距离和最小距离圆面积的均值和方差也是小于标准 DDPG 算法，说明所提出的基于多特征策略梯度优化方法的多机器人导航具有较高的可靠性。

将上述 2 种算法在模拟厂区环境中的路径轨迹标绘在图中可以得到图 15。

从图中可以看出，多特征策略优化算法控制下的多机器人轨迹相对平滑，具有较小的波动，且多机器人在相遇时具有较为规范的避让动作，保证了多机器人系统在执行过程中的控制稳定性，而标准 DDPG 算法在导航过程中则具有较多的控制波动，说明算法在执行过程中存在决策稳定性差的情况，且在多个机器人相遇时，机器人轨迹波动严重，这说明算法在多机器人相遇时处理动态障碍物能力较差，这对多机器人系统而言是致命的，而统计结果也表明了所提出的多特征策略梯度优化方法在多机器人系统中的动态避障和导航问题中相对标准 DDPG 算法具有显著优势。



(a) 特征策略优化算法轨迹



(b) 标准 DDPG 算法轨迹

图 15　路径轨迹

Fig. 15　Path trajectory

此外，为保证本算法所训练模型不受机器人数目的影响，使用在前述算法所训练的模型在 8 个机器人的模拟工厂环境中进行了寻路测试，由于实验机器人数目增加，为保证每个机器人的控制频率符合模型的控制频率，本次实验中 ROS 系统的控制频率为 90 Hz，获得寻路轨迹如图 16 所示。

从图中可以看出，多个机器人在执行任务过程中在车流量较高的中心区域可以正常通行，且均遵循右手通行原则，保证了多机器人系统运行时的安全性，测试结果表明所提出的算法模型不受多机器人系统中机器人数目的影响，因而提高了该方法在多机器人系统中的应用范围。

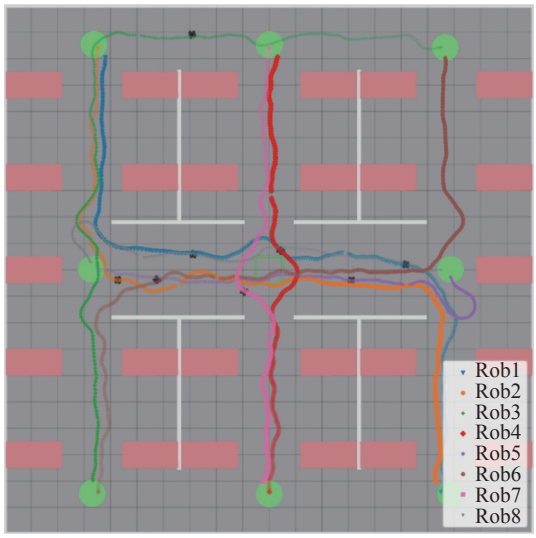


图 16　8 机器人寻路路径轨迹
Fig. 16　8 Robot path trajectory

5.4　实体机器人实验

本次测试使用的控制主机与前述相同, 场地参数$\xi=5$, 运行速度 0.4 m/s。机器人的抵达目标的感知半径为 0.5 m, 本次实验共计使用了 2 台远程主机和 1 台板载主机, 3 台主机均连接至同一无线网络, 2 台 Turtlebot3 机器人具有自己的 IP 地址, 并由 2 台远程主机控制, 为保证能够对 2 台 Turtlebot3 进行区分, 也方便板载机器人获取其位置信息, 在机器人顶端分别安装了红色和黄色的标识盒。

在行进图 17(a) 中黄色小车从下方开始从左侧运动并在达到边界后向右上侧运动, 最终抵达目标点, 而红色小车则在 13 s 与板载主机机器人相遇, 而后红色机器人感知到板载机器人位置, 并执行避让动作, 并在红色机器人与板载机器人之间出现可以通行的空间时, 从板载主机机器人侧后方向前行驶, 避免了与板载主机机器人的碰撞, 二者在最后分别抵达目标点。

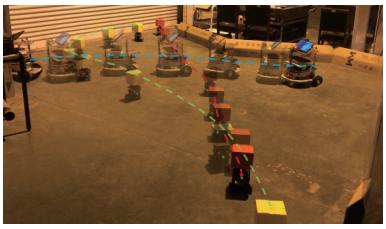
而在行进图 17(b) 中, 黄色机器人运行时感知到板载主机机器人位置, 选择从右侧行进, 但在运行过程中判断与板载主机机器人发生碰撞的可能行较高, 因此在 33 s 时选择机器人板载机器人侧后方行进, 并在脱离板载主机机器人碰撞范围后向目标点行进, 最终抵达目标, 而板载主机机器人则探测到其左侧存在红色机器人并向其运动方向前方运行, 因此选择向红色机器人后侧转弯, 红色机器人则选择向左前方运行, 以避让板载主机机器人, 而板载主机机器人在红色机器人离开后向目标点导航, 最终二者顺利抵达目标点。

(a) 多机器人实验车辆测试 1

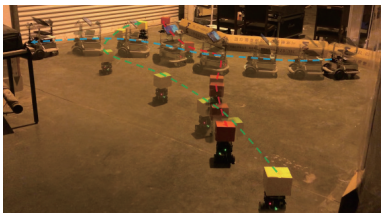
(b) 多机器人实验车辆测试 2

图 17　多机器人动态避障导航实验
Fig. 17　Dynamic obstacle avoidance navigation experiment of multiple robots

其具体轨迹如图 18 所示, 轨迹图 18(a) 和 18(b) 中红色机器人从图中上方向下方移动, 而黄色机器人则从下方向上方移动, 而轨迹图 1 中板载机器人则从左侧移动至右侧, 而在 18(b) 中则从右侧移动至左侧, 图中蓝色虚线表示板载主机机器人的路径轨迹, 绿色虚线表示黄色机器人的路径轨迹, 红色虚线则表示红色机器人的路径轨迹信息。



(a) 多移动机器人实机测试轨迹 1



(b) 多移动机器人实机测试轨迹 2

图 18　机器人轨迹
Fig. 18　Robot trajectory

实验结果显示, 3 辆移动机器人平台在实验过程中可以感知到其他机器人的移动位置并执行相应的回避动作, 最终安全抵达实验预定的目标

地点。结果证明，本文所述的基于强化学习的多机器人导航算法可以应用于多移动机器人平台在不依靠精确地图环境下的导航及动态避障。

6 结束语

本文针对现有多机器人系统中依赖精确地图的人机协同场景下存在的问题提出了基于深度强化学习算法的多特征策略优化算法作为机器人导航模型中的决策中心，搭建了基于 ROS 的强化学习虚拟训练环境，并将所得动态避障导航算法模型应用在实际机器人中。实验结果表明，本文所提出的导航算法可以应用于多移动机器人平台在不依靠精确地图环境下的导航及动态避障。但算法的训练时间较长，且在训练后期稳定性欠佳是该任务中尚存的问题，应当注意在未来工作中对于强化学习算法稳定性的分析和改进，保证强化学习算法的收敛稳定性。

参考文献：

[1] SHI Huiyuan, SU Chengli, CAO Jiangtao, et al. Nonlinear adaptive predictive functional control based on the Takagi-sugeno model for average cracking outlet temperature of the ethylene cracking furnace[J]. Industrial & engineering chemistry research, 2015, 54(6): 1849–1860.

[2] MELLINGER D, KUSHLEYEV A, KUMAR V. Mixed-integer quadratic program trajectory generation for heterogeneous quadrotor teams[C]//2012 IEEE International Conference on Robotics and Automation. Saint Paul: IEEE, 2012: 477−483.

[3] KHATIB O. Real-time obstacle avoidance for manipulators and mobile robots[M]//Autonomous robot vehicles. New York: Springer New York, 1986: 396−404.

[4] ZHANG Pengpeng, WEI Changyun, CAI Boliang, et al. Mapless navigation for autonomous robots: a deep reinforcement learning approach[C]//2019 Chinese Automation Congress. Hangzhou: IEEE, 2019: 3141−3146.

[5] CHEN Yufan, LIU Miao, EVERETT M, et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]//2017 IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017: 285−292.

[6] TAI Lei, PAOLO G, LIU Ming. 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. Vancouver: IEEE, 2017: 31−36.

[7] MINSKY M. Theory of neural-analog reinforcement systems and its application to the brain-model problem[M]. New Jersey: Princeton University, 1954.

[8] BELLMAN R. Dynamic programming[J]. Science, 1966, 153(3731): 34–37.

[9] FAN Tingxiang, LONG Pinxin, LIU Wenxi, et al. Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios[J]. The international journal of robotics research, 2020, 39(7): 856–892.

[10] BARTH-MARON G, HOFFMAN M W, BUDDEN D, et al. Distributed distributional deterministic policy gradients[EB/OL]. New York: arXiv, 2018: (2018−04−23)[2021−06−25].https://arxiv.org/abs/1804.08617.

[11] NA S, NIU Hanlin, LENNOX B, et al. Universal artificial pheromone framework with deep reinforcement learning for robotic systems[C]//2021 6th International Conference on Control and Robotics Engineering. Beijing: IEEE, 2021: 28−32.

[12] HUANG Liang, BI Suzhi, ZHANG Y J A. Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks[J]. IEEE transactions on mobile computing, 2020, 19(11): 2581–2593.

[13] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[C]//Proceedings of the 32nd International conference on machine learning. New York: PMLR, 2015: 1889−1897.

[14] WANG Yuhui, HE Hao, TAN Xiaoyang. Truly proximal policy optimization[C]// Proceedings of the 35th Uncertainty in Artificial Intelligence Conference. New York: PMLR, 2020: 113−122.

[15] 赵冬斌, 邵坤, 朱圆恒, 等. 深度强化学习综述: 兼论计算机围棋的发展 [J]. 控制理论与应用, 2016, 33(6): 701–717.
ZHAO Dongbin, SHAO Kun, ZHU Yuanheng, et al. Review of deep reinforcement learning and discussions on the development of computer go[J]. Control theory & applications, 2016, 33(6): 701–717.

[16] AGOSTINELLI F, HOCQUET G, SINGH S, et al. From reinforcement learning to deep reinforcement learning: an overview[M]//Braverman readings in machine learning. Key ideas from inception to current state. Cham: Springer, 2018: 298−328.

[17] NIELSEN M A. Neural networks and deep learning[M]. San Francisco: Determination press, 2015.

[18] HU Junyan, NIU Hanlin, CARRASCO J, et al. Voronoi-

based multi-robot autonomous exploration in unknown environments via deep reinforcement learning[J]. IEEE transactions on vehicular technology, 2020, 69(12): 14413–14423.

[19] CHRISTIANOS F, SCHÄFER L, ALBRECHT S V. Shared experience actor-critic for multi-agent reinforcement learning[J]. Advances in neural information processing systems, 2020, 33: 10707–10717.

[20] GAO Junli, YE Weijie, GUO Jing, et al. Deep reinforcement learning for indoor mobile robot path planning[J]. Sensors, 2020, 20(19): 5493.

[21] JAKOB Foerster, GREGORY Farquhar, TRIAN T AFYLLOS Afouras, et al. Counterfactual multi-agent policy gradients[C]//Proceedings of the AAAI conference on artificial intelligence. New Orleans: PKP, 2018, 32(1).

[22] LOWE R, WU Yi, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[EB/OL]. New York: arXiv, 2017. (2017−06−07) [2021−06−25].https://arxiv.org/abs/1706.02275.

作者简介：

欧阳勇平，硕士研究生，主要研究方向为智能自主无人系统。

魏长赟，副教授，博士，荷兰代尔夫特理工大学人工智能专业博士，英国卡迪夫大学机器人及自主系统实验室访问学者，主要研究方向是智能自主无人系统。发表学术论文 30 余篇。

蔡帛良，英国卡迪夫大学博士，主要研究方向为多机器人协作、智能无人系统。

第七届交互协作机器人国际会议
7th International Conference on Interactive Collaborative Robotics

交互协作机器人国际会议始于 2016 年，原是国际语音与计算机会议（SPECOM）的卫星会议之一。会议聚集了来自不同领域的专家学者，共同探讨人机协作在工业、社会、医疗、教育等不同方面的挑战。本次会议的主要议题是在不确定性和环境变异性条件下，一个或多个机器人在配备嵌入式传感器网络和云服务的操作环境中与人进行物理交互的协作行为的基础和手段。

重要日期：

论文截止日期：2022 年 9 月 1 日

录用通知时间：2022 年 9 月 30 日

参会登记时间：2022 年 10 月 10 日

会议日期：2022 年 10 月 21-23 日

详情请关注：

第七届交互协作机器人国际会议官网：https://icr2022.gaitech.net/