控制与决策
Control and Decision

移动机器人运动规划中的深度强化学习方法

孙辉辉， 胡春鹤， 张军国

引用本文:
孙辉辉, 胡春鹤, 张军国. 移动机器人运动规划中的深度强化学习方法[J]. 控制与决策, 2021, 36(6): 1281−1292.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0470

移动机器人运动规划中的深度强化学习方法

孙辉辉[1,2], 胡春鹤[1†], 张军国[1]

(1. 北京林业大学 工学院，北京 100083；2. 华北科技学院 机电工程学院，河北 廊坊 065201)

摘 要: 随着移动机器人作业环境复杂度的提高、随机性的增强、信息量的减少,移动机器人的运动规划能力受到了严峻的挑战. 研究移动机器人高效自主的运动规划理论与方法,使其在长期任务中始终保持良好的复杂环境适应能力,对保障工作安全和提升任务效率具有重要意义. 对此,从移动机器人运动规划典型应用出发,重点综述了更加适应于机器人动态复杂环境的运动规划方法 —— 深度强化学习方法. 分别从基于价值、基于策略和基于行动者-评论家三类强化学习运动规划方法入手,深入分析深度强化学习规划方法的特点和实际应用场景,对比了它们的优势和不足. 进而对此类算法的改进和优化方向进行分类归纳,提出了目前深度强化学习运动规划方法所面临的挑战和亟待解决的问题,并展望了未来的发展方向,为机器人智能化的发展提供参考.

关键词: 移动机器人；运动规划；强化学习；深度强化学习

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0470 **开放科学(资源服务)标识码(OSID):**

引用格式: 孙辉辉, 胡春鹤, 张军国. 移动机器人运动规划中的深度强化学习方法 [J]. 控制与决策, 2021, 36(6): 1281-1292.

Deep reinforcement learning for motion planning of mobile robots

SUN Hui-hui[1,2], *HU Chun-he*[1†], *ZHANG Jun-guo*[1]

(1. School of Technology，Beijing Forestry University，Beijing 100083，China；2. School of Mechanical and Electrical Engineering，North China Institute of Science and Technology，Langfang 065201，China)

Abstract: The motion planning ability of mobile robots are facing a severe challenge with complex environment and less prior information. It is important to study the motion planning method and theory for a mobile robot, so that the mobile robot could adapt to complex environment in a long-running and ensure the work security and task efficiency. This paper mainly summarizes the method based on deep reinforcement learning (DRL) which can deal with the dynamic and complicated obstacles better. The DRL methods, which are based on value, policy and actor-critic, are introduced respectively. Then, the typical robot application in simulation environment and complex real world environment are analyzed based on DRL. After comparing the advantages and disadvantages in detail, the improvement and optimization direction for the DRL method are classified, and the challenges faced by motion planning method are put forward respectively. Finally, the prospects in the field of mobile robot motion planning method with DRL are discussed, which will provide reference for the development of intelligent robots.

Keywords: mobile robot；motion planning；reinforcement learning；deep reinforcement learning

0 引 言

在现代科技飞速发展的今天,智能移动机器人以其小巧灵活、操纵简单、功能多样等特点[1],始终处于科学研究的前沿,一直引领着高新技术发展的重要方向[2]. 它们通过搭载各类传感设备来代替人类执行繁杂而危险的任务[3],在城市救援、生命探测、安全巡防等方面发挥着举足轻重的作用,同时也被广泛应用于工业、农林、医疗教育等行业. 随着人工智能和计算机大数据时代的到来,人类总是期望移动机器人能够具有更加强大的自主化能力,以代替人类在更多的领域完成更加复杂危险的探索操作任务. 为了实现这一目标,核心要求之一就是需要移动机器人必须具备优良的运动规划能力,使机器人在无人干预的条件下也可以在未知环境中有目的地、准确高效地完成任务[4].

目前,被广泛应用的运动规划算法主要有基于环

收稿日期: 2020-04-24；**修回日期:** 2020-11-08.

基金项目: 国家自然科学基金青年科学基金项目(61703047)；中央高校基本科研业务费专项资金项目(2016ZCQ08).

责任编委: 方勇纯.

†通讯作者. E-mail: huchunhe@bjfu.edu.cn.

境模型的A*算法[5]和D*算法[6]、基于搜索的随机路径图法(PRM)[7]和快速探索随机树法(RRT)[8]、基于策略的模糊逻辑法[9]和动态窗口法[10],以及基于仿生规划算法的遗传算法[11]、蚁群算法[12]和蜂群算法[13]等. 一般地,在地图已知、障碍静态、环境简单的条件下,这些运动规划算法可以通过环境建模或者概率搜索等方式来完成简单的任务[14]. 但是,传统规划方法多为定制型算法,还存在程序体积庞大、通用性差、功耗高等诸多难题. 设计具有自主决策能力的智能化机器人运动规划方法,进而弥补传统运动规划方法的缺陷,提高移动机器人运动规划方法的鲁棒性和泛化能力,是移动机器人目前亟待解决的问题之一.

近年来,借助于强化学习的快速发展,强化学习技术以强大的学习能力迅速应用于机器人领域,成为了研究者关注的热点,为移动机器人复杂环境中运动规划问题提供了新的思路和方向[15]. 基于强化学习的运动规划方法可以将任务环境的状态空间与自身运动参数相关联,通过与环境的持续交互进行试错迭代以获取奖励或惩罚,从而优化运动策略[16]. 另外,强化学习不需依赖环境模型以及任何先验知识,仅需通过自主学习和试错训练就可以完成策略的升级,对解决移动机器人在非结构环境中的路径规划,提高移动机器人未知环境的自适应性[17]的问题有着重要作用. 众多研究机构和知名大学[18-19]均对强化学习的运动规划方法投入了大量精力,并且已经取得了较好的效果.

本文以强化学习算法为基础,重点讨论移动机器人运动规划领域的强化学习方法,分别从轮式机器人、无人机、足式机器人及多机器人系统4个方面进行深入的应用分析,总结国内外学者在此领域做出的主要工作和最新研究进展,并探讨目前尚存在的问题和深度强化学习方法在移动机器人运动规划领域中的未来研究方向.

1 机器人运动规划中的强化学习

首先给出建立模型所用的变量(见表1).

表1 建立模型所用变量

符号	变量名称	符号	变量名称
$Q(s_t,a_t)$	动作价值函数	δ	TD误差
$V(s_t)$	状态价值函数	θ	actor网络参数
π	策略函数	w	critic网络参数
G_t	累计收获	J	目标函数
γ	折扣因子	$H(\alpha)$	策略的熵
α	迭代步长	$\hat{\alpha}$	正则化系数

基于强化学习的运动规划是一种基于数据的非监督式机器学习方法,集成了感知和规划于一体,通过策略学习实现端到端的运动规划. 其借鉴人类试错的思想,利用机器人与动态环境的反复交互,以获得最大奖励为目标不断优化机器人的动作选择,从而规划得到最优策略,完成在未知环境中的自主运动规划. 由于其不依赖完备的环境先验知识,能够通过自主学习提升自身的运动规划策略,尤其是在地图残缺或环境未知的任务中,比常规方法更为有效.

基于强化学习的机器人运动规划通过强化学习方法构建传感器端到动作执行端的模型网络,实现环境感知到机器人动作的直接映射,从而使得环境响应速度得以加快. 其学习过程可以概括为如图1所示的运动样本采集、策略优化、动作选择规则以及策略执行等环节.

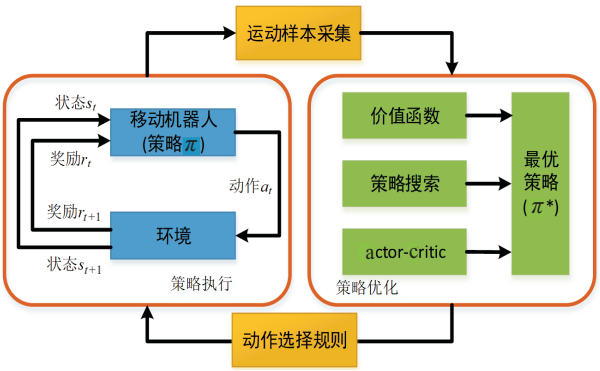


图1 移动机器人强化学习运动规划[18]

具体而言,在策略执行过程中,移动机器人运动规划策略可以分为两个部分:策略执行和策略优化. 在策略执行阶段,机器人根据当前状态s_t获得奖励r_t,然后根据策略π规划出下一步的执行动作a_t,如直行、旋转、加减速等,从而获得奖赏值r_{t+1}并转移至新的状态s_{t+1}. 同时,机器人在此过程中不断采集运动样本,送入策略优化部分,通过最优化移动机器人动作选择策略,制定机器人动作选择规则. 获取最优策略的方式包括两种:一种是通过最大化价值函数$Q(s_t,a_t)$来间接优化策略π;另外一种方式是通过策略搜索的方式,直接进行优化动作选择策略π,无需中间过程. 机器人得到最优策略则标志着机器人动作选择规划的完成,接下来移动机器人只需执行最优策略去完成目标任务.

在机器人规划下一步动作的过程中,依据是否存在价值函数的指导,可将方法分为三类:基于价值的、基于策略的和基于行动者-评论家框架的运动规划算法.

1.1 基于价值的运动规划方法

以价值函数为基础的强化学习方法常被用于机器人离散动作空间的运动规划. 机器人的动作常被分解为离散化运动控制指令的组合,然后根据价值函数选择最优动作集,常用的方法主要包括蒙特卡罗算法(MC)[20]、Q-leaning[21]及SARSA[22]等几种.

1) 蒙特卡罗算法(MC).

蒙特卡洛算法是一类利用自身经验来估计环境模型的方法[20],利用随机采样方式来近似求解每个状态的动作-价值函数$Q(s_t,a_t)$. 移动机器人通过与环境的交互获得机器人的状态信息s_t(位置、速度、周围障碍物分布)、机器人自身所采取的动作a_t以及即时奖励r_t,组成若干实际或模拟完整的状态序列样本,通过迭代进行策略评估计算机器人整个训练过程中的每个状态所对应的动作价值函数$Q(s_t,a_t)$,然后基于ϵ-贪婪法更新机器人动作选择策略π,直到得到最优价值函数. $Q(s_t,a_t)$的更新策略[23]为

$$Q(s_t,a_t)=Q(s_t,a_t)+\frac{1}{N}(G_t-Q(s_t,a_t)). \quad (1)$$

其中:G_t为某个状态s的奖励收获评估,N为累计更新计数次数. 由于蒙特卡洛算法每次采样都需要一个完整的状态序列才能进行策略更新,在动态随机的环境下很难获得完整序列的样本,使得算法存在机器人动作选择策略更新困难以及运动规划效率低的问题.

2) Q-learning算法.

Q-learning是一种使用时序差分(TD)算法来求解强化学习运动规划的方法. 不同于蒙特卡洛算法,Q-learning是一种单步更新的离线学习方法,即机器人每走一步,就可以对移动机器人的状态估值一次,同时只需要部分状态序列经历就可以完成价值函数的估计. 算法采用状态-动作值函数$Q(s_t,a_t)$作为评估函数. 首先,根据-贪婪策略选择机器人动作a_t;然后,利用贪婪策略从不同动作的估计值中选择最大Q值进行更新,即

$$Q^*(s_t,a_t)=$$
$$Q(s_t,a_t)+\alpha(r_t+\gamma\max_a Q(s_{t+1},a_t)-Q(s_t,a_t)). \quad (2)$$

其中:α为迭代步长,r_t为在当前状态下的奖励,s_{t+1}为机器人下一个状态,γ为0-1之间的折扣因子.

Q-leanring算法比蒙特卡罗算法更加灵活,学习能力更强,是当前普遍适用的求解强化学习问题的方法,在多种未知任务环境中都表现出良好的效果,已成功地应用于移动机器人在线自主导航、动态、静态障碍的碰撞避免和未知目标搜索以及多目标收集等任务[24].

3) SARSA算法.

SARSA也是基于时序差分(TD)算法的一种,采用了与Q-learning类似的结构,但其始终使用相同策略来更新价值函数和选择新的动作. 更新过程中,状态-动作价值函数$Q(s_t,a_t)$采用了实际值,而非估计值,即

$$Q(s_t,a_t)=$$
$$Q(s_t,a_t)+\alpha(r_t+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)). \quad (3)$$

机器人运动规划过程严格根据当前策略π对价值函数$Q(s_t,a_t)$进行更新,动作行为选择与价值函数迭代同时进行,较之Q-learning具有更快的收敛速度,已被验证可用于加快机器人训练速度、减少规划时长的问题[25],并成功地应用于城市交通车辆的运动规划和基于移动边缘计算的城市任务卸载和资源分配任务[26-27]. 但是,SARSA相比于Q-learning运动算法会更加保守,训练中价值函数的估计方差较小,变化平滑,规划结果不会产生大幅度的跳跃,但也有可能陷入局部最优的困境.

基于价值的强化学习算法在移动机器人运动规划领域有着简单高效的特点,但其是通过价值函数的迭代来优化动作选择策略π,主要体现以下局限性: 1) 对连续动作的处理能力不足; 2) 对受限状态下的问题处理能力不足; 3) 无法解决随机策略问题.

1.2 基于策略的运动规划方法

有别于基于价值的方法,基于策略的强化学习方法直接尝试优化策略函数π实现规划. 策略函数π给出的是移动机器人动作决策的概率分布,依照概率分布选定机器人动作,输出动作一般是运动机构的驱动扭矩或者动作参数. 为了优化动作选择策略π,移动机器人在策略函数中引入了策略参数θ来构建概率分布的形态,从而把对策略函数π的优化转化为对参数θ的优化. 然后,用蒙特卡罗法计算序列每个时间位置t的状态价值$V(s_t)$,再使用梯度上升法更新策略函数的参数θ,即

$$\theta=\theta+\alpha\nabla_\theta\log\pi_\theta(s_t,a_t)V(s_t). \quad (4)$$

将返回的策略参数传递给策略函数,即可进行策略函数的更新. 利用策略梯度算法,人们实现了水下航行器的运动策略规划[28]和无人机的目标探测与跟踪[29]等任务.

基于策略的强化学习方法是通过随机方式获取机器人运动序列进行策略更新,其策略参数θ是一个

无偏估计量,具有较大噪声,而且需要完全的序列样本才能进行策略函数的更新,学习效率较低,缺乏随机探索能力.

1.3 基于行动者-评论家的运动规划方法

行动者-评论家(actor-critic, AC)算法利用价值函数评估动作策略函数,保证学习过程同时兼顾价值函数与策略函数,解决了蒙特卡洛策略梯度算法依赖完整样本序列遍历、无法同时探索和训练更新的问题. 算法中的行动者(actor)提供策略函数,负责移动机器人动作选择,然后与环境交互;评论家(critic)等同于TD算法中的价值函数,负责计算每一步的价值,然后评估行动者的动作选择策略,并以此指导行动者选择下一步的动作,更新网络参数,有

$$\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t)\delta, \tag{5}$$

其中δ为TD计算误差. 移动机器人根据行动者选择的动作获得机器人当前状态和下一状态的特征向量,并通过评论家计算机器人当前状态的价值函数$V(s_t)$,行动者利用$V(s_t)$迭代时的TD误差更新机器人策略函数的参数θ,进而优化选择机器人动作,并得到反馈和新的状态,之后critic网络使用反馈和机器人新的状态继续更新自己的网络参数w.

基于AC策略框架的强化学习算法被用于实现连续动作空间的移动机器人规划,解决移动机器人室内导航[30]和目标搜索[31]等任务.

2 运动规划中的深度强化学习

深度强化学习(DRL)的出现,解决了强化学习对于高维信息感知能力缺乏的问题. 在强化学习的基础上,DRL增加了深度学习神经网络结构,并借助其对图像和高维数据的强大处理能力,对传感器输入的环境状态进行特征提取,将环境状态映射到动作价值函数[32];根据采用的DRL算法框架的不同,可分为基于价值的DRL运动规划和基于actor-critic的DRL运动规划方法.

2.1 基于价值的DRL运动规划

深度Q网络(DQN)[33]是基于价值的深度强化学习算法,它将深度神经网络与传统强化学习Q-learning算法相结合,通过将当前机器人所处的环境变量s_t作为价值网络输入,输出机器人动作所对应的动作价值函数的$Q(s_t, a_t)$值. 以获得最大累计奖励为目标,从经验回放池中取出机器人已经存储的样本数据进行估值更新,并以此为标签,用损失函数Loss不断拟合,让输出的动作价值$Q(s_t, a_t)$向价值评估回溯后的目标价值函数y_i不断靠近,直至逼近真实$Q(s_t, a_t)$收敛,完成价值网络的参数学习. 最后,根据动作选择策略以合适的探索率选择机器人动作.

采用DQN强化学习算法进行机器人运动规划时存在两个问题:一个是会过高估计Q值,移动机器人会陷入局部路径最优的困境;另一个是机器人的连续运动所采集的样本本身之间存在较强的相关性,容易导致机器人价值网络更新梯度消失,使机器人无法继续学习. 针对前一个问题:Hasselt等[34]提出了基于双神经网络的Double DQN算法,包括一个最优动作选择网络和一个价值函数估值网络,解决了Q值过高估计的问题. 针对后一个问题,Schaul等[35]提出了基于优先经验回放的DQN(prioritized replay DQN)算法,结构如图2所示. 其利用记忆回放单元使得神经网络更新更有效率,打破了采集样本之间的相关性,实现了Q网络上的TD误差的绝对值最大化.

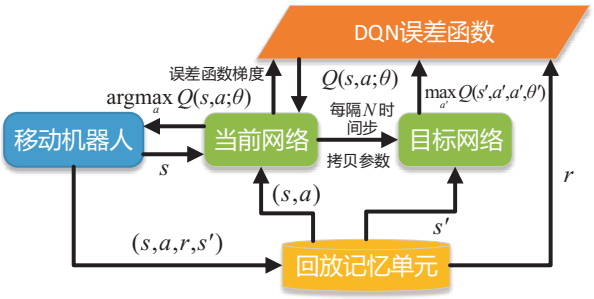


图2 基于经验回放的DQN算法网络结构[35]

DQN及其改进算法的不足之处在于只能处理移动机器人离散的动作集合,对于连续的动作空间则体现出较大的局限性.

2.2 基于actor-critic的DRL动作运动规划

在连续动作空间的机器人运动规划任务中,机器人输出的不再是有限的离散的动作集合,而是在连续动作域中每个动作的执行概率. 如轮式机器人作用于主动轮上的速度和加速度的大小变化,腿式机器人作用于每个关节上力与力矩的输出. 针对这类任务,基于actor-critic的DRL运动规划方法有着更强的适用性,常用的算法包括DDPG[36]、TRPO[37]、PPO[38]、A3C[39]、SAC[40]等,下面逐一介绍.

1) DDPG.

针对连续型动作空间的移动机器人运动规划,Lillicrap等[36]基于行动者-评论家(AC)算法框架,提出了一种无模型的深度确定性策略梯度算法(DDPG),具备了移动机器人连续动作规划的能力. DDPG算法为移动机器人构造了4个神经网络,同时对策略函数和价值函数进行估计,分别包括actor、critic的目标网络和当前网络. actor当前网络负责机

器人策略网络参数θ的迭代更新, actor目标网络中的参数θ'根据固定步长从当前网络中复制θ; critic当前网络负责价值网络参数w的迭代更新并计算当前Q值, critic目标网络参数w'定期从w中复制. 移动机器人运动规划方法如图3所示.

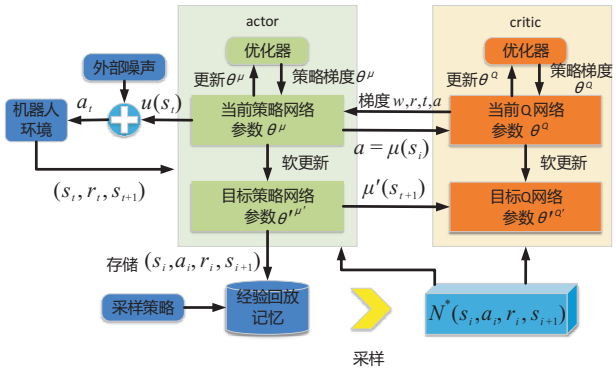


图 3　基于DDPG算法的运动规划方法[36]

DDPG通过记忆回放单元进行随机选择, 打破了数据的相关性, 提高了算法效率, 采用确定性策略直接输出移动机器人最大价值函数所对应的动作, 并将DDPG算法的优化目标函数定义为

$$J(\theta^{\mu})=E_{\theta^{\mu}}[r_1+\gamma r_2+\gamma^2 r_3+\ldots], \tag{6}$$

其中μ是产生确定性动作的策略网络的参数.

DDPG较好地处理了移动机器人连续控制的问题, 同时通过双网络结构和优先经验回放机制解决了actor-critic难以收敛的问题, 已成功地应用于移动机器人的运输、搜索、追踪和自动驾驶等任务[41-42].

2) TRPO.

在DDPG策略梯度的优化过程中, 更新步长将直接决定移动机器人是否能够快速准确地到达目标点. 为了确定合适的步长, Schulman 等[37] 提出了一种基于信赖域策略优化的强化学习算法(TRPO). TRPO以计算新策略与老策略之间的KL离散度范围为基础, 以最大化动作价值函数和状态价值函数的差值, 即以优势函数为目标, 定义目标函数为

$$J_{\mathrm{TRPO}}^{\theta'}(\theta)=E_{\pi_\theta}[\pi_\theta/\pi_{\theta'}A^{\theta'}(s_t,a_t)], \tag{7}$$

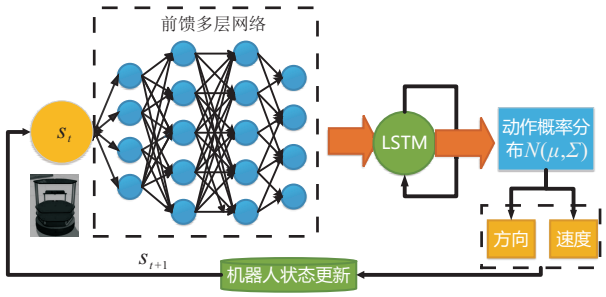


图 4　基于TRPO算法的运动规划方法

其中θ、θ'分别为新旧策略的网络参数.

图4表达了一种基于TRPO算法的移动机器人的运动规划方法, TRPO输出一种动作概率分布, 移动机器人可以根据这个概率分布优选自身的运动参数.

TRPO解决了策略梯度方法中更新步长难以确定的问题, 但在实现上进行了种种近似, 这些近似导致计算过程繁琐, 出现较大的误差.

3) PPO.

针对TRPO算法计算过程过于繁杂的问题, 从而导致机动机器人运动规划路径与最优路径产生偏差的问题, Schulman 等[38] 提出了一种近端策略优化的强化学习算法(PPO), 使移动机器人在动作选择时具有更好的探索性. PPO算法直接将新旧策略的KL(θ,θ')散度作为惩罚项, 其目标函数被更新为

$$J_{\mathrm{PPO}}^{\theta}(\theta)=E_{\pi_j}\left[\frac{\pi_\theta}{\pi_\theta}A^{\theta}(s_t,a_t)\right]-\beta\mathrm{KL}(\theta,\theta'), \tag{8}$$

其中β为离散度更新参数. 相比于TRPO强化学习算法, PPO算法大幅度简化了TRPO的计算步骤, 同时保留了随机策略的探索方式, 在采样样本满足最大似然概率情形下, 机器人的运动规划方法将具有更好的探索性和鲁棒性.

4) A3C.

为了打破移动机器人数据样本之间的相关性, Mnih 等[39] 借鉴异步强化学习的优势, 提出了异步优势行动者-评论家强化学习算法(A3C). A3C算法让机器人在相同的环境进行不同步的交互学习, 每个线程中的机器人都在随机探索, 共同更新共享策略模型参数θ的权重. 其策略参数的梯度更新方式为

$$\theta=\theta+\alpha\nabla_\theta\log\pi_\theta(s_t,a_t)A(s_t,\mathrm{a}_t)+c\nabla_\theta H(\pi_\theta). \tag{9}$$

其中: α为更新步长, H为策略π的熵项, c为系数. 基于A3C算法的机器人运动规划原理如图5所示.

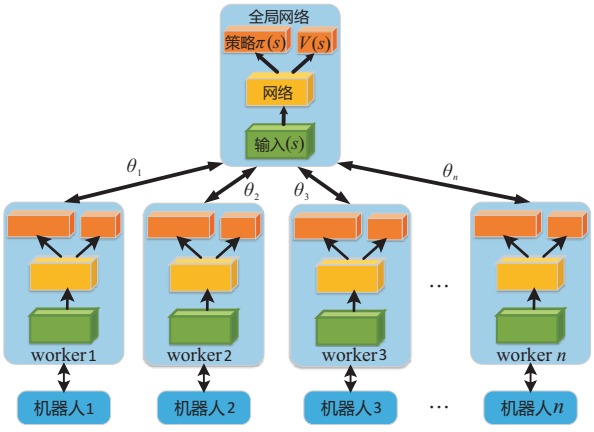


图 5　基于A3C算法的运动规划方法[39]

基于A3C算法的移动机器人运动规划方法,可以应用多线程同时训练,效率较高;但是训练数据过于庞大,需要多个机器人同时异步工作,在现实场景中应用成本较高.

5) SAC.

SAC[40]是一种基于最大熵的离线学习强化学习方法,移动机器人也是通过随机策略来进行选择动作,既可以学习过去的经验,又可以学习其他的任务经验. 如图6所示,基于SAC的强化学习运动规划方法通过最大熵的方法来最大化目标函数,其中基于最大熵的目标函数为

$$J(\theta)=E_{\pi}\left[\sum Q(s_t,a_t)-\hat{\alpha}\log\pi(a_t|s_t)\right], \quad (10)$$

其中$\hat{\alpha}$为熵正则化系数,包含了来自机器人真实系统的动力学参数.

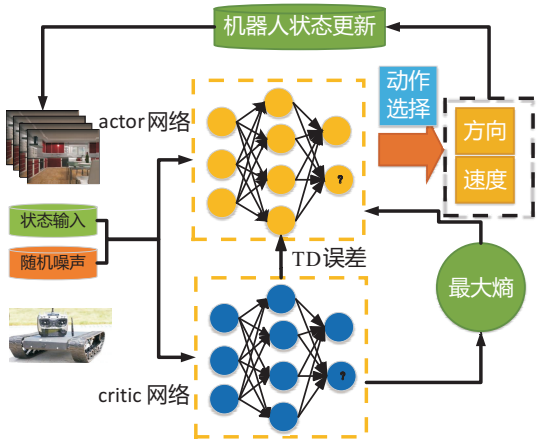


图 6 基于SAC算法的机器人运动规划原理

SAC运动规划算法在熵的约束下,通过一个鲁棒的学习框架来最小化超参数调节的需求,使用正则化系数来代替超参数更新移动机器人动作选择策略网络,实现了移动机器人目标函数的最大化预期回报,改善了策略梯度算法对超参敏感性过大、收敛性脆弱的问题,使SAC运动规划算法性能更加稳定,更加适用于真实环境,可实现腿式等移动机器人的行走技能的快速学习与规划[43].

3 DRL 运动规划的应用

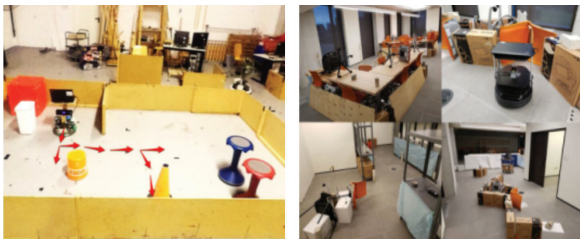
近几年,各种类型的机器人开始应用深度强化学习的运动规划算法来代替常规算法,实现了从仿真到真实场景、从地面机器人到无人机、从轮式到腿式机器人等各个领域的广泛研究.

3.1 轮式移动机器人运动规划

基于DRL的轮式移动机器人运动规划直接以外部传感数据作为输入,通过神经网络的映射,输出为驱动轮的转动速度值或转向轮的转向角度. 在对机器人速度变化要求不高的场景下,通常采用DQN算法,将输出参数离散化,以有限的动作数量驱动机器人完成导航;在对机器人的速度控制和运动精度有严格要求的环境下,一般采用基于actor-critic的DDPG或A3C算法,以连续的参数输出,使运动轨迹更加平滑.

在仿真环境条件下,Tai等[44]基于DQN强化学习算法,使用深度图像作为Q网络的输入,实现了在含有多种静态障碍物的仿真环境下移动机器人的运动规划. Barron等[45]使用两层及以上的深度神经网络架构,采用可见光RGB图像作为DQN网路的输入,在3D环境中实现了较好的运动规划效果.

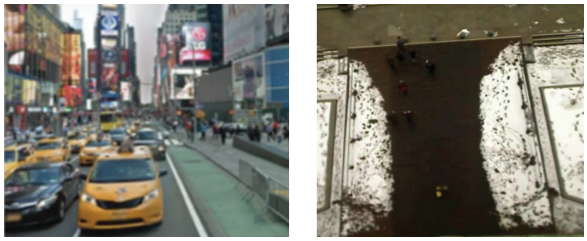
在室内环境中,Zhang等[46]利用迁移强化学习方法,以深度图像作为神经网络输入提取环境特征,实现了移动机器人在现实环境中的导航控制,如图7(a)所示. Niroui等[47]利用RGB-D深度信息相机和机器人里程计数据生成环境的二维占用网格,使机器人在无地图导航环境中获得更好的样本效率,实现了机器人在真实未知复杂三维环境中的避障导航,如图7(b)所示.



(a) (b)

图 7 轮式移动机器人室内环境运动规划[46-47]

在室外环境中,Mirowski等[48]应用一种交互式导航环境,直接使用图8(a)所示的真实街景作为视频输入进行训练,完全模仿了真实环境,并成功迁移到无人车,实现了跨城市导航. Li等[49]利用从大量真实人群数据集中收集的地图和行人轨迹,重新建立了一个有效的模拟学习环境,解决了移动机器人在真实居住环境中与人类进行社交导航的问题,如图8(b)所示.



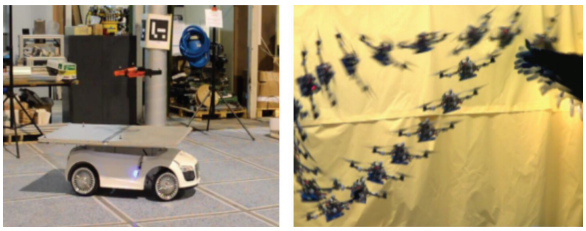
(a) (b)

图 8 轮式移动机器人室外环境运动规划[48-49]

3.2　无人机的运动规划

相比于地面移动机器人,无人机的运动规划将更加复杂,无人机的飞行不仅涉及到自身飞行路径的规划,而且涉及到在不同环境中飞行机动特性等. 不同于传统方法需要对无人机进行建模,深度强化学习将无人机的运动状态直接映射到无人机的飞行控制参数,通过反复训练,实现无人机自主飞行规划.

在解决无人机机动动作学习任务中,Rodriguez-Ramos 等[50] 基于确定性策略强化学习算法,设计了一种通用的强化学习框架,解决了无人机在移动平台上的着陆操作,如图 9(a) 所示. Hwangbo 等[51] 应用强化学习方法训练无人机起飞的策略网络,使其能够较快地响应阶跃信号,在非常苛刻的条件下,实现了无人机的手动抛飞并空中悬停,如图 9(b) 所示.



(a)　　(b)

图 9　强化学习训练无人起飞降落[50-51]

在解决无人机自主导航避障任务中,Kang 等[52] 提出一种真实数据与模拟数据混合学习的强化学习算法,用来训练无人机在室内环境中穿过走廊,实现了无人机仅使用单目相机就可以在室内避免碰撞. Maciel 等[53] 针对特殊天气条件下森林环境中的无人机自主导航的问题,将周围信息和距离目标的距离作为强化学习的双状态输入,通过改变奖励函数和网络模型,实现了无人机在多个不可预测的森林环境和恶劣天气下的导航与避障.

在解决无人机的规划决策类任务中,Wang 等[54] 根据部分可观测目标位置的量子概率模型以及自身的能量消耗,提出了一种用于跟踪和搜索的在线分布式强化学习算法,解决了无人机在目标搜索和运动规划最佳路线问题. Kulkarni 等[55] 利用强化学习训练无人机辅助搜索和救援,通过感知距离受害者信号源的距离,将其作为训练的输入,在目标搜索中实现了更加稳定的导航效果.

3.3　足式机器人的运动规划

足式机器人的运动规划关键问题在于解决机器人的步态规划. 深度强化学习不必完全依赖于精确的动力学建模,通过仿真环境训练,最终获得符合多足机器人结构、满足性能需求的规划策略. Haarnoja 等[56] 应用 SAC 算法,如图 10 所示,在不依赖仿真或者示教的情况下,将传感器数据通过神经网络直接映射到低级动作,实现了足式机器人复杂数据的自主学习,所得到的策略对环境中的适度变化具有较好的鲁棒性.

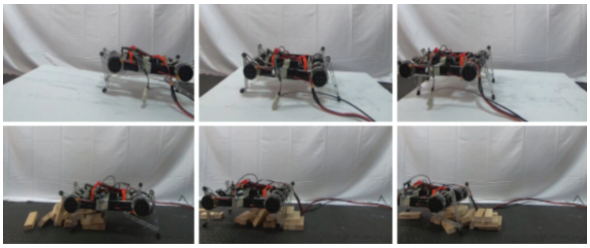


图 10　真实环境中训练的机器人越障[56]

深度强化学习足式机器人运动规划对于解决含有多种障碍的环境问题具有较强的适应能力,在面对不同的步态需求时不需要重新设计. Xie 等[57] 利用深度强化学习方法训练双足机器人在不同的环境中行走,使其掌握了多种行走方式,如图 11 所示的正向、倒向以及侧向等. 吴晓光等[58] 利用基于 DDPG 的深度强化学习算法,融合分布式优先经验回放机制,提高了训练样本的利用率,缩短了学习时长,实现了双足机器人在斜面上的稳定行走.

图 11　双足机器人 Cassie 在传送带行走训练[57]

深度强化学习足式机器人运动规划具备容错能力,可以处理容易出现的结构损伤、执行机构故障等问题,使机器人具备危险条件下的特殊机动能力. Hwangbo 等[59] 在仿真环境中训练 5 足机器人的行走技能,并在现实中实现了 4 足机器人 ANYmal 稳定行走,如图 12 所示,而且跌倒后具备自行恢复能力. 另外,面对足式机器人腿部意外损伤而无法正常行走的情况,Chattunyakit 等[60] 利用基于价值的强化学习方

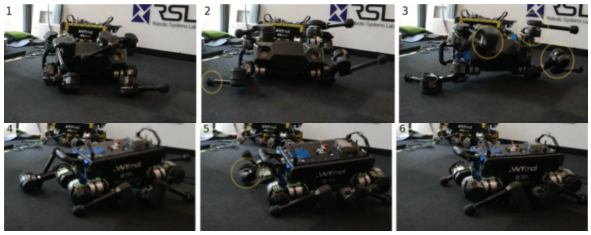


图 12　在真实世界中训练 4 足机器人 ANYmal 行走[59]

法训练腿部损伤机器人的非常规行走策略,实现了4足机器人多种腿部损伤情况下的自适应恢复并保持正常运动.

深度强化学习还适应于复杂特种结构的足式机器人,如弹射-腿结构[61]、轮腿结构[62]等.

Chen等[62]基于一种改进的PPO算法训练轮腿机器人的运动技能. 该策略将高维地图图像直接映射到马达命令,将复杂的导航任务分解成若干可管理的导航行为,并且提出了一种领域随机化技术,使机器人能够克服数据效率低、奖励稀疏等问题.

3.4 多机器人协同运动规划

机器人的任务需求与任务环境日趋复杂,依靠单个机器人的工作方式存在关键性短板,如感知能力有限、任务可靠性低、执行效率低下等. 近10年来,开展多机器人协同、协调、协作的方式,共同执行任务更具优势,逐渐成为机器人领域研究共识与研究热点[63]. 多机器人合作的工作方式对团体中的个体间运动规划提出了更多挑战,如何有效地开展协同运动规划,已成为这一领域有别于单机器人运动规划的独有特征. 多移动机器人强化学习运动规划系统的主要结构可分为集中式[64]和分布式[65]两类.

集中式强化学习以多机器人的共同任务为训练目标,通过集中计算单元获取所有机器人的状态和传感器信息,由集中计算单元进行集中策略的训练并负责分发. 集中式强化学习的优势是可以直接获得任何机器人的实时位置、速度和目标信息,全局指挥所有个体机器人完成最优条件下的任务目标分配、合作机制、路径协同等,在保证任务完成的同时尽可能地减少能量、时间以及其他自身的损失;但是面对大规模的移动机器人系统时,计算消耗将变得十分巨大. 为了解决这一问题,Chen等[66]通过预计算值函数来估计到达目标的时间,将集中式的在线计算降解为分布式的离线计算过程,得到一种计算更加实时高效的多机器人运动规划系统. 除了计算消耗,集中式强化学习通讯困难将随着机器人数量呈指数增加,无法保证信息获取的准确性与实时性. Everett等[67]为了解决深度强化学习中多机器人运动规划系统随着机器人数量的增加而偏离现实目标的问题,基于A3C框架的强化学习方法来模拟机器人之间复杂的交互与合作,利用LSTM循环神经网络实现了在不预设其他机器人行为规范下的自主运动规划.

分布式强化学习系统,首先需要将整体任务进行实时分割,然后将分割后的子任务派送给单个机器人,机器人收到任务后,作为单独的个体,以自身最大奖励为目标的同时兼顾整体总回报,通过协作实现整体任务与规划. 分布式强化学习的优势是可以充分调动每个机器人的资源,提高硬件利用率,每个机器人是独立个体,容错性和冗余性强.

分布式强化学习多机器人系统,虽然调动了每个机器人的训练资源,但是,单个机器人容易出现过分追求自身“奖励”的最大化,不能兼顾其他机器人的任务目标,造成过分占用系统资源,效率低下. 由于这种“竞争”关系的存在,机器人个体之间还容易发生冲突,造成相互碰撞等问题. 针对这些问题,Malus等[68]让单个机器人根据各自的观察值以及自身的位置和当前计划,学习对订单进行投标,实现了多个移动机器人进行物流运输调度任务的高效运转. Shi等[69]利用交互特征的注意机制捕获各无人车之间的相互作用信息,促进多无人车的信息共享策略,提高了无人驾驶车辆交通运输的效率和安全性.

针对碰撞避免的问题,Long等[70]基于策略梯度强化学习算法,提出了一种传感器级分布式强化学习算法,将包含速度信息的原始传感器数据映射到转向命令,成功实现了多种场景中的碰撞避免. Semnani等[71]通过引入新的混合性奖励函数,设计了混合式强化学习避碰方法,在不同环境复杂度下进行算法层级的自动切换,提高了机器人对外部环境的响应速度,降低了多个机器人间发生碰撞的概率.

4 强化学习运动规划未来方向

1) 复杂环境状态表征的强化学习运动规划.

移动机器人对环境的感知能力是深度强化学习运动规划的关键因素之一. 目前,大多数移动机器人所依赖的光学传感器普遍存在采样效率不足、实时反馈迟滞、环境信息缺失、精确程度不足等通病[72],加之机器人所处的真实环境存在光照、季节、天气等条件变化,进一步制约了移动机器人对环境观测与识别效果,进而对机器人的运动规划带来巨大的挑战. 如何在复杂环境中,保证深度强化学习对时变真实环境特征的准确感知,以及具备语义描述等抽象推理化表征,进而使机器人具备复杂场景的自动推理与自主运动决策能力,是移动机器人在复杂环境中顺利执行任务的关键问题.

2) 异构系统空天地协同的运动规划.

随着移动机器人空间立体化作业任务需求优势日渐明显,任务环境向着空天地异构机器人协同方向逐步迈进. 当前,基于强化学习的移动机器人运动规

划多用于解决低维平面的地面环境下导航与避撞问题,而针对空间立体环境中的协调式运动规划研究则较少,例如,陆地海洋两栖机器人、陆空协同作业仿生机器人系统的运动规划等. 未来,以基于AC策略的强化学习算法为基础,结合大规模的深度神经网络,增强数据的预处理,进一步提高深度强化学习对于输入特征的感知和提取能力,以不同的数据优先级来应对异构系统中激增的数据信息量,保证异构系统中智能体的学习训练效率,是异构系统空天地协同的运动规划的重要需求.

3) 多源域迁移的强化学习运动规划.

深度强化学习运动规划模型的训练过程耗时严重、效率低,因此,具有良好通用性和迁移能力的模型将会大大提高机器人运动规划效率. 然而,在实际应用中,由于原始数据和目标域的不同以及机器人系统互相之间的差距,实际环境中的模型迁移面临着巨大的问题. 目前,大多数强化学习任务迁移效果大多依赖于源领域与目标领域的相关程度,并且需要保证源域和目标域处于同一数据空间. 未来,研究如何根据移动机器人不同源域任务的环境样本,自适应衡量多种源域与目标域特征的差异,提取出共同特征或参数,并在统一的状态空间进行表征,最终获得多源域间不变的特征参数,实现强化学习运动规划算法中多源域的任务迁移、参数迁移以及特征迁移,是移动机器人在未来更加多样化的环境中执行任务的重要保证.

4) 数据-模型混合的强化学习运动规划.

深度强化学习运动规划算法一直面临样本利用率非常低、奖励函数难以设计的问题,以致算法易陷入局部极优,无法找到最优规划模型. 基于模型的运动算法表现出了比强化学习算法更加高效的性能,但是面对复杂任务又会出现建模过于复杂、动态特性无法表征的问题. 将基于模型的运动规划算法与无模型强化学习运动规划算法相结合,可以先从数据中学习环境模型,然后基于学到的模型对策略进行优化,并反向更新和完善模型,从而充分利用环境采样样本来逼近模型,提高训练数据使用效率,缩短机器人的学习过程. 这种无模型与基于模型互补的混合式强化学习运动规划算法,是未来强化学习高效运行的关键途径之一.

5 结 论

基于强化学习的机器人运动规划方法是一种通过环境交互不断提升自身的无模型运动规划方法,它可使机器人具备良好的环境分析和学习能力,尤其适用于地图残缺、复杂多变的动态非结构化环境. 该方法降低了编程的复杂程度,脱离了对地图模型和先验知识的依赖,提高了机器人在任务执行过程中的航迹规划、实时避障、目标搜索和自我决策等能力,让机器人向着更加智能化的方向发展. 就目前发展现状而言,深度强化学习运动规划算法大多还处于实验室阶段,与真实世界中的运动规划的条件还有较大的差距,例如行走地面的非结构化、通信和数据流延迟、机械结构可靠性不足等,这些都为强化学习“试错”类型的训练方式带来不小的挑战,也是下一步深度强化学习在移动机器人运动规划领域深入应用时亟待解决的问题.

参考文献(References)

[1] Wang X C, Wang X L, Wilkes D M. Reinforcement learning for mobile robot perceptual learning[M]. Berlin: Springer, 2019: 253-273.

[2] 曹风魁, 庄严, 闫飞, 等. 移动机器人长期自主环境适应研究进展和展望[J]. 自动化学报, 2020, 46(2): 205-221.
(Cao F K, Zhuang Y, Yan F, et al. Long-term autonomous environment adaptation of mobile robots: State-of-the-art methods and prospects[J]. Acta Automatica Sinica, 2020, 46(2): 205-221.)

[3] Panda M, Das B, Subudhi B, et al. A comprehensive review of path planning algorithms for autonomous underwater vehicles[J]. International Journal of Automation and Computing, 2020, 17(3): 321-352.

[4] Hu Y J, Yao Y, Ren Q, et al. 3D multi-UAV cooperative velocity-aware motion planning[J]. Future Generation Computer Systems, 2020, 102: 762-774.

[5] Zhan W W, Wang W, Chen N C, et al. Path planning strategies for UAV based on improved A* algorithm[J]. Geomatics and Information Science of Wuhan University, 2015, 40(3): 315-320.

[6] 陈靖, 辜丽川, 李倩倩, 等. 基于D*算法的农用履带机器人路径规划研究[J]. 安徽理工大学学报: 自然科学版, 2019, 39(1): 31-37.
(Chen J, Gu L C, Li Q Q, et al. Rsearch on path planning of agricultural robot based on D* algorithm[J]. Journal of Anhui University of Science and Technology: Natural Science, 2019, 39(1): 31-37.)

[7] Song P L, Huang J F, Mansaray L R, et al. An improved soil moisture retrieval algorithm based on the land parameter retrieval model for water‐land mixed pixels using AMSR-E data[J]. IEEE Transactions on Geoence and Remote Sensing, 2019, 57(10): 7643-7657.

[8] Mashayekhi R, Idris M Y I, Anisi M H, et al. Informed RRT*-connect: An asymptotically optimal single-query path planning method[J]. IEEE Access, 2020, 8: 19842-19852.

[9] Ge Y X, Song B F, Pei Y, et al. A fuzzy logic based method for fault tolerant hierarchical load management of more electric aircraft[J]. Proceedings of the Institution of Mechanical Engineers, 2019, 233(10): 3846-3856.

[10] Yu X Y, Zhu Y C, Lu L, et al. Dynamic window with virtual goal (DW-VG): A new reactive obstacle avoidance approach based on motion prediction[J]. Robotica, 2019, 37(8): 1438-1456.

[11] Xu Y, Liu X, Hu X, et al. A genetic-algorithm-aided fuzzy chance-constrained programming model for municipal solid waste management[J]. Engineering Optimization, 2020, 52(4): 652-668.

[12] Luis Fernando de Mingo López, Nuria Gómez Blas, Angel Luis Castellanos Peñuela, et al. Swarm intelligence models: Ant colony systems applied to BNF grammars rule derivation[J]. International Journal of Foundations of Computer Science, 2020, 24(5): 3141-3154.

[13] Zhao X, Wang C, Su J, et al. Research and application based on the swarm intelligence algorithm and artificial intelligence for wind farm decision system[J]. Renewable Energy, 2019, 134: 681-697.

[14] Wang J K, Meng M Q H, Khatib O. EB-RRT: Optimal motion planning for mobile robots[J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(4): 2063-2073.

[15] 张天泽. 基于强化学习的四旋翼无人机路径规划方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2018: 22-30.
(Zhang T Z. Research on path planning method of four rotor UAV based on reinforcement learning[D]. Harbin: Harbin Institute of Technology, 2018: 22-30.)

[16] Rudenko A, Kucner T P, Swaminathan C S, et al. THOR: Human-robot navigation data collection and accurate motion trajectories dataset[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 676-682.

[17] 张汕璠. 基于强化学习的路径规划方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2018: 1-69.
(Zhang S F. Research on path planning method based on reinforcement learning[D]. Harbin: Harbin Institute of Technology, 2018: 1-69.)

[18] 刘乃军, 鲁涛, 蔡莹皓, 等. 机器人操作技能学习方法综述[J]. 自动化学报, 2019, 45(3): 458-470.
(Liu N J, Lu T, Cai Y H, et al. A review of robot manipulation skills learning methods[J]. Acta Automatica Sinica, 2019, 45(3): 458-470.)

[19] 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.

[20] Ni J N, Zhou Y, Li S J. Hamiltonian Monte Carlo-based D-vine copula regression model for soft sensor modeling of complex chemical processes[J]. Industrial Engineering Chemistry Research, 2020, 59(4): 1607-1618.

[21] Park K H, Kim Y J, Kim J H. Modular Q-learning based multi-agent cooperation for robot soccer[J]. Robotics and Autonomous Systems, 2001, 35(2): 109-122.

[22] Ramachandran D, Gupta R. Smoothed sarsa: Reinforcement learning for robot delivery tasks[P]. United States: US8326780. 2012-04-12.

[23] Ncubukezi T. A proposed: Integration of the Monte Carlo model and the Bayes network to propose cyber security risk assessment tool for small and medium enterprises in South Africa[J]. International Journal of Computer Ence and Information Security, 2020, 3(18): 152-155.

[24] Goswami I, Das P K, Konar A, et al. Extended Q-learning algorithm for path-planning of a mobile robot[C]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010: 379-383.

[25] 宋宇, 王志明. 基于改进SARSA(λ)移动机器人路径规划[J]. 长春工业大学学报, 2019, 40(1): 55-59.
(Song Y, Wang Z M. Path planning based on improved SARSA (λ)[J]. Journal of Changchun University of Technology, 2019, 40(1): 55-59.)

[26] Feng Y M, Wu Y R. Environmental adaptive urban traffic signal control based on reinforcement learning algorithm[J]. Journal of Physics: Conference Series, 2020, 1650: 032097.

[27] Alfakih T, Hassan M M, Gumaei A, et al. Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on SARSA[J]. IEEE Access, 2020, 8: 54074-54084.

[28] Sun Y S, Cheng J H, Zhang G C, et al. Mapless motion planning system for an autonomous underwater vehicle using policy gradient-based deep reinforcement learning[J]. Journal of Intelligent & Robotic Systems, 2019, 96(3/4): 591-601.

[29] You S X, Diao M, Gao L P. Deep reinforcement learning for target searching in cognitive electronic warfare[J]. IEEE Access, 2019, 7: 37432-37447.

[30] Kulhánek J, Derner E, de Bruin T, et al. Vision-based navigation using deep reinforcement learning[C]. European Conference on Mobile Robots (ECMR). Prague: IEEE, 2019: 8870964.

[31] Li T G, Pan J L, Zhu D L, et al. Learning to interrupt: A

hierarchical deep reinforcement learning framework for efficient exploration[C]. IEEE International Conference on Robotics and Biomimetics (ROBIO). Kuala Lumpur: IEEE, 2018: 648-653.

[32] 徐晓苏, 袁杰. 基于改进强化学习的移动机器人路径规划方法[J]. 中国惯性技术学报, 2019, 27(3): 314-320.
(Xu X S, Yuan J. Path planning for mobile robot based on improved reinforcement learning algorithm[J]. Journal of Chinese Inertial Technology, 2019, 27(3): 314-320.)

[33] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[34] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]. AAAI Conference on Artificial Intelligence. Arizona: AAAI Press, 2016: 2094-2100.

[35] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[C]. International Conference on Learning Representations. San Juan: ICLR, 2016: 1-23.

[36] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[C]. International Conference on Learning Representations. San Juan: ICLR, 2016: 1-13.

[37] Schulman J, Levine S, Moritz P, et al. Trust region policy optimization[C]. International Conference on Machine Learning. Lille: IMLS, 2015: 1889-1897.

[38] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithm[J]. 2017, arXiv: 1707.06347v2.

[39] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]. International Conference on Machine Learning. New York: ICML, 2016: 2850-2869.

[40] 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: IMLS, 2018: 2989-2996.

[41] Zhang D, Bailey C P. Obstacle avoidance and navigation utilizing reinforcement learning with reward shaping[C]. Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II. California: SPIE, 2020: 114131H.

[42] Wang C, Wang J, Wang J J, et al. Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards[J]. IEEE Internet of Things Journal, 2020, 7(7): 6180-6190.

[43] Kim J I, Hong M, Lee K, et al. Learning to walk a tripod mobile robot using nonlinear soft vibration actuators with entropy adaptive reinforcement learning[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2317-2324.

[44] Tai L, Liu M. Mobile robots exploration through cnn-based reinforcement learning[J]. Robotics and Biomimetics, 2016, 3(1): 1-8.

[45] Barron T, Whitehead M, Yeung A. Deep reinforcement learning in a 3-d blockworld environment[C]. International Joint Conference in Artifical Intelligence. New York: IJCAI, 2016: 1-6.

[46] Zhang J W, Springenberg J T, Boedecker J, et al. Deep reinforcement learning with successor features for navigation across similar environments[C]. International Conference on Intelligent Robots and Systems. Vancouver: IEEE, 2017: 2371-2378.

[47] Niroui F, Zhang K C, Kashino Z, et al. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 610-617.

[48] Mirowski P, Grimes M, Malinowski M, et al. Learning to navigate in cities without a map[C]. The 32nd Conference on Neural Information Processing Systems. Montreal: Neural Information Processing Systems Foundation, 2018: 2419-2430.

[49] Li M, Jiang R, Ge S S, et al. Role playing learning for socially concomitant mobile robot navigation[J]. CAAI Transactions on Intelligence Technology, 2018, 3(1): 49-58.

[50] Rodriguez-Ramos A, Sampedro C, Bavle H, et al, A deep reinforcement learning strategy for UAV autonomous landing on a moving platform[J]. Journal of Intelligent & Robotic Systems, 2019, 93(1/2): 351-366.

[51] Hwangbo J, Sa I, Siegwart R, et al. Control of a quadrotor with reinforcement learning[J]. IEEE Robotics and Automation Letters, 2017, 2(4): 2096-2103.

[52] Kang K T, Belkhale S, Kahn G, et al. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight[C]. IEEE International Conference on Robotics and Automation. Montreal: IEEE, 2019: 6008-6014.

[53] Maciel-Pearson B G, Marchegiani L, Akcay S, et al. Online deep reinforcement learning for autonomous UAV navigation and exploration of outdoor environments[J]. 2019, arXiv: 1912.05684v1.

[54] Wang T, Qin R, Chen Y, et al. A reinforcement learning approach for UAV target searching and tracking[J]. Multimedia Tools and Applications, 2019, 78(4): 4347-4364.

[55] Kulkarni S, Chaphekar V, Chowdhury M M U, et al.

UAV aided search and rescue operation using reinforcement learning, systems and control[J]. 2020: arXiv: 2002.08415v1.

[56] Haarnoja T, Ha S, Zhou A, et al. Learning to walk via deep reinforcement learning[J]. 2019: arXiv: 1812.11103v3 .

[57] Xie Z, Clary P, Dao J, et al. Iterative reinforcement learning based design of dynamic locomotion skills for cassie[J]. 2019, arXiv: 1903.09537v1.

[58] 吴晓光, 刘绍维, 杨磊, 等. 基于深度强化学习的双足机器人斜坡步态控制方法[J]. 自动化学报, 2020, 46(x): 1-12.
(Wu X G, Liu S W, Yang L, et al. A gait control method for biped robot on slope based on deep reinforcement learning[J]. Acta Automatica Sinica, 2020, 46(x): 1-12.)

[59] Hwangbo J, Lee J, Dosovitskiy A, et al. Learning agile and dynamic motor skills for legged robots[J]. Science Robotics, 2019, 4(26): 5872.

[60] Chattunyakit S, Kobayashi Y, Emaru T, et al. Bio-inspired structure and behavior of self-recovery quadruped robot with a limited number of functional legs[J]. Applied Sciences, 2019, 9(4): 1-25.

[61] Yuan Y X, Li Z J, Zhao T, et al. DMP-based motion generation for a walking exoskeleton robot using reinforcement learning[J]. IEEE Transactions on Industrial Electronics, 2020, 67(5): 3830-3839.

[62] Chen X, Ghadirzadeh A, Folkesson J, et al. Deep reinforcement learning to acquire navigation skills for wheel-legged robots in complex environments[C]. International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 3110-3116.

[63] Yu P, Dimarogonas D V. A fully distributed motion coordination strategy for multi-robot systems with local information[C]. American Control Conference. Denver: IEEE, 2020: 1859-1864.

[64] Zhou X Y, Wu P, Zhang H F, et al. Learn to navigate: Cooperative path planning for unmanned surface vehicles using deep reinforcement learning[J]. IEEE Access, 2019, 7: 165262-165278.

[65] Wang D W, Fan T X, Han T, et al. A two-stage reinforcement learning approach for multi-UAV collision avoidance under imperfect sensing[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3098-3105.

[66] 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. Singapore: IEEE, 2017: 285-292.

[67] Everett M, Chen Y F, How J P. Motion planning among dynamic, decision-making agents with deep Reinforcement learning[C]. IEEE International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 3052-3059.

[68] Malus A, Kozjek D, Vrabi R. Real-time order dispatching for a fleet of autonomous mobile robots using multi-agent reinforcement learning[J]. CIRP Annals, 2020, 69(1): 397-400.

[69] Shi T, Sun L J. Towards efficient connected and automated driving system via multi-agent graph reinforcement learning[J]. 2020, arXiv: 2007.02794v2.

[70] Long P X, Fan T X, Liao X Y, et al. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning[C]. IEEE International Conference on Robotics and Automation. Brisbane: IEEE, 2018: 6252-6259.

[71] Semnani S H, Liu H, Everett M, et al. Multi-agent motion planning for dense and dynamic environments via deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3221-3226.

[72] Ejaz M M, Tang T B, Lu C K. Autonomous visual navigation using deep reinforcement learning: An overview[C]. IEEE Student Conference on Research and Development. Bandar Seri Iskandar: IEEE, 2019: 294-299.

作者简介

孙辉辉 (1989–), 男, 博士生, 从事智能机器人及其控制的研究, E-mail: cumtsunhui@126.com;

胡春鹤 (1986–), 男, 讲师, 博士, 从事无人机自主控制、多无人机协同控制及其应用等研究, E-mail: huchunhe@bjfu.edu.cn;

张军国 (1978–), 男, 教授, 博士生导师, 从事智慧林业监测与信息处理、无人飞行器及林业特种机器人等研究, E-mail: zhangjunguo@bjfu.edu.cn.

(责任编辑：李君玲)