控制与决策
Control and Decision

基于深度强化学习的机器人运动控制研究进展

董豪，杨静，李少波，王军，段仲静

引用本文:
董豪, 杨静, 李少波, 等. 基于深度强化学习的机器人运动控制研究进展[J]. 控制与决策, 2022, 37(2): 278−292.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1382

基于深度强化学习的机器人运动控制研究进展

董 豪[1], 杨 静[1,2], 李少波[1,2,3†], 王 军[1], 段仲静[3]

(1. 贵州大学 机械工程学院，贵阳 550025；2. 贵州大学 省部共建公共大数据国家重点实验室(筹)，贵阳 550025；3. 贵州大学 现代制造技术教育部重点实验室，贵阳 550025)

摘 要: 复杂未知环境下智能感知与自动控制是目前机器人在控制领域的研究热点之一,而新一代人工智能为其实现智能自动化赋予了可能. 近年来,在高维连续状态-动作空间中,尝试运用深度强化学习进行机器人运动控制的新兴方法受到了相关研究人员的关注. 首先,回顾了深度强化学习的兴起与发展,将用于机器人运动控制的深度强化学习算法分为基于值函数和策略梯度2类,并对各自典型算法及其特点进行了详细介绍;其次,针对仿真至现实之前的学习过程,简要介绍5种常用于深度强化学习的机器人运动控制仿真平台;然后,根据研究类型的不同,综述了目前基于深度强化学习的机器人运动控制方法在自主导航、物体抓取、步态控制、人机协作以及群体协同等5个方面的研究进展;最后,对其未来所面临的挑战以及发展趋势进行了总结与展望.

关键词: 复杂未知环境；人工智能；高维连续空间；深度强化学习；仿真至现实；机器人运动控制

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1382 **开放科学(资源服务)标识码(OSID):**

引用格式: 董豪, 杨静, 李少波, 等. 基于深度强化学习的机器人运动控制研究进展 [J]. 控制与决策, 2022, 37(2): 278-292.

Research progress of robot motion control based on deep reinforcement learning

DONG Hao[1], *YANG Jing*[1,2], *LI Shao-bo*[1,2,3†], *WANG Jun*[1], *DUAN Zhong-jing*[3]

(1. School of Mechanical Engineering，Guizhou University，Guiyang 550025，China；2. State Key Laboratory of Public Big Data，Guizhou University，Guiyang 550025，China；3. Key Laboratory of Advanced Manufacturing Technology of Ministry of Education，Guizhou University，Guiyang 550025，China)

Abstract: Intelligent perception and automatic control in a complex unknown environment is one of the current research hotspots of robots in the field of control, and a new generation of artificial intelligence makes it possible to realize intelligent automation. In recent years, the new method of robot control using deep reinforcement learning in high-dimensional continuous state-action space has attracted the attention of relevant researchers. Firstly, the rise and development of deep reinforcement learning are first reviewed. The deep reinforcement learning algorithms for robot motion control are classified into two categories: value-based functions and policy gradients, and their typical algorithms and their related features are detailly described. Then, for the learning process before simulation to reality, five kinds of simulation platforms for robot motion control are briefly introduced, which are often used for deep reinforcement learning. Moreover, according to different types of research, the research progress of the deep reinforcement learning approach of robot motion control is expounded in five aspects, including autonomous navigation, object grasping, gait control, human-robot collaborative and multi-robot cooperation. Finally, the future challenges and development trends are summarized and anticipated.

Keywords: complex unknown environment；artificial intelligence；high-dimensional continuous space；deep reinforcement learning；simulation to reality；robot motion control

0 引 言

随着机器人在制造、服务、医疗以及军事等领域的应用拓宽,对其环境感知、行动决策以及自主学习等能力的智能程度需求也日益增长. 2017年,国务院在《新一代人工智能发展规划》中也提出对复杂环境下机器人自主控制等智能技术研究的迫切要

收稿日期: 2020-10-08；**录用日期:** 2021-03-03.

基金项目: 国家重点研发计划项目(2018AAA0101803)；国家自然科学基金项目(51475097, 91746116)；工信部资助项目(工信部联装[2016]213号)；贵州省科技计划项目(黔科合人才[2015]4011)；贵州省重点实验室建设项目(黔科合平台人才[2016]5103))；贵州大学培育项目(贵大培育[2019]22号).

†通讯作者. E-mail: lishaobo@gzu.edu.cn.

求. 人工智能正处于创新发展的大好时代,将为新一轮科技与产业革命汇聚发展注入新动能,同时也为新一代智能机器人实现智能感知与自动控制赋予更多可能. 而如今,利用深度强化学习(deep reinforcement learning, DRL)方法进行机器人运动控制正作为一个新兴研究领域吸引着越来越多研究人员的注意.

本文首先从算法和平台两部分对相关研究基础进行介绍,阐述DRL的兴起与发展,介绍用于机器人运动控制的典型算法及其特点;其次,对常用于DRL的机器人运动控制仿真平台进行简要介绍;然后,根据研究类型的不同,对基于DRL的机器人运动控制方法在自主导航、物体抓取、步态控制、人机协作以及群体协同等5个方面的研究进展进行综述;最后,结合DRL的发展方向,对其解决复杂未知环境下的机器人运动控制问题所面临的挑战以及未来发展趋势进行总结与展望.

1　深度强化学习

1) 强化学习. 受到任何环境下“适者生存”的生物启发,强化学习(reinforcement learning, RL)[1]利用试错机制与环境进行交互,旨在通过最大化累积奖励(return)的方式学习最优策略[2],其中“奖励”用于量化动作的价值. 不同于监督学习,强化学习没有人工标记等辅助手段,其学习过程中由环境提供的强化信号是对所产生的执行动作进行效用评估,而不是去指示智能体如何执行正确的动作. 由于外部环境直接提供的信息很少,强化学习当中决策的智能体(agent)必须通过与环境的试错性交互,不断地从行动-评价过程中优化控制策略以提升系统的控制性能. 因此,强化学习本质上是通过参数化的函数逼近“状态-动作”的映射关系,以求解决策问题的最优策略.

2) 深度强化学习. 强化学习受自身结构与学习能力的约束,多以解决低维问题为主,在处理高维连续状态-动作空间(以状态信息连续、执行动作连续、涉及维度庞大的机器人运动状态为主)下的控制问题时,难以有效求解,且无法通过人工设定对高维数据进行合适的特征表达. 因此,通过引入神经网络对特征进行有效表示,将高维连续的状态与动作进行离散降维,以达到降低任务复杂度、提高学习速度的效果,使得强化学习拓展至高维空间成为可能. 而早在DRL兴起之前,便已有学者开展了与机器人运动控制的相关工作,Benbrahim等[3]利用小型神经网络对输入的传感器信息进行预训练,以缩短RL算法在双足机器人行走任务中的学习时间;Moussa[4]将专家神经网络应用到基于RL的抓取控制中,通过对输入物体图像的特征学习,使机械爪具有与人类相似的感知与决策能力,但由于训练数据和计算性能的欠缺,这些工作仅利用浅层神经网络对高维度输入数据降维,以便于传统的RL算法对其进行处理.

随着GPU计算速度的大幅提升,以自动特征提取为主的深度学习(deep learning, DL)与传统强化学习方法之间的融合得以推进. 在2013年,Google人工智能研究团队DeepMind提出深度Q网络(deep Q-networks, DQN)算法[5],首次将深度神经网络(deep neural networks, DNN)与Q学习算法相结合,并利用该算法让计算机学习策略性游戏的玩法,超越了专业人类玩家的表现,同时也证明了深度强化学习方法强大的自主学习能力,使其迅速成为人工智能领域的研究热点. 而近年来,深度强化学习(DRL)方法融合深度学习(DL)的感知能力与强化学习(RL)的决策能力,在诸多挑战性领域均有广泛的应用,如自动驾驶、计算机视觉、医疗诊断以及机器人控制等. 其原理框架如图1所示,在处理一系列环境感知及控制决策问题时,其学习过程具有一定的通用性[6],可表示为: 1)智能体与环境交互时刻进行,并通过DL方法感知和观察高维度目标,得到当前环境下具体的状态信息;2)基于预期回报来评价各动作的价值函数(以此激励智能体),并通过RL方法得到某种适应性策略,将当前状态映射为相应的动作;3)环境对该动作做出相应反馈,智能体以此进行下一时刻的观察. 通过以上过程的不断循环,智能体最终可以得到完成既定任务的最优行动策略.

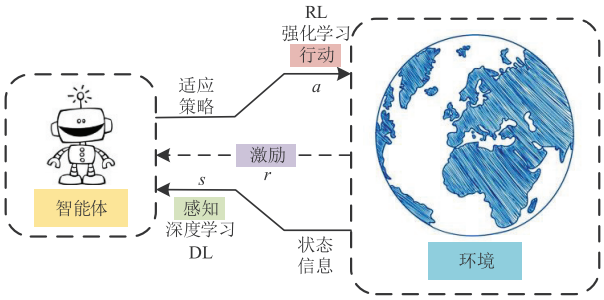

图1　DRL原理框架

2016年,深度强化学习方法开始逐渐应用于机器人运动控制领域,一些研究人员也将其引入多个现实任务中,Levine等[7]利用DRL方法对视觉感知和运动控制进行端到端联合训练,使机器人完成了对衣架、瓶盖等物体的特定放置任务;而面对密集人群中的机器人自主导航,Chen等[8]通过与注意力机制融合,提出基于改进DQN算法的移动路径规划方法,可令移动机器人根据实时图像信息获得控制策略,并在与人群交错时主动避让. 现如今,深度强化学习的研

究正处于快速发展阶段,本节接下来将用于机器人运动控制的DRL方法分为基于值函数和基于策略梯度两类,并针对各自典型算法进行详细介绍.

1.1 基于值函数

基于值函数的深度强化学习是用DNN逼近奖励值函数,以激励机器人等智能体获得最优行动策略,主要包括DQN及其改进方法.

1) DQN. 针对传统强化学习中依赖人工提取特征的问题,Mnih等[9]在Nature上发表了正式版DQN算法,旨在通过利用深度学习来自动提取海量输入数据的抽象表征,以此完成自我激励的强化学习,并优化控制问题的行动策略,其网络框架如图2所示. DQN训练过程中使用相邻的4帧游戏画面作为网络的输入,经过多个卷积层和全连接层,输出当前状态下可选动作的Q值,采用带有参数θ的卷积神经网络作为函数逼近器,并且定期从经验回放池中采样历史数据,利用随机梯度下降算法(stochastic gradient descent, SGD)更新网络参数,实现端到端的学习控制,而其在Atari视频游戏上也达到了人类职业玩家的控制效果.

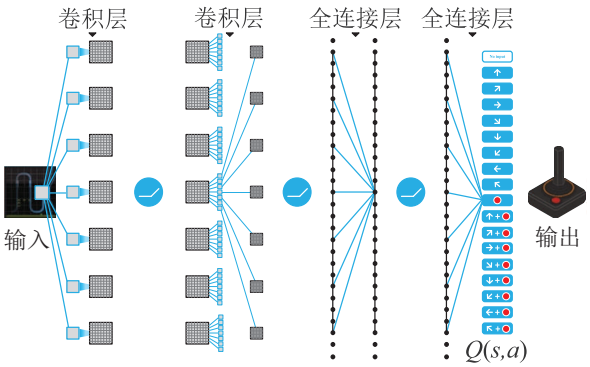

图 2 DQN网络结构[9]

DQN创新性地将深度神经网络和Q学习相结合,仅使用游戏的原始图像作为输入,并通过经验回放(experience replay)技术和固定目标Q网络,增加了历史数据的利用率,同时随机采样打破了数据间的相关性,有效解决了使用神经网络逼近非线性动作值函数所带来的不稳定和发散性问题,极大提升了强化学习的适用性. 此外,DQN通过截断奖赏和正则化网络参数,使梯度被限制到合适范围内,从而提升了训练过程的鲁棒性.

2) Double DQN. 在利用DNN所产生的目标Q值来近似表示值函数的优化目标时,DQN始终是选取下一状态中最大Q值所对应的动作,而在选择与评价动作过程中,其都是基于目标Q网络的参数,导致在学习过程中会出现过高估计Q值的问题.

Hasselt等[10]基于双Q学习算法(double Q-learning)和DQN提出双DQN(double DQN, DDQN)算法,与DQN训练流程相似,DDQN通过输入原始图像,经过多个卷积层以及全连接层,输出Q值以获得最优行动策略. 但与DQN选择和评价动作均基于同一个参数θ不同的是,DDQN将动作选择和策略评估分开,解决了DQN训练过程中过高估计动作值函数的问题. DDQN算法结构如图3所示,其借鉴双Q学习算法,对DQN算法进行改进: 在深度双Q网络中训练2个Q网络,采用在线Q网络参数θ来估计策略并以此进行将来动作的选择,而目标Q网络参数θ^-用来估计Q值以衡量动作价值,最终通过网络之间交替更新达到了解耦的效果.

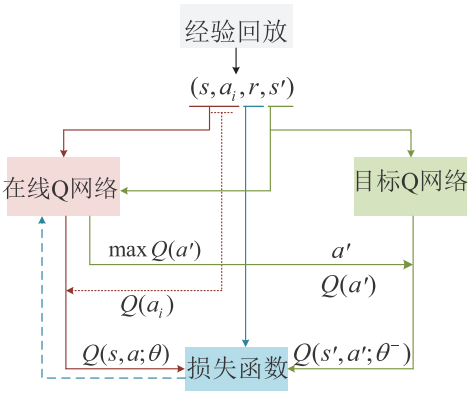

图 3 Double DQN算法结构示意图

3) Dueling DQN. 受不同动作的影响,在大多基于视觉感知的DRL任务中,状态与所对应动作之间的值函数不尽相同. 尤其是在某些特定状态下,值函数的大小与动作无关,便会出现价值与动作分离的问题.

为提升DQN算法对值函数近似估计的精确性,以获取更好的评估策略,Wang等[11]提出一种具有竞争网络结构的竞争深度Q学习(Dueling DQN)算法,其网络结构如图4所示. Dueling DQN保留原有DQN的卷积层以及全卷积层,并结合优势学习(advantage learning)的思想,将卷积层提取到的抽象特征分流到全连接层的两条支路中,分别代表状态值函数$V(s)$和某个状态下的动作优势函数$A(s,a)$,其中,动作优势函数$A(s,a)$指在状态s下,某动作a相对于平均状态动作而言的优势,用于衡量当前状态下各个动作的相对优劣程度. 最后,输出模块通过聚合操作将两条支路组合起来以得到各个动作的奖励值函数$Q(s,a)$. 由于此时构造的Q函数存在解不唯一的问题,Dueling DQN通过令贪婪动作选择时的优势函数为零,并将最大算子换成优势函数的平均值,以提高优化过程的稳定性. 尽管这种改进失去了V和A的

原始语义,但通过DQN与竞争网络结构的结合,使得智能体可在策略评估过程中更快地得到正确的行为.

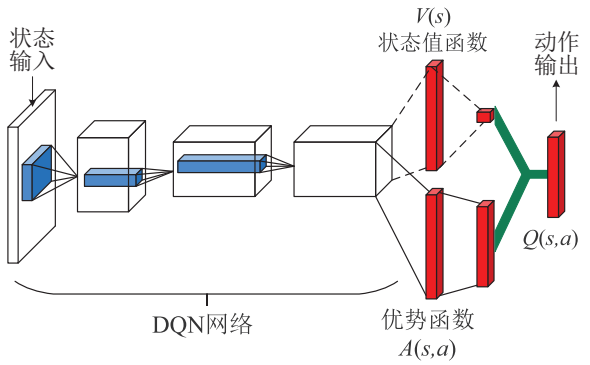

图 4　Dueling DQN网络结构图[11]

4) NAF. 虽然在以竞技类电子游戏为代表的离散控制问题上,DQN及其改进算法取得了优异的表现,但在动作空间较大或动作域连续的情况下,其最大化操作无法有效求解最优策略. 因此,面向离散控制的DQN算法难以应对真实环境中大量存在的高维连续控制问题.

2016年Gu等[12]提出一种归一化优势函数(normalized advantage functions, NAF)算法,以扩展DQN算法在连续控制问题中的应用. 与Dueling DQN的分流思路相同,NAF利用凸优势层对DQN算法进行修改,使用一个深度神经网络分别估计状态值函数$V(x)$与优势函数$A(x,u)$. 其中:利用优势函数衡量动作的质量,并通过特定层再将状态值函数与优势函数相结合,以此构建目标Q网络,$V(x)$和$A(x,u)$综合得到的目标值函数$Q(x,u)$可以继续用DQN算法解决问题. 得益于经验回放、目标网络以及优势更新,NAF在保证获得最大Q值的同时,通过利用输出的动作a更新连续变量Q值,其有效性在机器人操作类与行走类任务得到验证. NAF是基于值函数的DRL算法在连续控制问题上的首次尝试,但其设定过多约束和假设条件,目前并未提出一种在普遍意义上解决值函数与连续动作空间相容性的方法.

1.2　基于策略梯度

基于策略梯度的深度强化学习利用DNN逼近策略并利用策略梯度方法求得最优策略,其用于机器人控制的主要典型算法包括TRPO、PPO、DDPG和A3C.

1) TRPO. 传统策略梯度算法中更新步长的选取十分重要,其选取不当将会影响所得策略的优劣,尤其是使用DNN来表示策略,其策略的更新易受到影响.

为保证策略更新中策略性能的提升,受混合策略更新方法的启发,John等[13]提出置信域策略优化算法(trust region policy optimization, TRPO)对策略进行改善,使回报函数单调递增. TRPO首先使用单路径采样(single-path sampling)或蜿蜒采样(vine sampling)得到一系列状态动作对,通过蒙特卡洛方法估计得到Q函数值;然后,利用所得Q值对样本求平均得到优化问题中目标和约束的估计;最后,采用共轭梯度和线搜索方法近似解决约束优化问题,并以此更新参数,实现对机器人运动任务的最优控制. 但当策略选用深层神经网络表示时,会使问题求解面临十分庞大的计算量,这也是TRPO算法目前所存在的主要缺陷.

2) PPO. 为降低TRPO的计算复杂度,Schulman等[14]在其与行动者-评论家(actor-critic, AC)框架结合的基础之上提出了近端策略优化算法(proximal policy optimization, PPO). PPO算法减少了TRPO中采用共轭梯度和线搜索方法近似求解的过程,提升了算法的训练速度和可实施性,其算法迭代过程为先通过执行当前策略来估计优势函数,然后通过优化代理函数来更新策略参数. 同时,为了使新旧策略更新相对接近,PPO算法中采用了两种解决方法: 1)在目标函数中添加了剪切(clip)项,当新旧策略之间的更新偏移量超过预先的设定区间而获得更大的目标函数值时,剪切项将剪切代理目标,使策略更新被限制在一定区间内,以防止策略更新过快而无法收敛或收敛过慢; 2)引入惩罚(penalty)项,通过自动调整惩罚系数来限制KL散度值的大小,以减小新旧策略中动作概率分布之间的差异性,从而代替TRPO中约束对于策略更新幅度的限制. 另外,PPO算法采用1阶近似代替TRPO中的2阶泰勒展开方法,使其在解决深层神经网络表示中的大规模复杂问题上有很好的效果.

3) DDPG. 与TRPO算法中随机性策略输出的动作概率不同,确定性策略输出的是动作,无法进行环境探索,而通过环境交互进行学习是强化学习的必要过程,如何将确定性策略应用于强化学习中成为一个难以解决的问题.

为此,Silver等[15]提出一种基于AC框架的确定性策略梯度算法(deterministic policy gradient, DPG),利用离线策略(off-policy)即不使用相同策略进行动作选择和评估,使得确定性策略可支持智能体进行环境探索行为,其中评估策略使用确定性策略,动作策略使用随机策略. 随后,Timothy等[16]在DPG算法的基础上提出深度确定性策略梯度算法(deep deterministic policy gradient, DDPG),利用DNN逼近价值函数和确定性策略,使得其在连续任务上的应用

得以扩展,可以看作DQN与AC框架的结合,其算法结构如图5所示.

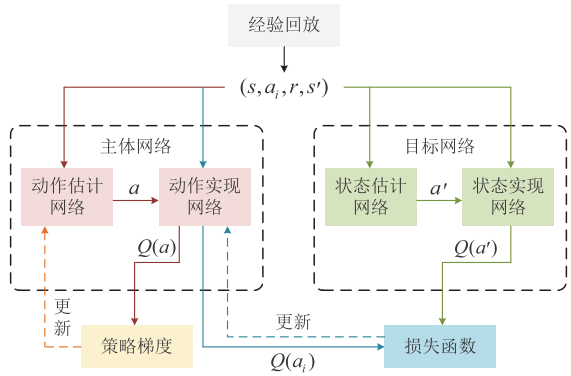

图5 DDPG算法结构示意[17]

DDPG包括4个神经网络,由主体网络与目标网络两部分组成,分别通过策略梯度和损失函数更新网络参数值以获得最优控制策略. 针对使用DNN逼近值函数而普遍存在的不稳定问题,其借鉴DQN的思路,在DPG算法中使用经验回放与目标网络,以减少数据间的相关性,并采取AC框架结构,用Critic最大化Q值,而Actor再利用Critic对动作的梯度进行学习,使得DDPG适用于解决机械臂运动等连续控制问题,具有较高的样本利用效率.

4) A3C. 针对值函数逼近过程中的不稳定性,除了经验回放机制以外,通过异步数据采集同样也可以解决. Mnih等[18]基于AC框架提出了异步优势动作评价算法(asynchronous advantage actor-critic, A3C),使用异步梯度下降法优化DNN控制器,其架构如图6所示. A3C通过子网络复制来收集样本计算的累积梯度,使得多个智能体实现并行训练,且子网络的更新参数与主网络共享.

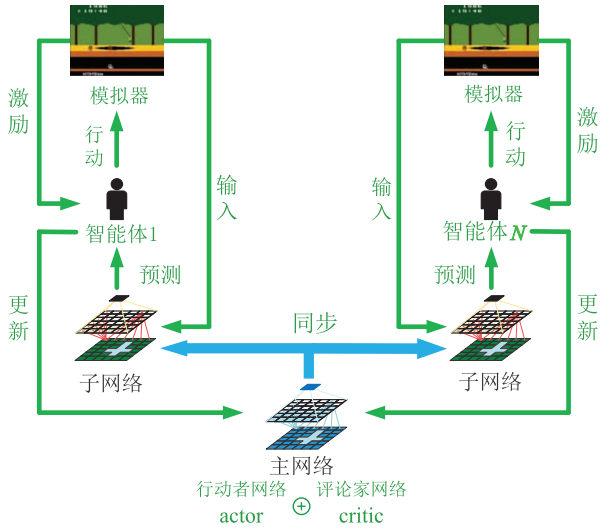

图6 A3C多线程并行架构[19]

多任务执行可以使算法在计算效率和样本利用率上都有所提高,A3C通过创建多个智能体,在不同环境中异步、并行学习,提高神经网络训练的稳定性. 其在行动者网络中使用多个线程并行运行,每个线程使用不同的探索策略,将样本收集的差异性最大化,以此降低数据间的相关性,并在策略的目标网络梯度中加入熵正则项,避免了过早收敛到次优策略. 多线程训练使得A3C缩短了训练时间,同时其在连续运动控制问题上取得不错效果,包括赛车游戏和随机3D迷宫导航任务,是目前较为通用和成功的DRL算法.

表1 用于机器人运动控制的DRL算法对比

算法		特点	AC框架	样本利用率	学习速度	应用		适用性
						仿真环境	真实环境	
基于值函数	DQN	经验回放	否	★★	★★	Atari Games[9]	机械臂运动控制[20]	离散空间任务,多用于较为单一的控制问题,如自主导航、物体抓取等
	Double DQN	双Q学习	否	★★★	★★	Atari Games[10]	机器人物体抓取[21]	
	Dueling DQN	竞争网络	否	★★	★★★	Atari Games[11]	移动机器人自主避障[22]	
	NAF	优势函数	否	★★★	★★★	MuJoCo Jaco机械臂控制[12]	机械臂操作控制[23]	
基于策略梯度	TRPO	置信域策略优化	是	★★	★★	MuJoCo 2D任务[13]	移动机器人自主导航[24]	连续空间任务,多用于大规模非线性控制问题,如人机协作、群体协同等
	PPO	近端策略优化	是	★★★★	★★★	Roboschool 机器人运动控制[14]	六足机器人步态控制[25]	
	DDPG	确定性策略梯度	是	★★★	★★★	MuJoCo 3D任务[16]	移动机器人自主导航[26]	
	A3C	异步梯度	是	★★★★	★★★★	MuJoCo Labyrinth[18]	AUV避障及目标追踪[27]	

通过以上对DRL各典型算法的介绍,如表1所示,基于值函数的DRL算法主要以改进DQN结构为主,在面对高维连续控制任务时,DQN及其改进算法若将连续动作变量进一步离散化,则会导致所涉及状态和行为的维数巨大,在如此高维空间下进行策略搜索举步维艰,且现有计算资源及能力也无法应对;而基于策略梯度的DRL算法改进则集中于采用AC框架、异步方式[17]以及策略优化方式,通过直接利用DNN逼近策略,更适合于连续控制任务. 显然,无论是结构优化或是算法改进,都使得算法的样本利用率和学习速度有所提高,而目前用于机器人运动控制的DRL方法仍以基于策略梯度为主.

2　仿真平台

仿真至现实(Sim2Real)是指将仿真环境中的运动策略迁移至真实环境中,而实现两者间的无差别转换是基于学习的机器人控制方法的最终目标. 仿真训练作为真实环境部署的“摇篮”,同时也是机器人自主学习与运行测试的首要步骤. 一般地,对于Sim2Real的主要流程(如图7所示),首先通过在仿真环境下训练学习以获取运动技能或控制策略(控制器),再将其引入真实环境,利用原始传感信息进行技能回放(或策略控制),使得机器人在复杂未知环境下作出最优的自主行动决策.相比于直接让实体机器人在真实环境中进行反复试错学习的不现实性,仿真训练为基于DRL的机器人运动控制方法提供了一种低成本、高效率、可观测的学习途径.

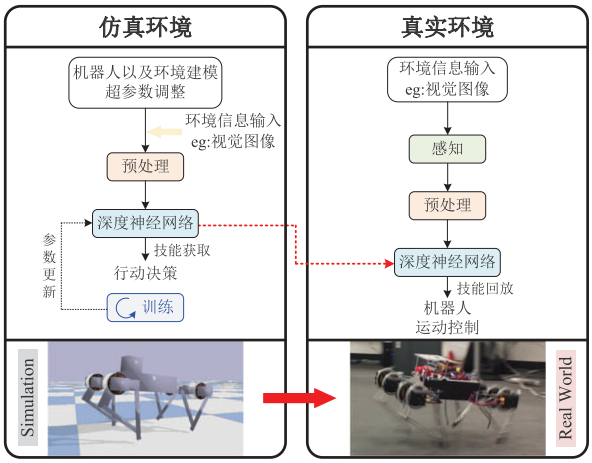

图7　Sim2Real流程

为此,本节将分别对Gazebo[28]、PyBullet[29]、MuJoCo[30]、V-REP[31]以及Webots[32]等5种常用于机器人运动控制中DRL方法研究的仿真平台进行简要介绍,其对应特点如表2所示.

表2　常用的机器人仿真平台对比

仿真平台	编程语言	操作系统	物理引擎	3D渲染引擎	ROS兼容性	适用场景	操作难度	是否开源	官网链接
Gazebo	C++/Python	Linux/Mac OSX	ODE/Bullet/Simbody/DART	OGRE	★★★★★	未知环境中移动机器人自主导航、多智能体交互控制等	★★★	是	http://www.gazebosim.org/
PyBullet	Python	Linux/Mac OSX/Windows	Bullet	TinyRenderer	★★★	机器人连续控制任务中的算法测试及验证	★★	是	https://pybullet.org/wordpress/
MuJoCo	C/C++/Python	Linux/Mac OSX/Windows	MuJoCo	OpenGL	★★	结构化环境下的机器人姿态控制以及动态力学分析	★★★★	是	http://www.mujoco.org/
V-REP	Matlab/C/C++/Python/Java	Linux/Mac OSX/Windows	ODE/Bullet/Vortex/Newton	Internal/External	★★★★	工业场景下的机器人运动规划,如物体抓取等	★★	是	https://www.coppeliarobotics.com/
Webots	Matlab/C/C++/Python/Java	Linux/Mac OSX/Windows	ODE	OGRE	★★★	仿生足类机器人运动仿真以及多机器人协同控制	★★★★	是	https://www.cyberbotics.com/

1) Gazebo. Gazebo是由南加州大学Nate等人开发的一种开源免费的高性能仿真平台,集成了机器人操作系统ROS和机器人平台PR2,可用于机器人的设计、开发、仿真以及测试. 相比于其他4个仿真平台,Gazebo与ROS的兼容性最好,通常与ROS配套使用.

Gazebo具有高质量的图形界面以及便捷编程窗口,可支持多种物理引擎,并使用图形渲染引擎ORGE为研究人员提供具有照明、阴影和纹理的高仿真室内外环境,能够模拟不同环境下机器人与其他对象间的交互,因此适用于人机协作、群体协同等多智能体交互控制. Gazebo内附多种机器人仿真模型,并支持用户导入自己建立的模型,同时提供丰富类型的传感器以模拟环境的反馈,使用户可以构建不同类

型的高仿真机器人运动模拟环境(包括地面、空中以及水下环境),主要用于移动机器人对未知环境的路径规划以及环境探索.此外,Gazebo除了用户端还为用户提供云服务端,使得控制算法在仿真机器人上进行快速测试与验证成为可能.

2) PyBullet. PyBullet实质是基于物理引擎Bullet开发的一个Python模块,旨在为用户提供免费的Sim2Real研究工具,可用于机器人、游戏、视觉效果以及机器学习的物理模拟,与V-REP和Webots可支持多类型编程语言不同,PyBullet目前只支持Python作为编程控制语言,但作为目前AI领域较为流行的语言,这并不限制相关研究者的使用.

PyBullet提供了各种仿真测试,例如正、逆向动力学和运动学以及碰撞检测,还包括机器人仿真模型,同时也支持用户导入URDF、SDF、MJCF等格式的仿真类加载文件,因此,适合于研究人员对新式机器人(如仿生机器人)设计开发,以及在虚拟环境中进一步训练与验证. 除了物理模拟之外,PyBullet还支持环境渲染,包含有一个CPU渲染器和OpenGL可视化,可支持虚拟现实技术的研究.凭借其使用以及操作便捷,PyBullet是目前深度强化学习中较为流行的仿真器,常用于机器人连续控制任务中的DRL算法测试及验证.

3) MuJoCo. MuJoCo是由华盛顿大学Emo等开发的一款3D仿真软件,本身是一种用于机器人、生物力学以及图形等研究领域的物理引擎,结合DeepMind开源的强化学习环境control suite[33],相比OpenAI Gym[34]更适合于连续控制任务. 当然,MuJoCo也支持用户单独使用以进行机器人仿真,但考虑到MuJoCo本身作为物理引擎,不同于V-REP等仿真集成平台,其在构建丰富的虚拟环境并集成标准的仿真工具(如运动库、路径规划器等)时通常比较麻烦,因此不适合复杂大规模环境下的仿真任务.

作为第1个基于模型优化(尤其是通过接触进行优化)而设计的模拟器,MuJoCo可扩展计算密集型技术,包括优化控制、物理一致性状态估计等,使得其可应用于具有丰富交互行为的复杂动态系统. 而在机器人运动仿真过程中,其具备关节防卡死、多约束、多驱动以及细节化仿真等特点,适用于机器人姿态控制及机械臂运动仿真,尤其是控制任务中的动态力学分析. 相比于Gazebo和V-REP中常用的ODE以及Bullet物理引擎,MuJoCo在简单的机器人运动仿真测试中具有更好的速度与精度[35],广泛用于传统应用(如电子游戏、物理机器人部署前的控制方案检验),同时也是DRL算法测试与验证的常用平台之一.

4) V-REP. 虚拟机器人实验平台(virtual robot experiment platform, V-REP)是由Coppelia Robotics公司开发的一款商业性质的通用机器人仿真软件. V-REP基于分布式控制架构:每个模型或对象都可以通过远程API、ROS节点或插件进行单独控制,并允许用户使用多类编程语言开发独立的应用程序. 但V-REP依赖于线程之间通信,对于需要大量数据的外部应用程序来说,其使用速度较慢. 因此,为方便开发任务的进行,大多数用户选择Matlab作为外部应用程序,进行仿真控制.

通过结合多种用于机器人仿真的内、外部库,V-REP向用户提供具有高度可扩展性的3D机器人集成开发环境. 而为提高建模精度,V-REP配备多个物理引擎,同时还提供了大量的机器人仿真模型,尤其是机械爪模型,相比其他仿真平台更为丰富,可供用户进行合适的选择,多用于运动规划类问题. V-REP支持用户自主建模并允许将控制器和功能插件嵌入仿真模型,简化了用户的开发任务与实现复杂性,相比于Gazebo和Webots,其具有友好的用户操作界面,更适合初学者使用,可用于快速算法开发、验证以及机器人运动建模与仿真.

5) Webots. Webots是由Cyberbotics公司开发的一种开源的3维机器人仿真平台,集成了物理引擎ODE,可用于模拟刚体动力学并提供属性(如质量、形状、纹理和形状等),支持用户自主设计CAD机器人模型等和自定义室内或户外仿真环境的属性配置.

不仅提供常用的机器人仿真模型,Webots还配置了多种可选择的仿真传感器和驱动器,可通过控制器接口与机器人操作系统ROS进行连接;并且向使用者提供编译器,也允许将控制器和功能模块嵌入至仿真模型中,这有利于降低开发人员实现仿真任务的复杂性,使得自行设计的机器人原型和涉及复杂数据处理的控制算法可以进行快速测试和验证. 而相较于Gazebo和V-REP, Webots在执行仿真任务时的资源占用较少[36],仿真交互时的编程方式也较为灵活,且在外部API调用方面也比V-REP更方便. Webots为用户进行机器人建模、控制编程和仿真模拟提供了完整的开发环境,其通用性强但上手难度较大,多用于仿生机器人运动仿真以及多机器人协同控制任务.

仿真阶段不仅为DRL算法测试提供了支撑工具,也为具体的机器人运动控制任务提供了大量仿真数据和学习环境,是赋予机器人在复杂动态环境中自

主行动力的关键环节. 通过介绍常用于DRL机器人运动控制的仿真平台,以此为相关研究人员及初学者在开展基于DRL的机器人运动控制研究时,提供一些适用于自身研究工作的选择与参考.

3　基于DRL的机器人运动控制研究进展

为了解基于DRL的机器人运动控制研究现状,本文以“深度强化学习+机器人控制”为主题,在CNKI、web of science、Google scholar、engineering village、 arXiv等中英文数据库以及中国知识产权局(CNIPR)、美国专利商标局(USPTO)中检索得到文献统计数据,以形成本文研究的样本库(见参考文献). 通过文献统计(如图8所示)发现,该方向的研究于2016年起逐年递增,其中不乏机器人研究领域中ICRA、IROS等国际顶级会议收录的论文.

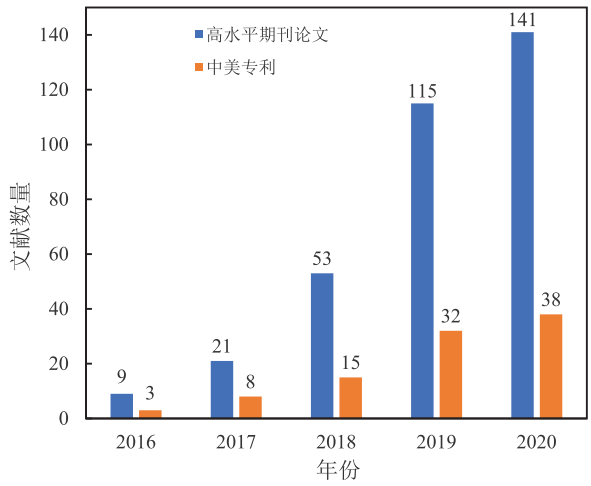

图8　“深度强化学习+机器人控制”主题文献统计

基于上述文献样本库,并根据研究类型及控制任务的不同,本节将从自主导航、步态控制、物体抓取、人机协作以及群体协同等5个方面对目前基于DRL的机器人运动控制相关研究进行综述.

3.1　自主导航

面对复杂未知的动态环境,移动机器人实现智能控制并自动完成规划任务的前提之一便是具备自主导航能力. 通过与环境之间的交互,机器人需要找到一条合适路径从当前位置移动到目标位置,避免与障碍物发生碰撞. 相比于传统方法,DRL通过对环境信息的有效利用,可在动态未知的环境下为机器人自主导航提供一种基于学习的低成本方案.

在不依赖地图信息的情况下,Li等[37]提出一种基于DQN和视觉伺服的移动机器人路径规划方法,以初始环境图像与目标图像为输入,通过训练建立彼此之间的对应关系和控制策略,以此完成室内自主导航任务,与传统视觉伺服方法相比,其具有较强鲁棒性以及泛化能力;而在具体的医疗诊断任务中,Hase等[38]同样使用DQN网络对机器人自主导航策略进行优化训练,通过实时超声图像信息引导KUKA机器人对人体进行超声波检测,初步实现在不同人体上对脊柱、骶骨等目标部位的自主搜寻,但两者所采用的DQN算法,只能使机器人输出有限的执行动作,其运动轨迹也并不平滑. 而王珂等[39]利用A3C算法对移动机器人进行路径规划,在通过运动学约束来优化状态空间搜索以及训练速率的同时,克服DQN的局限性,使机器人可以在连续动作域中输出平滑轨迹,但面对动态环境,依然缺乏移动避障的能力.

在复杂真实环境中,移动机器人应具备避障技能,这是完成其他复杂任务的先决条件. 对此,Fan等[40]提出一种基于DRL和PID控制器的混合移动避碰方法,通过利用激光雷达而非相机传感器进行环境状态的测量,在降低输入数据维度的同时,避免仿真环境与真实环境之间“差距”的扩大,使仅依靠仿真数据训练得到的避碰策略能较为稳定地部署至不同类型的真实机器人中;类似地,Shi等[41]通过利用额外奖励方式激励智能体探索陌生环境,同样将激光测距结果转换为运动控制策略以实现机器人无地图避障,使由内在好奇心驱动下的移动机器人自主导航,提升了仿真训练时的学习效率. 但由于目前大多研究工作都属于局部路径规划方法,这使得在自主导航过程中机器人移动路线规划并不一定是最优的.

3.2　物体抓取

物体抓取[42]是作业型机器人最为基础和普遍的运动形式之一,同样也是其获取、移动、运输目标物体的前提. 传统抓取控制以分析法(即硬编码方法[43])为主,通过运动学分析和物理模型建立,针对目标物抓取进行编程控制. 而在面对未知且种类繁多的物品时,进行相应程序的逐一编写并不现实,基于DRL的机器人物体抓取控制方法,通过学习与探索最优策略以完成一系列抓取行为,可减少或完全去除手动建模的繁琐计算,使机器人在抓取任务中可进行自主学习.

为了验证深度强化学习方法在机器人抓取任务中的有效性,Andrea等[44]通过对简单目标及任务的设定,在仿真环境中训练机械臂对物体拾取和放置任务的操作策略,展示了NAF、TRPO以及DDPG算法在连续控制任务中的可行性,但在真实环境测试过程中,机械臂需要依靠识别人工标记的特征信息才能完成对目标物体的抓取,且抓取环境比较单一;为进一步提升其在杂乱环境下自主抓取的决策能力,Ahn

等[45]提出一种结合DQN和目标检测算法的物体抓取技术,利用高、低级分层的控制结构分别对动作选择和执行策略进行有序训练,使得智能体能够更为有效地学习,而其中的目标识别过程,则采用Mask R-CNN算法[46]进行实例分割(区分目标与周围障碍物),以确定物体形状及位置信息,尽管其通过使真实世界的物体状态与仿真训练过程中的输入信息相似,缩小了两者之间的差异,但受到机械臂初始抓取姿态的固有设定所限制,即便周围障碍物排列规整,可能也会导致目标物抓取失败.

不同于上述两项工作,在实际抓取场景当中可能存在目标距离较远或者目标物体处于移动状态的情况. 对此,考虑到目标距离超出抓取范围时需要进行移动抓取,Wang等[47]搭建了一种基于PPO算法的移动机器人抓取系统,包括机械手、移动基座和视觉系统,利用深度目标姿态估计算法(deep object pose estimation, DOPE)[48]进行目标信息感知,并结合机器人当前状态作为系统输入来指导机器人进行自主移动及物体抓取,而仿真训练获得的移动抓取策略在实体机器人上也有较为稳定的迁移效果,但由于该系统只在移动基座上配置了车载摄像头,当移动至目标较近距离时会导致机械手无法进行局部检测,使得抓取物被遮挡并随之出现两者间运动解耦的情况;而对于可移动目标物抓取的高维连续控制问题,Du等[49]提出一种基于DDPG算法的空间目标捕获控制方法,通过随机运动获取状态信息,较短时间内对仿真数据进行预训练,在不更新网络参数的同时将所得的状态动作序列存储到经验回放区,提升了实际训练过程中的学习效率,相比于传统方法,其无需手动建模,这极大降低了对复杂控制器设计的工作量,但由于缺乏非结构化环境下动态目标捕获的真实样本信息,目前仍停留在仿真阶段.

3.3 步态控制

与传统轮式机器人相比,仿生足类机器人在山地丛林等复杂地形环境下具有更灵活的行走能力、越障能力以及广阔的工作空间. 步态控制(gait control)则是令仿生足类机器人拥有行动能力的关键技术之一,传统步态控制方法通常需要将接触点选择、轨迹优化以及操作空间控制等工作分模块进行,且面对不同运动任务需要调整设计,而DRL也为解决机器人步态控制中复杂动力学建模和运动轨迹跟踪精度不足等问题,提供另一个可行的解决方案.

目前,由于仿生足类机器人的类型丰富,对其DRL步态控制方法研究也较为广泛. 在仿人机器人方面,施群等[50]通过随机引入离散动作来采集训练数据,并初步建立姿态辨识模型以作为离线估计器,使得实体机器人在DDPG网络中进行在线学习时具备先验知识的指导,在提高样本利用率与训练效率的同时,相比于传统PID与MPC控制方法,人形机器人的姿态跟踪控制偏差较小,在平滑至障碍路段中步行具有相较稳定的效果;与前者思路相似,Xi等[51]利用分层高斯过程(hierarchical Gaussian processes,HGP)算法[52]对样本数据进行离线预训练,以获得初始运动估计器,再基于DDPG算法对步态控制策略进行在线优化,使双足机器人在仿真动态平台上得以稳定行走. 为提升样本量和训练速度,上述研究都采取了“离线训练-在线优化”的分阶段学习模式,但由于前后两阶段之间的交互数据存在分布不匹配的问题,使得在线学习可能需要耗费大量时间去寻找离线训练过程中相对应的最优控制策略.

除了稳定行走以外,其他多类型的步态学习主要集中在仿生多足机器人的研究上. 针对四足仿生机器人,Tan等[53]提出一种基于PPO算法的敏捷步态学习方法,将运动控制任务转换为马尔可夫决策过程,通过增强虚拟环境的逼真度以及场景随机性,使得仿真训练阶段所获取的步态控制器在实体机器人上具有较好的鲁棒性,但由于设置的奖励机制较为简单,机器人无法在运动过程中进行速度调整,在复杂地形结构中也并不适用. 尽管在集中式控制策略中,由单一智能体执行运动决策避免了多足之间的协调问题,但由于任务固有的难度(尤其面对高维连续状态-动作空间),其不能够快速地训练以及收敛. 为此,Sartoretti等[54]采取分布式的运动控制方法,将铰接式六足机器人的各足视为单一智能体,利用A3C多线程并行训练的优势,使多智能体集中学习、分散控制,在与环境交互过程中各个智能体的运动参数会根据共享奖励来更新,以实现机器人整体的爬行运动. 该方法的有效性也在结构化环境中进行了验证,但由于各足之间在无交互状态下进行协作,在执行复杂任务时无可避免地会出现运动解耦问题. 而面对周期性任务,Zhang等[55]提出一种基于MDGPS算法[56]的滚动步态学习方法,通过引导策略搜索优化高维参数空间下的策略梯度计算,将任务分散并先进行局部采样以学习滚动行为,再以此引导张拉球机器人(tensegrity robot)对连续滚动步态的全局最优策略的搜寻,相比于开环策略和人工设计控制器具有更优的学习效率与泛化性. 但综上分析,现阶段的研究都主要以机器人自身状态信息来输出控制,而缺乏对外

部环境信息的有效利用,使得其在运动过程中不具备移动转向及避障等能力.

3.4 人机协作

人机协作(human-robot collaborative, HRC)[57]的长期目标即是提高协同工作效率,以顺应工业4.0的发展趋势. 其利用人的灵活性与机器人的高效性,旨在让机器人与人类更为紧密地合作,协同高效地完成复杂任务. 但面对一些复杂任务,由于人的运动轨迹不规律性,人机协作模型难以建立,导致传统控制方法的协作效率不高、稳定性差. 而针对人体控制策略建模,DRL方法通过预测人类某个状态时刻可能执行的动作,为人机协作的机器人运动控制建立较为直观的认知模型,使人机协作更接近人类之间的互动.

与既定任务中的机器人单体执行不同,基于DRL的人机协作方法按学习顺序的不同可分为: 1)任务优先; 2)人类优先; 3)共同学习. 而为了探索3种不同方式的学习效果,Tao等[58]将人机协作任务分解为任务学习与人类行为学习两部分,通过设置对应的分层奖励机制来判别人机协作任务中的团队表现,并考虑训练时长以及人员参与程度来区分不同学习方式下的优缺点. 若训练时长较为重要,则机器人应当首先学习人类操作; 若优先考虑团队表现和人员参与,则应先学会执行任务的最优行动策略; 而由于对人类对象的过于依赖,在面对复杂任务时人机共同学习[59],与前两者相比不具备优势,这也为相关研究者提供了不同训练策略的参考. 但由于未考虑人-机任务执行之间的分工,使其缺乏对协作过程中机器人是否具备伙伴意识与自我意识的判别.

进一步地,将具体协作过程中不同角色分配给人与机器人,又可由协作关系分为2类: 1)人机主从; 2)人机平等. 在人机协作的主从关系下,大多数研究集中在以人为中心的“人主-机从”协作模式. 为了提升人机协作过程中的安全性,Clegg等[60]采用PD控制器来模拟人类接受辅助穿衣的动作策略,而机器人辅助穿衣的行动策略则由TRPO算法利用电容式传感而非视觉传感信息进行训练与输出,通过协同优化这两者之间的策略,使得在保证辅助穿衣过程的正常进行以外,避免了遮挡状态下机器人与人体接触较近而发生不当行为,但受惩罚机制和动作缩放的限制,当人类出现不确定行为时,协作任务完成所需的时间会相对较长. 为进一步平衡执行过程中任务耗时与安全性之间的关系,Ghadirzadeh等[61]通过运动捕捉来提高机器人对人类行为认知能力以避免协作延迟,并构建行为树作为先验知识,以此在深度循环Q网络中进行物品包装任务的学习,增强了其在协助过程中的伙伴意识以及主动性. 但上述两项研究都采取了共同学习的方式,存在着当协助对象变更后机器人适应性较差的问题.

在人机平等协作方面,考虑到任务整体的协作效率,金哲豪等[62]通过人的随机行为采集数据并以高斯过程回归(Gaussian process regression, GPR)方法来拟合人为控制预测模型,使机械臂利用先前所学习到的人类行为认知结果作为DDPG算法建立球杆系统MDP的一种环境状态信息,来学习人机协作过程中的自适应控制策略,以人机平等的形式提升了机器人在协作过程中的自我意识. 但由于受到环境噪声以及对人为操作预测的误差影响,人机协作时球杆系统控制存在一定的不稳定性.

3.5 群体协同

群体协同(multi-robot cooperation, MRC)[63]相比于单体机器人,在面对多变的外部环境时具有较强鲁棒性,适用于执行更加复杂多样的耦合型任务. 但多机器人之间的相互影响对协同控制系统的同步性要求很高; 同时,关节约束、密集障碍识别与规避、协作伙伴间信息传递等问题对于群体协同控制的发展提出了挑战. 而DRL方法通过减少传统协同控制中的人为分析以及推导建模过程,为多机器人之间的群体协同控制开辟了一条新式发展道路.

群体协同控制与人机协作的不同点在于,其不存在对协作伙伴行为认知与预测的难点,可看作是在执行任务时对单体机器人数目的扩增,通过多机器人协同作业的形式,突破单体行动能力以及范围限制,以扩展机器人的应用领域. 而群体协同下的多机器人运动控制DRL方法研究主要分为集中式和分布式两类,集中式方法是将运动控制定义为优化问题,Bae等[64]以传统路径规划算法下的最短路径搜索时长作为优化目标,利用卷积神经网络提取图像特征,并作为全局信息输入多层DQN训练网络中,使多机器人在共享权重的同时,进行最优移动路径的探索. 相比传统路径规划方法,具有较好的环境适应性,但不具备避障能力,且在一定学习进度以后,机器人之间的学习效果相同,导致进一步学习的效率较慢. 与集中学习的方式不同,分布式方法可使多机器人并行学习,让各自独立选择所要执行的动作,Semnani等[65]采用改进A3C算法作为分散学习网络,并额外设定碰撞惩罚机制来增强避障性能,而为了避免机器人之间运动过于保守所导致的卡死现象,其利用FMP算法[66]根据各自的相对位置信息来调节机器人的运动

速度,降低了拥堵环境下多机器人移动碰撞概率,但由于未考虑复杂的运动约束问题,使得离散动作输出下的机器人运动灵活性较差. 同时,基于DRL的群体协同控制也受到机器人之间通信交互的限制[67],而使得机器人通常只能根据局部信息作出决策,导致在部分可观测环境中的协作性能不佳.

而针对群体协同的具体任务,Zhang等[68]将领土防御任务看作零和博弈问题,以入侵者与领土间的距离信息作为奖励,使多个防御机器人在DQN网络中学习联合防御策略,协同任务的目标则是保持领土与入侵者之间的距离相对安全,其有效性在简单的拦截实验中得以验证,但需要依靠入侵者所携带的GPS信息来判断领土安全性,一旦入侵者位置信息阻断,协同系统将无法生效. 在目标包围任务中,Ma等[69]利用基于AC框架的DDPG网络对多个智能体进行并行训练,通过局部伙伴信息(包括包络半径、角速度以及机器人之间的相邻距离),使得分布式学习方式下各机器人可自主进行策略搜索并作出相应独立决策,而优先级奖励函数的设置可对多机器人包络问题进行约束,以在其协同包围时既保持队形也能实现避障. 不足的是,由于只涉及到二维连续空间的协同控制,对于实际当中的高维非线性控制问题有待进一步研究. 而在协调编队任务中,Wang等[70]首次提出一种基于DRL的多机器人编队控制方法,利用自动编码器对可观测的高维状态空间信息进行压缩表示以加快学习进度,并以预定模式的拟合程度作为奖励,使多机器人在观察(其他伙伴位置信息)、计算(自身移动方向及速度)以及行动的反复试错过程中逐步学习到预设编队模式的行动策略,相比于现有方法具有更好的编队效率,也证明了DRL方法在连续模式形成问题上的可行性,但当协作机器人的原有数目发生变化时,先前训练好的模型不再适用且需要重新训练,这也是大多数群体协同控制方法所需面临的问题.

4 挑战与展望

综上所述,深度强化学习凭借自身优势受到了研究人员的广泛关注,虽然已初显成效,但作为机器学习的一个新兴领域,其正处于发展阶段,在解决实际中多领域任务的机器人运动控制问题时,仍面临许多挑战:

1) 样本数据不足. 由于机器人在真实环境下的高维数据采集困难,其可训练数据不足,致使基于数据驱动的DRL方法在特征提取、参数优化等方面的优势发挥并不充分.

2) 学习效率较低. 面对复杂未知环境,奖励机制的设置不合理会导致在策略寻优时出现稀疏奖励问题,进而影响机器人在部分可观测性环境下探索的学习效率以及积极性.

3) 泛化能力不强. 目前,大多研究仍停留在仿真与测试环节,不仅是训练模型在仿真至现实时的迁移性能存在一定偏差,且很难将相同任务的执行策略应用至不同实体机器人之中.

4) 推理认知能力弱. 现阶段的DRL方法还只是依靠大量数据训练,拟合得到运动模型,虽然缓解了机器人传统控制方法难以建模的问题,但其智能程度仍停留在计算智能.

5) 环境交互通道单一. DRL依赖于大量环境信息的交互,尤其是在面临动态未知环境时,机器人需要进行频繁的信息交互才能作出较优决策,而目前研究主要局限于单一传感信息,一旦环境观测阻隔,任务的执行便会中断.

针对以上问题,并结合深度强化学习发展趋势,对DRL中机器人运动控制的未来发展及研究方向作出如下几点总结:

1) DRL与模型结合. 现阶段以model-free为代表的DRL方法仅依靠数据驱动来实现最优控制策略的学习,虽然在处理已知静态任务时具有较好表现,但容易直接忽略被控对象的动力学特征. 智能动力学模型是未来机器人实现自主控制的关键之一,在面对未知动态任务,model-based的DRL方法通过模型去理解外部世界的因果关系,可将一般学习问题转化为优化问题,无需更多次的探索,便可以通过模型预测未知状态值,进而快速地对值函数进行评估或者直接优化策略,以提高样本数据利用率,相比基于无模型的方法具有更好的泛化能力和稳定性,必定是未来基于DRL的机器人控制方法的主要发展方向之一.

2) 创新奖励方式. 奖励函数是衡量策略性能的重要指标,而其大多数是由先验知识人为给定,并在实验过程中不断调整选优. 在结构与任务单一的问题中,奖励函数的给定较为容易. 但在相对复杂的复合任务中,奖励函数并不是显式的,人类也不能给出明确标准的奖励函数. 而人为地添加一些监督信号(如内在奖励或辅助任务),可在一定程度上提高机器人在具体任务环境下的探索能力. 此外,逆向强化学习(inverse reinforcement learning, IRL)、模仿学习(imitation learning, IL)以及生成式对抗网络(generative adversarial networks, GAN)分别从奖励函数表达形式(深度神经网络)、学习方式(专家示范)以

及无奖励方式(对抗形式)等不同角度,为奖励方式的创新提供了一些新思路.

3)通用控制策略. 目前,基于DRL的机器人运动控制研究取得了一定的进展,在经过仿真环境中的训练学习,合适的控制策略可以转移至真实环境下,实现机器人对于特定任务的执行. 尽管在面对“现实差距(reality gap)”,相关研究人员着重于增强虚拟环境逼真度、场景随机化以及添加环境噪声等,以增强模型鲁棒性,但不足以适应各种环境与任务,而通过结合迁移学习(transfer learning, TL)、元学习(meta learning, ML)或持续学习(continual learning, CL),研究学习策略的转移以及已学知识的再利用,使机器人不必从零开始学习来提升DRL模型的通用性,有待后续研究人员的关注.

4)高效学习机制. 在执行长时间任务时,基于学习的机器人控制方法容易在训练阶段耗费大量时间并陷入探索的“死胡同”,因此仅依靠计算能力提升的学习并不是真正意义上的高效学习. 新式DRL方法也正在不断创新与发展,图神经网络(graph neural network, GNN)、记忆神经网络 (memory neural networks, MNN)、多智能体DRL (multi-agent DRL, MADRL)以及分层DRL (hierarchical DRL, HDRL)分别从增加记忆组件、协同学习与任务分解等方向,提高了机器人运动控制中DRL模型的记忆推理能力和自主学习效率,而如何分散学习并聚合多机器人训练所收集到的环境及行动决策信息,增强机器人学习能力的有效性、延续性,进而使所学知识具有记忆性和可拓展性,将是其未来发展的主要推动力.

5)多模态融合. 现有绝大多数基于学习的机器人运动控制研究都是以单一视觉传感图像进行数据输入、特征学习和策略输出,但在执行人类社会中更深层次的任务(如情感识别、人机交互等)时,仅通过单一模态的机器人运动控制无法满足未来所需. 而通过整合多种交流模态信息(如听觉[71]、触觉[72]以及自然语言等)进行多通道表示学习,可增强基于DRL的机器人运动控制的稳定性、准确性以及任务执行的高效性、多样性. 但由于多模态表示存在数据间的异质性和模式间的偶然性,如何解决多模态信息数据采样、多模态对齐与融合以及多流时间建模等问题,实现深度多模态表示学习(multimodal representation learning, DMRL)[73]与DRL在机器人控制领域的融合,将是该方向的主要发展趋势之一.

此外,通过目前的文献调研来看,基于DRL的机器人运动控制研究还存在一定的局限性:

1)所执行的任务过于单一. 大多数研究仅使得机器人具备较为基础的运动能力,但这些都只能针对性地完成单一任务,考虑到实际任务的复杂性,如何使机器人在陌生环境下学会“摸爬滚打”,在运用一系列运动技能来完成任务的同时,具备额外的能力以避免执行过程的意外终止,将是研究者关注的热点之一.

2)所涉及的行动领域过于单一. 众多研究者将机器人的行动局限于地面范围,如何扩展机器人的行动范围而非限制活动区域,如水下探测机器人、地下搜救机器人以及空中作业机器人的运动控制研究,并使单体机器人具备多领域的运动能力,是未来DRL机器人运动控制研究需要解决的问题.

3)运动控制的各项指标难以平衡. 基于DRL的机器人运动控制作为学习式方法,必然受到学习过程(训练效率)和执行过程(安全性、完成程度)中的因素影响,而如何平衡其在训练效率、行动安全性以及任务执行度等因素之间的关系,也是相关研究工作需要考虑的问题;

4)群体协同作业局限于同类多机器人协同. 目前群体协同的大部分研究以扩展单体机器人的数目为优势,通过群体活动来提升执行任务的效率,而如何克服不同类型机器人之间的模型异质性,利用DRL方法学习并实现异类多机器人协同控制,以进一步提升执行任务的多样性,同样是未来可研究方向之一.

5 结　论

深度学习与强化学习的结合使机器人的感知与决策能力得到大幅提升,利用深度强化学习方法解决高维连续状态与动作空间下的机器人运动控制问题,已使得机器人初步脱离人类的全过程操控,并能够相对“独立”地完成既定任务,在一定程度上提高了机器人的学习能力和智能化水平. 而目前基于DRL的机器人运动控制研究正处于快速发展的阶段,无论是在学术研究或是工程运用方面都有较多空间亟待拓展. 本文主要从研究基础和研究现状两方面展开综述,在研究基础方面,详细介绍了常用于DRL机器人运动控制的典型算法、仿真平台及其各自特点;而在研究现状方面,则根据研究类型的不同对现阶段研究进展进行了归纳、分析与总结,并结合目前所面临的挑战,对其未来发展趋势进行了展望,以期为相关研究人员提供详实有效的参考.

参考文献(References)

[1] Naeem M, Rizvi S T H, Coronato A. A gentle introduction to reinforcement learning and its application in different fields[J]. IEEE Access，2020, 8: 209320-209344.

[2] 陈宗海, 杨志华, 王海波, 等. 从知识的表达和运用综述强化学习研究[J]. 控制与决策, 2008, 23(9): 961-968.
(Chen Z H, Yang Z H, Wang H B, et al. Recent researches on robot autonomous grasp technology[J]. Control and Decision, 2008, 23(9): 961-968.)

[3] Benbrahim H, Franklin J A. Biped dynamic walking using reinforcement learning[J]. Robotics and Autonomous Systems, 1997, 22(3): 283-302.

[4] Moussa M A. Combining expert neural networks using reinforcement feedback for learning primitive grasping behavior[J]. IEEE Transactions on Neural Networks, 2004, 15(3): 629-638.

[5] Volodymyr M, Koray K, David S, et al. Playing atari with deep reinforcement learning[EB/OL]. (2013-12-19)[2020-09-01]. https: //arxiv.org/pdf/1312.5602.pdf.

[6] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
(Liu Q, Zhai J W, Zhang Z C, et al. A survey of deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.)

[7] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies[J]. Journal of Machine Learning Research, 2016, 17(39): 1-40.

[8] 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. Vancouver: IEEE, 2017: 1343-1350.

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

[10] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning[C]. Association for the Advancement of Artificial Intelligence. Phoenix: AAAI, 2016: 2094-2100.

[11] Wang Z, Tom S, Matteo H, et al. Dueling network architectures for deep reinforcement learning[C]. The 33rd International Conference on Machine Learning. New York: International Machine Learning Society, 2016: 1995-2003.

[12] Gu S, Timothy L, Ilya S, et al. Continuous deep Q-learning with model-based acceleration[C]. The 33rd International Conference on Machine Learning. New York: International Machine Learning Society, 2016: 2829-2838.

[13] John S, Sergey L, Philipp M, et al. Trust region policy optimization[C]. The 32nd International Conference on Machine Learning. Lille: International Machine Learning Society, 2015: 1889-1897.

[14] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization Algorithms[EB/OL]. (2017-08-28)[2020-09-01]. https://arxiv.org/pdf/1707.06347.pdf.

[15] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms[C]. The 31st International Conference on Machine Learning. Beijing: International Machine Learning Society. 2014: 387-395.

[16] Timothy P L, Jonathan J H, Alexander P, et al. Continuous control with deep reinforcement learning[EB/OL]. (2019-07-05)[2020-09-01]. https://arxiv.org/pdf/1509.02971.pdf.

[17] 多南讯, 吕强, 林辉灿, 等. 迈进高维连续空间:深度强化学习在机器人领域中的应用[J]. 机器人, 2019, 41(2): 276-288.
(Duo N X, Lv Q, Lin H C, et al. Step into high-dimensional and continuous action space: A survey on applications of deep reinforcement learning to robotics[J]. Robot, 2019, 41(2): 276-288.)

[18] Mnih V, Adrià P B, Mehdi M, et al. Asynchronous methods for deep reinforcement learning[C]. The 33rd International Conference on Machine Learning. New York: International Machine Learning Society, 2016: 1928-1937.

[19] Babaeizadeh M, Frosio I, Tyree S, et al. Reinforcement learning through asynchronous advantage actor-critic on a GPU[EB/OL]. (2017-03-02)[2020-09-01]. https://arxiv.org/pdf/1611.06256.pdf.

[20] Fangyi Z, Jürgen L, Michael M, et al. Towards vision-based deep reinforcement learning for robotic motion control[EB/OL]. (2015-11-13)[2020-09-01]. https://arxiv.org/pdf/1511.03791.pdf.

[21] Shariq I, Jonathan T, Thang T, et al. Toward sim-to-real directional semantic grasping[EB/OL]. (2020-03-05)[2020-09-01]. https: //arxiv.org/pdf/1909.02075.pdf.

[22] Xie L H, Wang S, Markham A, et al. Towards monocular vision based obstacle avoidance through deep reinforcement learning[EB/OL]. (2017-06-29)[2020-09-01]. https://arxiv.org/pdf/1706.09829.pdf.

[23] Gu S, Holly E, Lillicrap T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]. IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017: 3389-3396.

[24] Mingming L, Rui J, Shuzhisam G, et al. Role playing learning for socially concomitant mobile robot navigation[EB/OL]. (2017-05-29)[2020-09-01]. https://arxiv.org/pdf/1705.10092.pdf.

[25] Qin B, Gao Y, Bai Y. Sim-to-real: Six-legged robot control with deep reinforcement learning and curriculum learning[C]. The 4th International Conference on Robotics and Automation Engineering. Singapore: IEEE, 2019: 1-5.

[26] Jesus J C, Bottega J A, Cuadros M A S L, et al. Deep deterministic policy gradient for navigation of mobile robots in simulated environments[C]. The19th International Conference on Advanced Robotics. Belo Horizonte: IEEE, 2019: 362-367.

[27] Cao X, Sun C Y, Yan M Z. Target search control of AUV in underwater environment with deep reinforcement learning[J]. IEEE Access, 2019, 7: 96549-96559.

[28] Koenig N, Howard A. Design and use paradigms for

Gazebo, an open-source multi-robot simulator[C]. The IEEE/RSJ International Conference on Intelligent Robots and Systems. Sendai: IEEE, 2004: 2149-2154.

[29] Erwin C, Yunfei B. Pybullet, a Python module for physics simulation for games, robotics and machine learning[EB/OL]. (2020-06-03)[2020-09-01]. http://pybullet.org.

[30] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control[C]. The IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura: IEEE, 2012: 5026-5033.

[31] Rohmer E, Singh S P N, Freese M. V-REP: A versatile and scalable robot simulation framework[C]. The IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo: IEEE, 2013: 1321-1326.

[32] Michel O. Webots: Professional mobile robot simulation[J]. International Journal of Advanced Robotic Systems, 2004, 1(1): 39-42.

[33] Yuval T, Yotam D, Alistair M, et al. DeepMind control suite[EB/OL]. (2018-01-02)[2020-09-01]. https://arxiv.org/pdf/1801.00690.pdf.

[34] Greg B, Vicki C, Ludwig P, et al. OpenAI gym[EB/OL]. (2016-06-05)[2020-09-01]. https://arxiv.org/pdf/1606.01540.pdf.

[35] Erez T, Tassa Y, Todorov E. Simulation tools for model-based robotics: Comparison of Bullet, Havok, MuJoCo, ODE and PhysX[C]. The IEEE International Conference on Robotics and Automation. Seattle: IEEE, 2015: 4397-4404.

[36] Ayala A, Cruz F, Campos D, et al. A comparison of humanoid robot simulators: A quantitative approach[C]. The 10th International Conference on Development and Learning and Epigenetic Robotics. Valparaiso: IEEE, 2020: 1-6.

[37] Li Y, Jana K S. Learning view and target invariant visual servoing for navigation[C]. The IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 658-664.

[38] Hase H, Azampour M F, Tirindelli M, et al. Ultrasound-guided robotic navigation with deep reinforcement learning[EB/OL]. (2020-04-07)[2020-09-01]. https://arxiv.org/pdf/2003.13321.pdf.

[39] 王珂, 卜祥津, 李瑞峰, 等. 景深约束下的深度强化学习机器人路径规划[J]. 华中科技大学学报: 自然科学版, 2018, 46(12): 77-82.
(Wang K, Bu X J, Li R F, et al. Path planning for robots based on deep reinforcement learning by depth constraint[J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2018, 46(12): 77-82.)

[40] Fan T, Long P, Liu W, et al. Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios[J]. International Journal of Robotics Research, 2020, 39(7): 856-892.

[41] Shi H, Shi L, Xu M, et al. End-to-end navigation strategy with deep reinforcement learning for mobile robots[J]. IEEE Transactions on Industrial Informatics, 2020, 16(4): 2393-2402.

[42] 刘亚欣, 王斯瑶, 姚玉峰, 等. 机器人抓取检测技术的研究现状[J]. 控制与决策, 2020, 35(12): 2817-2828.
(Liu Y X, Wang S Y, Yao Y F, et al. Recent researches on robot autonomous grasp technology[J]. Control and Decision, 2020, 35(12): 2817-2828.)

[43] Wang C, Zhang X, Zang X, et al. Feature sensing and robotic grasping of objects with uncertain information: A review[J]. Sensors, 2020, 20(13): 1-30.

[44] Andrea F, Elisa T, Nicola C, et al. Robotic arm control and task training through deep reinforcement learning[EB/OL]. (2020-05-06)[2020-09-01]. https://arxiv.org/pdf/2005.02632.pdf.

[45] Ahn K H, Song J B. Image preprocessing-based generalization and transfer of learning for grasping in cluttered environments[J]. International Journal of Control, Automation, and Systems, 2020, 18(9): 2306-2314.

[46] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. The IEEE International Conference on Computer Vision.Venice: IEEE, 2017: 2980-2988.

[47] Wang C, Zhang Q, Tian Q, et al. Learning mobile manipulation through deep reinforcement learning[J]. Sensors, 2020, 20(3): 1-18.

[48] Tremblay J, To T, Sundaralingam B, et al. Deep object pose estimation for semantic robotic grasping of household objects[EB/OL]. (2018-09-27)[2020-09-01]. https://arxiv.org/pdf/1809.10790.pdf.

[49] Du D, Zhou Q, Qi N, et al. Learning to control a free-floating space robot using deep reinforcement learning[C]. The IEEE International Conference on Unmanned Systems. Beijing: IEEE, 2019: 519-523.

[50] 施群, 吕雷, 谢家骏. 可变环境下仿人机器人智能姿态控制[J]. 机械工程学报, 2020, 56(3): 64-72.
(Shi Q, Lü L, Xie J J. Intelligent posture control of humanoid robot in variable environment[J]. Journal of Mechanical Engineering, 2020, 56(3): 64-72.)

[51] Xi A, Chen C. Walking control of a biped robot on static and rotating platforms based on hybrid reinforcement learning[J]. IEEE Access. 2020, 8: 148411-148424.

[52] Xi A, Mudiyanselage T W, Tao D, et al. Balance control of a biped robot on a rotating platform based on efficient reinforcement learning[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(4): 938-951.

[53] Tan J, Zhang T N, Coumans E, et al. Sim-to-real: Learning agile locomotion for quadruped robots[EB/OL]. (2018-05-16)[2020-09-01]. https://arxiv.org/pdf/1804.10332.pdf.

[54] Sartoretti G, Paivine W, Shi Y, et al. Distributed learning of decentralized control policies for articulated mobile robots[J]. IEEE Transactions of Robotics, 2019, 35(5): 1109-1122.

[55] Zhang M, Geng X, Bruce J, et al. Deep reinforcement

learning for tensegrity robot locomotion[C]. The IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017: 634-641.

[56] Montgomery W, Levine S. Guided policy search as approximate mirror descent[EB/OL]. (2016-07-15)[2020-09-01]. https://parxiv.org/df/1607.04614.pdf.

[57] Liu Q, Liu Z, Xu W, et al. Human-robot collaboration in disassembly for sustainable manufacturing[J]. International Journal of Proguction Research, 2019, 57(12): 4027-4044.

[58] Tao L, Bowman M, Zhang J, et al. Learn task first or learn human partner first? deep reinforcement learning of human-robot cooperation in asymmetric hierarchical dynamic task[EB/OL]. (2020-03-01)[2020-09-01]. https://arxiv.org/pdf/2003.00400.pdf.

[59] Jonas T, Ali S, Faisal A A. Human-robot collaboration via deep reinforcement learning of real-world interactions[EB/OL]. (2019-12-02)[2020-09-01]. https://arxiv.org/pdf/1912.01715.pdf.

[60] Clegg A, Erickson Z, Grady P, et al. Learning to collaborate from simulation for robot-assisted dressing[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2746-2753.

[61] Ghadirzadeh A, Chen X, Yin W J, et al. Human-centered collaborative robots with deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 566-571.

[62] 金哲豪, 刘安东, 俞立. 基于GPR和深度强化学习的分层人机协作控制[J]. 自动化学报, DOI: 10.16383/j.aas.c190451.
(Jin Z H, Liu A D, Yu L. Hierarchical human-robot cooperative control based on GPR and DRL[J]. Acta Automatica Sinica, DOI: 10.16383/j.aas.c190451.)

[63] Huang B, Ye M, Hu Y, et al. A multirobot cooperation framework for sewing personalized stent grafts[J]. IEEE Transactions on Industrial Informatics, 2017, 14(4): 1776-1785.

[64] Bae H, Kim G, Kim J, et al. Multi-robot path planning method using reinforcement learning[J]. Applied Sciengces-basei, 2019, 9(15): 3057.

[65] Semnani S, 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.

[66] Semnani S H, Ruiter D, Liu H. Force-based algorithm for motion planning of large agent teams[EB/OL]. (2019-09-10)[2020-09-01]. https://arxiv.org/ftp/arxiv/papers/1909/1909.05415.pdf.

[67] Freed B, Sartoretti G, Choset H. Simultaneous policy and discrete communication learning for multi-agent cooperation[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2498-2505.

[68] Zhang H, Li D, He Y. Multi-robot cooperation strategy in game environment using deep reinforcement learning[C]. The IEEE International Conference on Robotics and Biomimetics. Kuala Lumpur: IEEE, 2018: 886-891.

[69] Ma J, Lu H, Xiao J, et al. Multi-robot target encirclement control with collision avoidance via deep reinforcement learning[J]. Journal of Intellicent & Robotics Systems, 2020, 99(2): 371-386.

[70] Wang J, Cao J, Stojmenovic M, et al. Pattern-RL: Multi-robot cooperative pattern formation via deep reinforcement learning[C]. THe 18th IEEE International Conference on Machine Learning and Applications. Boca Raton: IEEE, 2019: 210-215.

[71] Lathuilière S, Massé B, Mesejo P, et al. Deep reinforcement learning for audio-visual gaze control[C]. The IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 1555-1562.

[72] Church A, Lloyd J, Hadsell R, et al. Deep reinforcement learning for tactile robotics: Learning to type on a braille keyboard[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6145-6152.

[73] Guo W, Wang J, Wang S. Deep multimodal representation learning: A survey[J]. IEEE Access, 2019, 7: 63373-63394.

作者简介

董豪 (1996−), 男, 硕士生, 从事智能控制与智能制造的研究, E-mail: gs.hdong19@gzu.edu.cn;

杨静 (1991−), 男, 副教授, 博士, 从事视觉计算与触觉感知等研究, E-mail: jyang23@gzu.edu.cn;

李少波 (1975−), 男, 教授, 博士生导师, 从事智能制造、制造物联等研究, E-mail: lishaobo@gzu.edu.cn;

王军 (1995−), 男, 硕士生, 从事机器视觉与智能制造的研究, E-mail: gs.wangjun19@gzu.edu.cn;

段仲静 (1992−), 男, 硕士生, 从事产品质量在线管控与深度学习的研究, E-mail: duan_zhongjing@163.com.

(责任编辑：孙艺红)