文章编号:1672-6553-2023-21(12)-053-017

DOI:10.6052/1672-6553-2023-134

基于强化学习的机器人多接触交互任务控制*

柯恺宸[1,2]　金士博[1]　高博扬[4]　黄行蓉[1,3†]

(1.北京航空航天大学 中法工程师学院/国际通用工程学院,北京　100191)

(2. 北京航空航天大学 杭州创新研究院(余杭),杭州　310023)

(3. 北京航空航天大学 发动机研究院,北京　100191)

(4. 哈尔滨工业大学 计算学部,哈尔滨　150001)

摘要　作为自动化和智能化时代的代表,机器人技术的发展成为智能控制领域研究的焦点,各种基于机器人的智能控制技术应运而生,机器人被越来越多地应用于实现与环境之间的复杂多接触交互任务.本文以机器人复杂多接触交互任务为核心问题展开讨论,结合基于强化学习的机器人智能体训练相关研究,对基于强化学习方法实现机器人多接触交互任务展开综述.概述了强化学习在机器人多接触任务研究中的代表性研究,当前研究中存在的问题以及改进多接触交互任务实验效果的优化方法,结合当前研究成果和各优化方法特点对未来机器人多接触交互任务的智能控制方法进行了展望.

关键词　强化学习,　智能控制,　机器人,　多接触交互任务

中图分类号:TP242　　**文献标志码**:A

A Survey of Robot Intelligent Control Method in Contact-Rich Tasks*

Ke Kaichen[1,2]　Jin Shibo[1]　Gao Boyang[4]　Huang Xingrong[1,3†]

(1.Sino-French Engineering School, Beihang University, Beijing　100190, China)

(2. Hangzhou Innovation Institute(Yuhang), Beihang University, Hangzhou　310023, China)

(3. Research Institute of Aero Engine, Beihang University, Beijing　100190, China)

(4. Faculty of Computing, Harbin Institute of Technology, Harbin　150001, China)

Abstract　As a representative of the automation and intelligence era, robot technology has become the focus of the intelligent control research, Various robot intelligent control technologies have been developed, Robots are more and more used to achieve some complex contact-rich interaction tasks. This paper chooses robot complex contact-rich interaction tasks as a topic, combining the research of robot reinforcement learning, to make a survey of research about the robot contact-rich interaction tasks based on reinforcement learning. We review some representative researches of implementing reinforcement learning in robot contact-rich interaction tasks, analyze the existing problems in these researches, summarize the optimization methods of solving these problems to improve the experimental effects, and finally make a prospect on the future of robot contact-rich tasks.

Key words　reinforcement learning,　intelligent control,　robot,　contact-rich tasks

2022-06-19 收到第 1 稿,2022-12-02 收到修改稿.

*国家自然科学基金资助项目(52105083,52175071,U2341231)和两机基础科学中心国际合作项目(P2022-C-III-001-001), National Natural Science Foundation of China (52105083, 52175071, U2341231) and the Science Center for Gas Turbine Project(P2022-C-III-001-001).

† 通信作者 E-mail:huangxingrong @buaa.edu.cn

引言

近年来,机器人技术得到了蓬勃发展,许多先进的机器人控制技术相继出现:如带有阻抗力控制技术的机器人[1,2],能够实现人机协作的带有精确控制技术的协作式机器人[3,4]等.基于这些技术,如今的机器人拥有更加灵巧的运动能力,相比于原先固定重复执行机械式操作任务的机器人,其能够完成更多更加复杂的、原先只能由人类自身完成的任务,可以有效的实现与环境间的接触交互.

现有研究将机器人或其末端执行器与交互物体、环境之间持续频繁接触的任务定义为机器人多接触交互任务[5].这类任务要求机器人不仅能够精确控制轨迹,还可以有效控制接触交互动作,自主调节控制与环境间的相互作用力,完成与环境间的安全交互,如同人类与物体间接触时表现出的对接触力的反馈控制能力[6].多接触交互任务涉及类型从服务类的机器人开门[7]、擦拭桌面[8],到传统工业中的销孔装配[9]及医疗领域的手术、康复等[10].此外,在接触交互的基础上还存在动态交互任务,这类任务基于交互作用力控制改变交互物体动力学特性,主要包括机器人对目标物体的投掷、打击、动态抓取等动作[11,12],任务需要对交互物体动力学特性的研究,效果更具随机性,针对动态交互任务的概述详见参考文献[13].本文针对机器人一般多接触交互任务为核心内容展开综述.

机器人多接触交互任务的难点是接触时的稳定安全控制.经典控制方法往往通过力或力/位置混合控制[14]结合比例积分(PI)控制器使得机器人与环境的交互力处于预期设置状态;以阻抗、导纳控制[15]为主的间接力控制思想使机器人在交互中表现出机械顺从性;基于力跟踪的自适应控制使机器人能与未知环境交互并适应交互环境发生的变化[16].然而,经典控制方法控制机器人完成任务的过程中,机器人的总体运动规划仍需依靠人类操作实现,机器人难以更自主地针对目标任务进行决策,存在智能化层面的不足,且其在未知的复杂交互任务的背景下能表现出的效果有限.与之相较,应用机器学习方法的人工智能技术在此问题中赋予机器人决策实现任务的能力[17],成为实现机器人动态机械系统智能化控制的一种选择.在机器人接触交互任务中,目前常见的智能学习技术是模仿学习与强化学习:模仿学习基于人类或专家的示例来学习任务[18,19],强化学习基于经验误差负反馈迭代的机器学习方法,智能体通过与环境的不断交互的经验中更新学习策略,以自主规划出能完成预期任务目标的策略[20,21].相较而言,模仿学习的优势是其学习过程的稳定性和收敛性,其更容易收敛到良好的策略,且由于它是基于人类或专家示例的数据,能避免强化学习中的探索问题,通常不需要过多的样本数据.但模仿学习由于需要示例数据,且仅依赖于已有示例数据,通常缺乏探索,无法自主探索新的策略或解决新问题,学习的效果好坏取决于示例数据的好坏.强化学习的优势则是其自主学习和探索的能力,并能在各种复杂多样环境中使用,劣势在于其通常需要大量的试验和样本数据来学习有效的策略,且训练出的策略依赖于奖励函数和超参数设置,具有不稳定性.但通过强化学习,理论上能够通过自主学习过程规划出比人类专家演示更加优秀的策略[22],是一个极具前景的研究方向.本文将强化学习作为机器人智能控制的主要讨论内容.

近年来,随着基于神经网络的深度强化学习算法取得的技术性突破,如今的强化学习算法具备处理大规模动作空间问题的能力[23],推动了强化学习的广泛应用.基于深度强化学习的研究展示出这种人工智能算法的智能决策优势,研究者们通过强化学习训练 AlphaGo 与人类进行围棋博弈并击败人类顶尖棋手[24]、训练智能体以高分自动通关 atari 主机游戏[25]、训练智能体在星际争霸游戏中击败 99.8%的人类玩家[26]、训练电脑能自主根据用户需求生成推荐系统[27]、训练汽车实现自动驾驶[28]、训练机器人自主避障[29]、抓取物品[30]等.越来越多的研究者尝试将强化学习和人工智能技术应用于机器人研究中[31],包括用强化学习解决机器人复杂多接触交互任务的问题,例如在接触稳定的条件下通过强化学习训练机器人开门[32]、训练机器人完成销孔装配[33]、训练机器人实现柔和地表面擦拭[34]等.

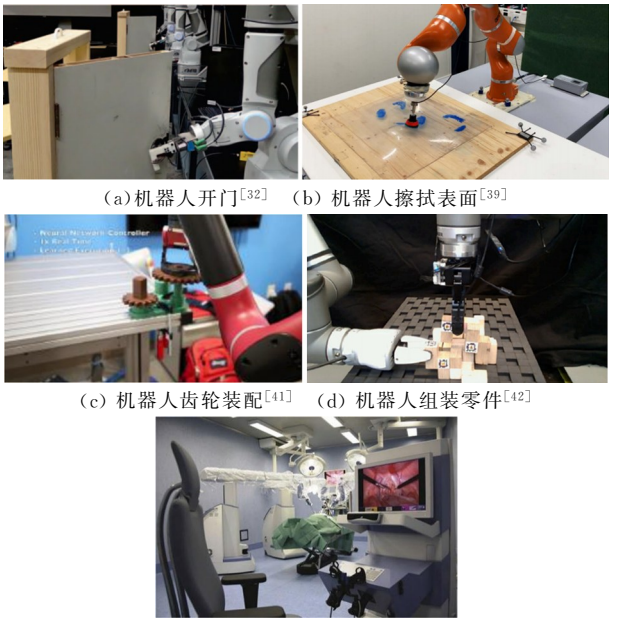
本文围绕如何借助强化学习控制机器人实现多接触交互任务,介绍了当前的主流研究进展,展望了未来的可能发展方向.以下第一节介绍机器人多接触交互任务的类型和特点;第二节介绍强化学习的基本概念以及强化学习算法类型;第三节介绍

强化学习在机器人多接触交互任务上的研究进展;第四节指出当前使用强化学习训练机器人做多接触交互任务存在的问题,讨论强化学习方法当前存在的局限性,提出可通过将力控、学习类方法、机器视觉等方法相结合,以提升机器人多接触交互智能控制效果.

1　机器人多接触交互任务概述

多接触交互任务需要机器人或其末端执行器保持与交互物体或环境频繁长时间接触,机器人动力学控制效果将影响机器人与环境间交互的安全性,影响机器人对环境变化的自适应能力,决定目标任务的可行性.

多接触交互任务的种类十分广泛,从操作性质可将机器人多接触交互任务分为以下三类[5]:(1)铰接式物体交互任务:包括开门[35]、开抽屉[36]、开瓶子[37]、转动气阀[38]等.机器人所交互的物体为铰接式物体,这类物体只能沿着预定义的路径移动,在交互过程中机器人某些方向上的自由度将由于铰接式物体运动学结构受到限制.机器人在规划这类任务时需要对所交互物体的动力学结构进行精确建模,否则不准确的运动学模型将带来僵硬的交互接触,使机构运动产生较大的内力,导致交互操作困难以及物体损坏.(2)表面接触操作任务:如表面擦拭[8](擦桌子、擦黑板等)、表面处理[39,40](抛光、打磨、雕刻、写字等).这类任务要求机器人具备精确的力控制和表面适应能力,以顺应交互环境表面的情况,避免交互过程中发生碰撞和产生不必要的过度力,以保证机器人平稳均匀地在物体表面执行运动轨迹.(3)工件对准操作任务,如销孔装配[9]、零件装配[41],组装[42]等.这类任务相对而言较为复杂,要求装配时机器人所持零件与目标之间的相互对准,在特定工业零件装配中还对任务完成的精度和准确性存在要求.操作过程需要引导机器人根据两个物件之间的结构的配对情况和装配时的受力调整末端所持零件的位姿.本文将对这几类交互任务的强化学习控制进行总结,此外,还存在一些医疗康复任务[43],如通过机器人辅助进行手术操作、外骨骼机器人辅助人类康复训练等.这类任务中机器人将与人类进行人机接触交互,需要极其精准的机器人运动控制和安全的接触交互,也将在后续有所讨论.图 1 展示了 4 种具有代表性的机器人多接触交互任务实验.



(a)机器人开门[32]　(b) 机器人擦拭表面[39]

(c) 机器人齿轮装配[41]　(d) 机器人组装零件[42]

(d) 手术机器人[43]

图 1　机器人多接触交互任务示例

Fig.1　Examples of robot contact-rich manipulation (a) Robotic door opening[32] (b)Robotic surface wiping[39] (c) Robotic assembly of gear[41] (d)Robotic assembly of part[42] (e)Surgical robot[43]

2　强化学习算法概述

2.1　强化学习概述

强化学习是一种类似于人类经验学习的人工智能方法,近年来在机器人领域得到了广泛应用.强化学习的本质是在交互中学习,并进行交互过程的最优化[23].

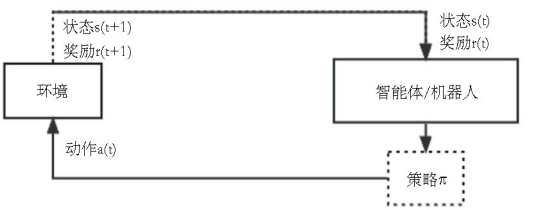


图 2　强化学习观察－动作－学习的循环过程

Fig.2　The perception-action-learning loop of reinforcement learning

强化学习过程可以被描述为一个马尔科夫决策(MDP)过程[20],每一次学习迭代过程在整个过程中将被视作由一组参数组成的一个元组 (s,a,r,T,γ):s 代表智能体和交互环境的状态;a 代表智能体的动作;r 代表智能体获得的奖励;T 代表传递函数,描述在状态 s 下执行动作 a 后的发生的状态变化;γ 代表未来获得奖励的折扣因子,范围在 0 到 1 之间.在整个的学习过程中,智能体通过观

察当前状态 st 进行相应的动作 $a(t)$，算法将根据智能体在环境中的每一步行为向智能体提供奖励 $r(t)$.智能体再根据收到的奖励 $r(t)$ 进行下一步的交互行为动作 a_{t+1}(图 2).

智能体通过学习获得的策略由 π 表示，代表从状态到动作或从状态到动作的概率分布的映射，这取决于算法类型的不同.每一次智能体动作的执行获得的奖励 r 将会累积到总奖励值 R 中：

$$R=r_0+\gamma r_1+\gamma^2 r_2+\gamma^3 r_3+\cdots\cdots \tag{1}$$

强化学习最终的目标是通过不断进行迭代寻优找到一种最优策略 π^*，其能使智能体根据不同的环境状态自适应输出一组动作，以获得全过程最大累计奖励期望 E：

$$\pi^*=\operatorname{argmax} E[R|\pi] \tag{2}$$

2.2 强化学习算法类型

强化学习是一个经验学习的过程，智能体通过反复循环探索过程获得对环境的认识.根据是否在训练前基于预先设置的经验模型对所交互环境建模，强化学习分为有模型(Model-based)和无模型(Model-free)两类算法：(1)Model-based 算法在训练前将针对目标任务和交互环境构造预测模型，智能体在训练时通过读取该经验预测模型，预测出后续训练过程中每一步交互的状态和所能获得的奖励.在 Model-based 算法中，智能体通过参考读取预测模型，只需要采样较少的数据即可实现训练过程收敛，样本效率高，训练速度快.但其算法复杂性更高，且需要基于所交互背景提前建模，导致方法的泛化性有限，且训练好坏取决于预测模型建模的准确性.(2)Model-free 算法则不需要对环境进行建模，智能体从交互中直接进行学习和迭代.而 Model-free 算法不需要对环境的建模预测，相对更简单，应用更加方便，但算法中智能体只能基于随机采样进行学习，需要大量的样本数据，样本效率较低，学习时间长.

除此之外，根据马尔科夫决策过程求解问题的方式，强化学习算法还分为以下三类[23]：(1)基于值函数(Value-based)，(2)基于策略搜索(Policy-based)，(3)将二者结合，既使用价值函数，又采用策略搜索机制的演员-评论家算法(Actor-Critic).下面详细介绍这三类算法原理.

(1) Value-based 类算法

Value-based 价值函数法基于对处于给定状态的价值的估计.在强化学习中，将会构造一个与状态和动作相关联的质量函数 Q，记作 $Q\pi(s|a)$.Q-learning 是 Value-based 算法的典型代表，其基本原理在于建立一个 Q 表格，在表格中为任务每步动作设定 Q 值.其大小将基于动态规划不断进行更新，最终收敛到各个策略、状态、动作所能获得的奖励值.在学习策略过程中，算法便会根据当前状态下 Q 值大小选择 Q 值最大的动作策略：

$$\pi(s|a)=\operatorname{argmax} Q^{\pi}[s|a] \tag{3}$$

因此，Q-Learning 的算法核心在于 Q 函数的设置，目前常见的 Value-based 类算法有两种：(1)深度 Q-Learning 算法(DQN)[44]，其通过设置神经网络来近似拟合 Q 函数，以解决复杂高维任务中 Q 函数的复杂度问题.但 DQN 通常适用于离散动作空间，对于机器人操作任务这种连续空间上的动作而言，文献[45]中还提出归一化优势函数法(NAF)，算法从原理上可视作 DQN 的连续空间扩展；(2)Soft Q-Learning 算法[46]，其通过基于能量模型的方式，在训练最大化奖励的同时加上最大化熵的目标，让学习训练过程能够探索除最大 Q 值动作外其他方向上的动作，以避免局部最优点的出现，增加学习策略的鲁棒性.

(2) Policy-based 类算法

与 Value-based 类相反，Policy-based 策略搜索方法不考虑值函数模型，而是直接搜索最优策略.通常在此方法中会将策略 π 参数化为参数 θ，训练过程是一个随机优化过程，智能体在每个状态下的下一步动作按照概率分布执行.算法根据每步动作获得的奖励期望，基于策略梯度上升法来优化概率模型，进而优化策略参数 θ，以最大化预期回报 R[47].

在每一步步长 α 下，策略参数的更新过程可表示为：

$$\theta_{t+1}=\theta_t+\alpha\sum\nolimits_a \Delta\pi_\theta(a\mid s_t)Q^{\pi}(s_t,a) \tag{4}$$

常见的用于机器人多接触交互任务的 Policy-based 类算法有以下三类：(1)路径策略改进算法(简称为 PI^2)[48]，该方法在早期的机械臂领域应用较为广泛，通过路径积分对机器人的运动轨迹进行搜索，其特点是适用于机器人运动，能规划出平滑的轨迹，显著缺点是完成任务的学习能力有限；(2)

Model-free 类型的信赖域策略优化法(TRPO)[49]和近端策略优化法(PPO)[50],这两种算法的基础是策略梯度上升法,解决了策略梯度法训练过程中对学习率敏感导致的更新幅度大的问题,以保证训练策略的合理性,学习性能好且学习速度更快,目前在无模型强化学习领域引起了广泛关注;(3) Model-based 算法中的基于模型预测的引导性策略搜索法(GPS)[51,52],通过预先设定引导学习策略向高回报区域搜索,训练出的策略能够避免陷入局部最优,较为稳定,且训练效率更高.

(3) Actor-Critic 算法

演员－评论家算法(Actor-Critic)将 Value-based 算法以及 Policy-based 算法相结合[53],算法中设置演员(Actor)和评论家(Critic)两个神经网络:Actor 网络用于预测行为的概率,学习一个能获得当前状态下高回报的策略,相当于基于策略梯度法的 Policy-based 类算法;Critic 网络则是用于估计预测在这个状态下的价值,相当于 Value-based 类算法的动态规划.

比较基于策略梯度法的 Policy-based 类算法以及 Value-based 类算法,基于策略梯度法的 Policy-based 类算法的优势是当策略具有大量参数时,由于具有更高的样本效率,能保证策略快速收敛.且其通过训练直接输出动作策略,更适合解决机器人多接触交互任务这类连续空间中的动作问题,但由于梯度法的局限性容易陷入局部最优解,方差较大,训练出的策略不够稳定可靠;Value-based 类能够输出稳定的策略,但在连续空间中的表现不够理想,且训练效率较低.Actor-Critic 算法通过 Actor 网络输出策略,由 Critic 网络根据交互中获得的奖励更新自身网络,并对 Actor 网络输出的动作策略进行评估(图 3),以结合两类算法的优势.

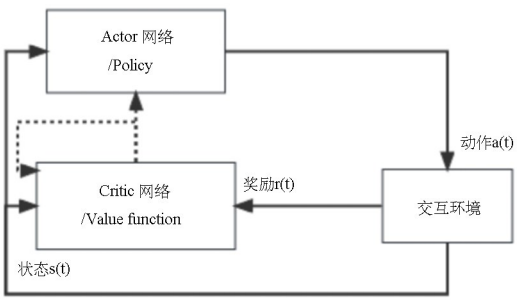


图 3　Actor-Critic 网络框图,虚线代表 Critic 网络根据从环境中获得的奖励值更新自身以及评估 Actor 网络策略
Fig.3　Actor－Critic network diagram

由于 Actor-Critic 算法可以结合两类算法的优势,近年来基于 Actor-Critic 框架的算法不断推陈出新,逐渐成为强化学习算法研究的热点趋势.目前常见的基于 Actor-Critic 框架的算法包括:① 深度确定性策略梯度算法(DDPG)[54],即在 DQN 算法的基础上结合 Actor-Critic 框架,能够在连续动作空间中更有效地学习;② Soft Actor-Critic 算法(SAC)[55,56],即结合 Soft-Q-Learning 算法中基于能量模型学习的思想,具备更强的探索能力和更加具有鲁棒性的学习策略,且通过结合 Actor 网络提高了训练的效率;③ 双延迟深度确定性策略算法(TD3)[57],即在 DDPG 算法的基础上进行改进,通过添加新的神经网络,以及 Actor 网络延迟更新的思想,避免 DDPG 算法中带来的误差估计问题;④ 优势演员－评论家算法(A2C)及异步优势演员－评论家算法(A3C)[58],即通过多个异步更新的 Actor-Critic 网络,实现并行的强化学习交互训练,大幅度提高训练的收敛速度和训练效率.

3　强化学习训练机器人多接触交互任务的研究进展

3.1　基于强化学习的机器人多接触交互任务研究

强化学习方法在近年来取得了很大程度的发展,通过将这些方法应用于训练机器人实现多接触交互任务,机器人在不同背景的交互任务中通过自主学习制定出合适的策略,已能初步表现出智能化控制效果.本节将介绍基于强化学习训练机器人多接触任务的研究进展.

文献[59]是最早进行强化学习训练机器人多接触交互任务实验的研究.早期由于技术的局限性,对机器人训练模拟器环境的开发并不完整,机器人只能在真实环境中直接进行训练,探索过程需保证运动的安全性,无法进行过多的随机采样,要求高训练样本效率.文中提出应用 PI^2 算法训练机器人进行开门任务的运动/力控制,训练机器人进行柔顺的开门.该方法能够实现开门动作,但由于训练条件和算法的局限性,在机器人受到训练后执行动作时的作用力曲线中,预期作用力效果与实际作用力相比存在较大误差.

针对真实环境下机器人训练样本效率问题,文献[60]提出在机器人零件装配任务中应用强化学习在 Model-based 类下的引导性策略搜索法

(GPS),通过 Model-based 算法的预测模型以减少训练收敛所需的采样数据,其学习过程表现出更高效率,学习出的策略更具鲁棒性:通过 20～25 个样本数量和 3～4 分钟的机器人交互能完成学习到成功的控制策略,能以 100%的任务成功率完成装配任务,训练出的控制策略在交互环境发生存在扰动变化的情况下能允许 1～2cm 的目标扰动偏差.但由于算法本身依赖于根据任务设定预测模型,对预测模型的准确设置要求也将导致该方法不易实现,且针对不同位置的交互任务都需要不同的预测模型,算法泛化性较差.文献[61]中还通过异步训练的方式实现机器人开门动作(图 4),即同时训练多个机器人并汇总各个体的探索数据,在一次训练中增加获得的样本数量,减少训练的次数和时间.在 GPS 算法的基础上进一步提高训练效率,缩短了 1 倍以上的达到最大奖励所用时间,但仍旧依赖于设置完整的预测模型实现固定任务.文献[62]尝试将强化学习与手术机器人的路径规划相结合,以创建出一种能在动态情况下避免碰撞的机器人运动控制方法.

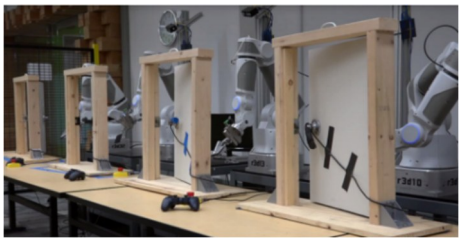


图 4 机器人通过异步训练学习开门动作[61]

Fig.4 Robot learning a door opening task through asynchronous training[61]

之后,随着强化学习算法的推进和机器人强化学习训练技术的发展,如 PPO、DDPG、SAC、TD3 等 Model-free 类算法提供了高性能的学习训练框架,简化了机器人多接触交互任务的训练过程,优化了实现效果;用于智能体模型连续运动训练的模拟器得到开发,为机器人提供了安全的训练探索方式:通过 Mujoco 物理引擎能让现实环境的物理交互过程在模拟器中实现[63],以 OpenAI gym 为代表的模拟器能建立机器人与交互环境的仿真模型,让机器人进行大量的强化学习训练,获得大量的训练样本数据以获得最优策略[64].文献[32]中即借助基于 Mujoco 环境下的 gym 模拟器进行训练(图 5),运用文中所提的归一化优势函数法(NAF)和 DDPG 算法训练机器人做开门动作,经过模拟器训练迭代最终获得百分百的开门成功率.文献[65]采用 PPO 算法在 ROS Gazebo 模拟环境中控制机器人学习完成轴孔装配任务;文献[66]同时对比了 PPO、DDPG、SAC 算法在同一机器人轴孔装配任务中的训练效果,提出使用 SAC 算法学习到的机器人装配策略具有更强的鲁棒性和泛化性;此外,如 NVIDIA 公司研发的 Isaac gym 平台还通过运用 GPU 加速进行并行多机器人训练(图 6),以提高学习训练效率[67].文献[69,70]还开发了能高效收集学习训练数据的手术机器人强化学习模拟平台,在这些模拟平台中能构建数十种手术环境和物理交互动作,进一步推进了将强化学习应用到手术机器人中.

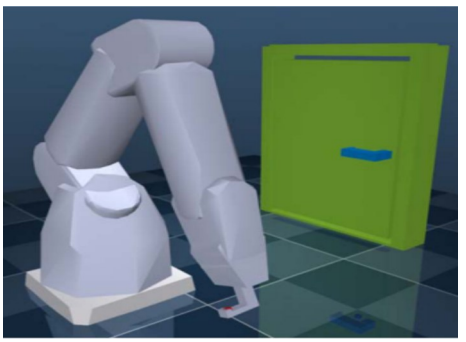


图 5 基于 gym 模拟器仿真训练机器人开门[32]

Fig.5 Robot door opening training based on Gym simulator[32]

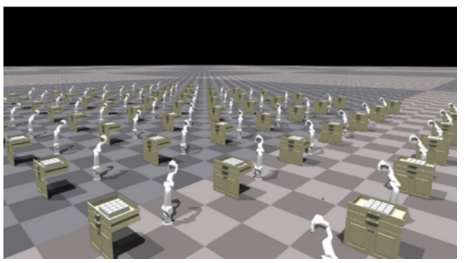


图 6 通过 Isaac gym 模拟器进行异步多机器人训练

Fig.6 Asynchronous multi-robot training in Isaac Gym simulator

3.2 基于多方法融合强化学习的机器人多接触交互研究

考虑到受力效果、交互环境变化、训练机器人实现多任务场景交互等问题,研究人员尝试通过将强化学习与其他技术方法相融合来提高强化学习在机器人多接触交互任务的智能控制效果.

首先介绍用于强化学习训练接触交互时优化机器人力学特性的相关研究成果.

文献[71]提出结合 GPS 算法和力传感器进行销孔装配任务,通过将力传感器数据集成到强化学习框架中,根据力传感器的反馈数据调整学习探索

过程中的运动轨迹,以自适应调整交互环境的误差变化(图 7).其在销孔装配任务中表现出相比于一般基于线性高斯控制器的 GPS 算法而言在孔位置存在误差偏移的情况下提升了成功率,在偏移量小于 5mm 时仍能使训练最终达到 100%成功,偏移量为 10mm 时保持 60%的成功率.

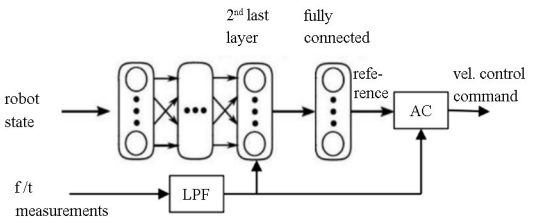


图 7　力传感器数据结合学习算法中的神经网络结构训练控制机器人运动[71]

Fig.7　Neural net architecture with force/torque measurement for robot motion training[71]

文献[41]中提出结合 GPS 算法和力控制器控制机器人做高精度齿轮装配任务,力控制器在机器人动力学模型基础上添加对机器人位置/速度的增益控制,学习出的策略将考虑并顺从交互过程中的外力变化.在高精度齿轮装配中相比于无法实现任务的基础 GPS 算法,其能表现出 60%～100%的成功率及对装配物体的存在的偏移量在 5cm 内都具有泛化性.在两篇文章中,待改进的问题在于可以将原始视觉和触觉输入添加到方法框架中,以实现装配交互的高度识别并改善学习到的策略.

文献[33]中运用 SAC 算法结合力控制方法解决机器人销孔任务,同时采用 PD 控制和阻抗控制(文献[15]、4.2.2)两种力控制框架与强化学习训练过程相结合,测试训练效果,通过训练效果对比确定两种力控制框架中的最优增益参数.所提出的方法能让机器人获取随任务阶段变化行为的策略,在交互柔顺性得到保证的条件下成功执行高精度销孔插入任务.该方法的不足在于机器人学习后的性能高度依赖于控制器超参数的选择,可以通过加入视觉识别及人类演示来获取最优控制器超参数.

文献[72]中提出结合 DDPG 算法和可变阻抗控制方法进行表面接触任务,可变阻抗控制是一种在阻抗控制基础上结合强化学习的力控方法(4.2.2、文献[73,74]),通过将阻抗增益加入强化学习动作空间中,让机器人能根据任务阶段不同表现出不同的顺从特性.该方法让机器人在未知环境下直接进行表面接触交互,相较于直接强化学习训练机器人运动和基于一般的固定阻抗控制方法表现得更具可靠性,最终学习出策略的奖励分数分布平均值更高,在相同训练时间内获得了 1 倍以上的学习奖励值,在交互环境位置、环境刚度、交互产生的摩擦力不同的交互任务中也能获得平均值更高的奖励.文献[75]还将这种可变阻抗控制方法应用在外骨骼机器人上,以减小外骨骼机器人实际与理想运动状态之间的跟踪误差和控制交互力大小,提高康复过程交互的安全性.

文献[42]中运用自适应 DDPG 算法结合力触觉传感器解决机器人组装零件问题,文献[76]中运用 TD3 算法和触觉识别控制机器人实现组装结构相对复杂的家具.通过机器人末端抓手上的触觉传感器获取交互任务信息(图 8),再将获得的交互信息传入强化学习框架中.该方法在所组装家具位置固定的情况下能够以高成功率满足预期效果.但训练过程是通过模拟器传递到现实控制中,其中存在不准确的接触建模误差;文献[77]中同样提出结合 TD3 算法及触觉传感器实现机器人开门,控制过程利用触觉传感器识别门把手位置,改善了学习出的策略中机器人开门动作和姿势.开门过程中能够激活所有触觉传感器上的单元,增加交互接触面积,以保证更好的抓握稳定性和更一致的开门结果.

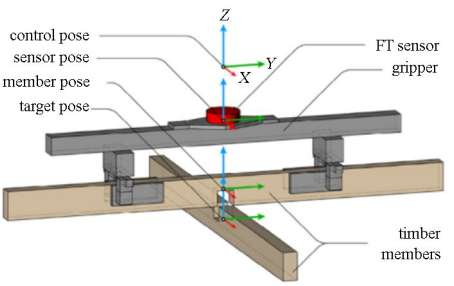


图 8　装备触觉传感器识别系统的机器人抓手案例[42]

Fig.8　Prototype of robotic gripper equipped with tactile sensor[42]

基于机器人与未知物体间的多接触交互训练问题和学习策略的泛化性的优化的相关研究成果也越来越受到重视.

文献[78]中提出结合 PI^2 算法与力控制器的思想,在机器人利用力控制器通过估计这类铰接系统的运动学约束情况,实现机器人开门、开抽屉等运动.机器人与几种固定状态下未知的铰接系统进行交互,通过学习估计出几种不同打开方式的铰接系统的约束模型.其在仿真和真实实验中都做到了

100%成功实现机器人开关门、抽屉,交互过程表现出一定的泛化性.方法的缺点是这将使缺少对机器人作用力的优化、对机器人柔顺性的考虑、以及对铰接系统位置发生变化情况下的讨论.

文献[79]中提出结合 SAC 算法和查询机制,让人类用户加入强化学习框架,对执行的交互任务类型和状态进行指示,强化学习训练根据指示初始化框架设置,算法在机器人开门任务中表现出 100%的成功率及能根据物体特征进行策略变化的泛化性.方法的局限性和待改进的方向在于难以确定复杂交互任务的查询标准,可以通过如神经网络这类拟合估计模型提供的数据处理功能来辅助量化.

文献[80]中提出使用示范扩大策略梯度算法(DAPG)[81]利用多指灵巧机械手实现机器人开门.DAPG 算法是一种 Model-free 类,基于策略梯度法的强化学习算法,其思想是在强化学习框架中结合人类示范来加速强化学习算法,多用于机器人灵巧手的交互训练.通过训练,其表现出对门初始位置的变化具备高鲁棒性,学习到的策略都能根据人类示范进行模仿训练.在泛化性提高的同时,其还能从非零奖励开始策略搜索,以更快的速度获得收敛的最优策略,相比于普通强化学习算法,减小了 2~3 倍以上的训练时间.

文献[52]中提出结合 GPS 算法、PI^2 算法和视觉识别技术进行机器人开门.通过视觉识别交互物体的状态,根据视觉输入读取预测模型,全局训练基于 GPS 算法,局部运动轨迹优化基于 PI^2 算法.相比于基于线性高斯控制器 GPS 算法,机器人学习到 100%成功开门的策略的迭代次数从 8 次下降到 3 次.同时,方法表现出的学习策略泛化能力大大提高,在所交互的门的位置改变的情况下通过视觉识别和强化学习训练仍能在 5 次迭代后达到 80%的成功率.但方法局限性在于任务的实现范围有限,需要通过预先人工演示来获得预测模型.

文献[82]中提出结合强化学习和视觉识别技术进行机器人开门、开抽屉任务,通过视觉识别使机器人表现出自主识别能力,能识别所交互铰接任务物体特征(图 9),通过对特征的定量估计,以识别交互的位置并进行奖励函数的设置,能同时实现多种铰接式任务,具备好的泛化性;文献[83]则先搭建了“眼在手配置”的机械臂视觉伺服系统,融合比例控制与滑模控制设计出一种基于图像的机械臂视觉伺服控制器.再通过深度强化学习 DDPG 算法自适应调整控制器伺服增益,减少伺服误差,在机械臂轴孔装配实验中表现出强鲁棒性和快速收敛的效果;文献[84,85]同样讨论强化学习结合视觉图像问题,机器人通过视觉识别神经网络,能够获取擦拭表面环境情况及所需擦拭的任务情况.再根据机器人视觉获取的要清洁的污垢污渍的位置/形状,自主学习清洁操作;文献[86]还讨论了强化学习训练轨迹的优化问题,基于 SAC 算法和视觉识别,并将强化学习策略与轨迹优化框架相结合,以执行和优化机器人的擦拭运动轨迹.

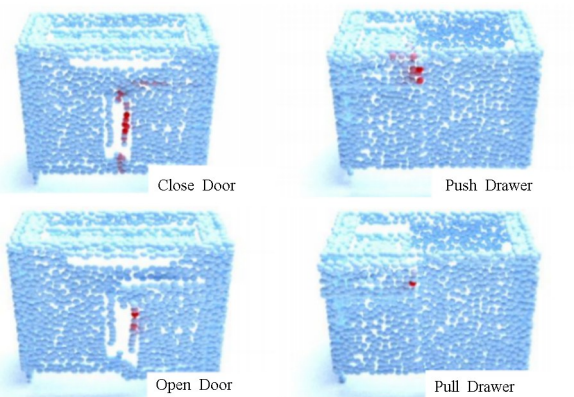


图 9 基于视觉识别点云模型以识别所交互物体特征[82]

Fig.9 Visual recognition point cloud model for recognizing features of objects[82]

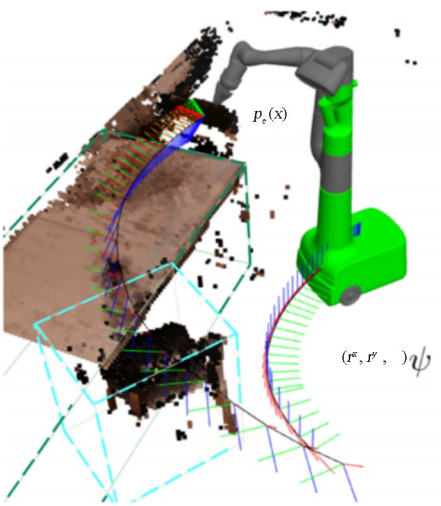


图 10 识别擦拭任务中所需擦拭污渍的特征以学习目标运动轨迹[86]

Fig.10 Robot wiping trajectory learning by sensing stains[86]

文献[34]中提出将强化学习 PPO 算法、可变阻抗控制及视觉识别技术三种方法相结合进行机器人表面擦拭任务.在视觉识别任务的基础上,通过可变阻抗控制,使机器人交互过程中获得对外界受力的顺从特性,能自主控制末端接触力大小.训练出的策略表现既表现出擦拭环境变化的泛化性,

还表现出一定的交互安全性,保证在成功完成擦拭任务的同时,机器人的最大作用力约为 5N,远小于安全接触阈值,且在受到突发外界力时仍保持末端位姿稳定.

4　强化学习训练机器人多接触交互任务中存在的问题和可能的优化途径

综上所述,强化学习算法和强化学习与机器人多接触交互任务结合的研究在当前取得了一定的成功,机器人能够通过强化学习训练中获得自主决策完成多接触交互任务的能力.但由于强化学习技术目前的局限性,以及现实中机器人多接触交互任务本身的复杂性等,基于强化学习的智能控制方式存在众多问题和不足.

4.1　强化学习训练机器人做多接触交互任务中存在的问题和挑战

基于强化学习实现机器人多接触交互任务控制中遇到的问题和挑战主要包括[87,88]:学习策略的稳定性和可靠性、学习的效率、仿真环境到实际机器人控制的误差、学习策略的泛化性、复杂机器人交互任务的奖励函数设置、真实环境下机器人交互的安全性等.这些问题会带来许多困难,主要包括以下五类:(1)强化学习中的奖励函数在复杂多接触交互任务中难以得到精确设置,使得训练过程只能使用稀疏奖励来引导.随着任务范围或行动维度的增加,算法找到高奖励策略的难度将成倍增加,使得强化学习的有效性和使用范围大打折扣.(2)复杂多接触交互任务的强化学习探索过程常会由于任务本身的复杂性导致其在初始低回报的动作中耗费大量时间,致使学习训练效率下降,训练时间过长.(3)多接触交互任务需要机器人表现出交互过程中可靠的安全稳定性,对学习训练要求高,难度大.(4)强化学习的探索采样过程需要由模拟器完成,再输出到真实操作环境中.但多接触交互任务的物理交互过程在模拟器环境下存在建模误差,从模拟器训练出的策略在真实环境下可能存在偏差.(5)泛化性差是强化学习算法本身存在的通病,只通过强化学习训练难以获得自适应性强的学习策略.

4.2　优化机器人多接触交互任务智能控制方法可能途径

对于上述问题的解决和优化,过去的研究除对强化学习算法本身的改进外,从以下四个方向出发(图 11):(1)机器学习类算法的融合,将强化学习与模仿示例学习等学习类算法相结合,基于模仿示例学习提高强化学习的训练样本效率,实现算法特性上的互补;(2)结合强化学习框架和自适应控制、阻抗顺从控制等经典力控制方法,以经典控制方法表现出的顺从特性,提高机器人操作的安全性和环境适应性,改进学习效果;(3)在强化学习过程中结合触觉、视觉等多传感器,根据多传感器的识别数据控制强化学习过程,以学习出具有泛化和自适应能力的策略;(4)进一步考虑从学习模拟器到真实环境的误差问题,以保证输出的运动控制策略的可靠性.

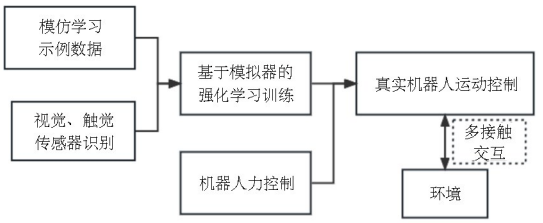


图 11　优化机器人多接触交互任务效果的智能控制方法框图
Fig.11　Block diagram of intelligent control method for robot contact-rich tasks optimization

4.2.1　学习类算法融合

进行学习类算法融合是强化学习算法改进的一种思路,可用于训练机器人的学习类方法分为三类[17]:基于强化学习、基于模仿学习、基于迁移学习.在机器人多接触交互任务中,迁移学习方法仍待研究,而以示例模仿学习(LfD)[18,19]为代表的模仿学习与强化学习在过去得到了广泛应用.

以示例学习为代表的模仿学习算法基本思想为:人通过握住机器人的末端执行器来引导机器人、给予机器人预期的参考轨迹数据,机器人根据参考轨迹能自主学习其中的过程.其通常基于动态运动基元算法来拟合示教的轨迹,使示教的轨迹参数化.过去通过示例学习机器人实现了自动擦拭桌面[8]、销孔装配任务[89]、机器人插插头[90]、机器人开门[91]等任务.

比较两种学习类方法的特点,示例模仿学习的框架从人类专家示例演示出发采集信息,根据示例表现出相对应的动作,相对于强化学习的自主学习过程更加简单、稳定;而强化学习在自主探索过程容易碰壁,且依赖于准确的奖励函数定义,但却能通过自主学习过程规划出比人类专家演示更加优

秀的策略.故过去的研究将两种学习类算法相结合,充分发挥二者的优势.

实现学习类算法融合存在两种途径:

一是用示例模仿学习的人类专家演示采集过程来简化强化学习的自主探索过程.文献[92]中提出将预先训练好的机器人数据集加入强化学习框架,以有效快速地学习新任务.在此基础之上,文献[93]提出使用示例与强化学习相结合,一起用于解决探索困难的复杂任务,从而能够在训练早期实现非零回报和合理的策略,缩短强化学习的随机探索阶段,提高训练的效率.

二是将示例模仿学习中的人类专家演示应用于奖励函数设置定义问题.根据从人类专家演示中获得预期的机器人规划效果、量化机器人探索过程的策略好坏,辅助完成奖励函数的设置.此思路的代表为逆强化学习算法(IRL)[94,95].逆强化学习的思想为:根据专家演示直接获取奖励函数,将直接获取的奖励函数用于强化学习的自主探索,以训练机器人找到完成预期目标的最优策略(图 12).

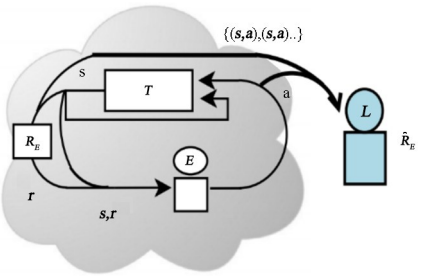


图 12 逆强化学习示意图[95]

Fig.12 Diagram of inverse reinforcement learning[95]

文献[96]基于逆强化学习算法实现了机器人销孔装配任务的交互实验,通过专家示例演示生成了完成任务所需的动作分布及奖励函数,通过该奖励函数运用 TRPO 算法进行训练,其相比普通的模仿学习表现出的任务平均完成度更高,证明了这种方法的可行性.

4.2.2 强化学习结合力控制方法

经典力控制方法是学习类方法外研究机器人多接触交互任务的另一方向,也是机器人实现接触交互任务的基础.受益于先前的研究中的机器人力控制理论,机器人做复杂接触交互任务的安全性和可行性得到了提高.对于强化学习而言,初始的强化学习算法并无交互力的反馈,训练出的策略缺少相对应的力控制约束,这容易导致安全问题,故将强化学习框架中结合经典力控制方法必不可少[97].

根据对机器人交互作用力的控制方式,经典机器人力控制方法分为显式或隐式控制:显式力控制直接控制机器人末端的作用力大小方向,但目前从技术上难以实现复杂交互任务的要求.隐式力控制设置机器人交互作用力与其动力学参数间的动态关系,间接实现交互力控制,能使机器人具有对外力的顺从机械特性,即机器人能在接触交互的过程中根据外力表现出顺应外力大小方向变化的动力学性能,成为接触交互安全的关键[5].

隐式控制力实现机器人顺从特性的代表是阻抗控制.阻抗控制的核心思想是为机器人的交互建立一种虚拟质量-弹簧-阻尼动力学模型[15]:

$$\boldsymbol{M}(\ddot{x}_r-\ddot{x}_d)+\boldsymbol{B}(\dot{x}_r-\dot{x}_d)+\boldsymbol{K}(x_r-x_d)=\boldsymbol{F}_e \tag{5}$$

式中,$\boldsymbol{M}$、$\boldsymbol{B}$、$\boldsymbol{K}$ 为交互方向上的惯性、阻尼、刚度矩阵,$\boldsymbol{F}_e$ 为交互方向上的作用力矩阵,x_r、x_d 分别为交互方向上的实际位置和预期位置.

通过将此虚拟模型设置到机器人各种交互点(如机器人末端、机器人各关节),并设定相关的弹簧刚度、系统阻尼系数.机器人在空间中仍会按照设定轨迹运动,但其会根据外力大小,与设定的增益成比例地偏离轨迹,基于模型参数表现出顺从的机械状态,以实现机器人的柔性控制.阻抗控制在开门任务[98,99],表面任务[100]和装配任务[33]中展现了安全的机器人接触交互.

此外,对于强化学习和阻抗力控制方法的结合,一种新颖的思路是文献[73,74]提出的基于强化学习的变阻抗控制.在阻抗控制中机器人表现出的柔顺性与阻抗增益大小相关,而根据所做交互任务、或任务阶段的不同,交互的环境刚度存在区别,机器人表现出的柔顺特性应相对应变化.在线实时调整阻抗增益的形式能保证交互过程中机器人动力学参数和末端作用力的稳定[101].而变阻抗控制通过强化学习,将阻抗增益加入训练策略动作空间,让阻抗增益根据学习奖励随任务过程而改变,以适应任务各阶段的要求.文献[102]在机器人表面擦拭、开门等几种多接触交互任务中,对比了是否采用阻抗控制进行强化学习训练,以及使用固定阻抗控制结合强化学习和变阻抗控制训练出的效果,验证了使用强化学习训练机器人做变阻抗控制的可行性和优势.文献[41]中还证明结合变阻抗控

制能为强化学习策略训练带来一定泛化性.

4.2.3　强化学习结合多传感器

以力触觉感知、视觉识别为代表的传感器技术是机器人领域中的一项重要技术:利用力传感器的测量功能,机器人能对其末端受力情况进行精确估计,引导机器人调整位姿[103],利用视觉摄像头的图像特征识别或触觉传感器的接触识别,机器人得以定量化周围所交互环境的特征参数,识别交互情况[104].

过去对传感器多数据的应用推动了机器人多接触交互任务的研究:文献[105]利用计算机视觉识别几种类型的门的特征(门把手旋转轴、门把手位置),以让机器人能够针对各种类型的门选择合适的操作策略;文献[106]还将视觉和触觉相结合,以多传感模式识别所装配目标的几何信息、位置信息,控制机器人以高成功率实现销孔装配任务.

因此,针对未来对机器人多功能智能化的需求和强化学习在泛化性和奖励函数设置的局限性问题,过去的研究提出强化学习与传感器数据结合的想法并证明了可行性:如文献[77]中机器人结合触觉传感器实现开门,机器人通过触觉传感器获得对门的动力学模型的准确认识,开门操作的性能得以提高;文献[107]提供了一种能够从机器人的高维观测(如图像和触觉反馈)中通过算法提取稠密奖励函数的方法,实现强化学习更快的收敛速度和性能;文献[82]利用视觉识别结合强化学习训练,让机器人能同时识别门、抽屉等铰接式物体的特征,自主选择对应奖励函数和交互动作实现目标操作;文献中[108]还提出将强化学习和触觉、视觉传感器相融合,通过基于视觉与触觉模态的双信息通道引导强化学习.相比于不使用传感器、只使用一种传感器而言,在多接触交互任务中表现出的策略泛化性能够大幅度提高.

4.2.4　模拟器－真实环境问题研究

从模拟器到真实环境的过渡问题是强化学习训练机器人所必须讨论的一个关键点,是一直以来机器人领域强化学习训练待解决的一大挑战.强化学习训练机器人操作任务需要仿真模拟器的支持,初期的自主探索过程需要在模拟器中对预期任务的进行建模以供交互而非直接在真实环境下操作,以避免机器人在训练探索过程的危险.模拟器能够对机器人训练样本的获取提供保障.

然而,由于目前模拟器中对现实物理模型建模技术的局限性,以及从模拟器到现实的数据传输、噪声等原因,在模拟器中难以对指定任务中的物理交互进行精准建模,特别是复杂度高的多接触交互任务.这会使得强化学习模拟仿真训练效果与真实世界之间存在误差,训练出的机器人策略在现实中的可靠性存在问题[109].强化学习在机器人多接触交互上的研究需时刻考虑模拟器到真实环境的误差问题,以保证真实环境下机器人操作的安全.

目前的研究虽无法解决模拟器到真实环境的误差问题,但为此提供了一些思路:文献[110]中提出在强化学习中结合空间控制操作框架(OSC),提高模拟向真实迁移的可能性.文献[111]中讨论了几种模拟到真实环境的方法,如在模拟器的物理系统建立一个精确的数学模型、将模拟高度随机化,以覆盖真实世界数据的真实分布、训练过程中添加环境扰动等.

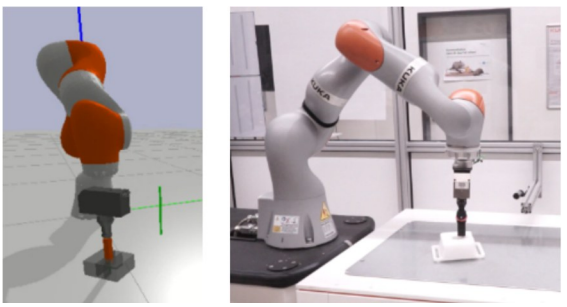


图 13　模拟器与真实环境下的机器人多接触交互任务[110]
Fig.13　Robot contact-rich tasks in real environment and simulator[110]

5　结论与讨论

本文通过调研基于强化学习的机器人多接触交互任务控制方法,介绍了当前强化学习算法在机器人多接触交互任务的研究进展,现有研究充分论证了强化学习应用于多接触交互任务智能控制方案的可行性,通过强化学习能够控制机器人自主探索智能化完成预期任务.

对于强化学习带来的交互安全性、泛化性、误差等开放性问题,以及用于训练机器人进行复杂交互任务的强化学习模拟器真实度问题,目前还有大量工作需要开展.对经典控制方法和模拟学习与强化学习的结合、多传感器与强化学习的结合以及强化学习模拟环境真实化的误差讨论,都将是未来改善强化学习效果的潜在方向.在未来多接触交互任

务的研究中,需进一步利用这些智能控制方法带来的优势和效果,并在强化训练框架中进行多方法的集成融合,以此提高机器人自主、自动、高效地实现更复杂多接触交互任务的效果.

参考文献

[1] ALBU-SCHÄFFER A, OTT C, HIRZINGER G. A unified passivity-based control framework for position, torque and impedance control of flexible joint robots [J]. International Journal of Robotics Research, 2007, 26(1): 23－39.

[2] ALBU-SCHAFFER A, OTT C, FRESE U, et al. Cartesian impedance control of redundant robots: recent results with the DLR-light-weight-arms [C]// 2003 IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2003: 3704－3709.

[3] CENCEN A, VERLINDEN J C, GERAEDTS J M P. Design methodology to improve human-robot coproduction in small-and medium-sized enterprises [J]. IEEE/ASME Transactions on Mechatronics, 2018, 23(3): 1092－1102.

[4] VYSOCKY A, NOVAK P. Human Robot Collaboration in industry [J]. MM Science Journal, 2016, 2016(2): 903－906.

[5] SUOMALAINEN M, KARAYIANNIDIS Y, KYRKI V. A survey of robot manipulation in contact [J]. Robotics and Autonomous Systems, 2022, 156: 104224.

[6] KLINGBEIL E, MENON S, GO K, et al. Using haptics to probe human contact control strategies for six degree-of-freedom tasks [C]// 2014 IEEE Haptics Symposium (HAPTICS). Piscataway, USA: IEEE, 2014: 93－95.

[7] NAGATANI K, YUTA S I. An experiment on opening-door-behavior by an autonomous mobile robot with a manipulator [C]// 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 1995, 2: 45－50.

[8] URBANEK H, ALBU-SCHAFFER A, VAN DER SMAGT P. Learning from demonstration: repetitive movements for autonomous service robotics [C]// 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2004: 3495－3500.

[9] NEWMAN W S, ZHAO Y, PAO Y H. Interpretation of force and moment signals for compliant peg-in-hole assembly [C]// 2001 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2001, 1: 571－576.

[10] JONES K C, DU W. Development of a massage robot for medical therapy [C]// 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics. Piscataway, USA: IEEE, 2003, 2: 1096－1101.

[11] BILLARD A, KRAGIC D. Trends and challenges in robot manipulation [J]. Science, 2019, 364(6446): eaat8414.

[12] RUGGIERO F, LIPPIELLO V, SICILIANO B. Nonprehensile dynamic manipulation: a survey [J]. IEEE Robotics and Automation Letters, 2018, 3(3): 1711－1718.

[13] WOODRUFF J Z, LYNCH K M. Planning and control for dynamic, nonprehensile, and hybrid manipulation tasks [C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2017.

[14] RAIBERT M H, CRAIG J J. Hybrid position/force control of manipulators [J]. Journal of Dynamic Systems, Measurement, and Control, 1981, 103(2): 126－133.

[15] HOGAN N. Impedance control: an approach to manipulation [C]// 1984 American Control Conference. [S.l.]: IEEE, 1984: 304－313.

[16] DUAN J J, GAN Y H, CHEN M, et al. Adaptive variable impedance control for dynamic contact force tracking in uncertain environment [J]. Robotics and Autonomous Systems, 2018, 102: 54－65.

[17] HUA J, ZENG L C, LI G F, et al. Learning for a robot: deep reinforcement learning, imitation learning, transfer learning [J]. Sensors, 2021, 21(4): 1278.

[18] 李帅龙, 张会文, 周维佳. 模仿学习方法综述及其在机器人领域的应用 [J]. 计算机工程与应用, 2019, 55(4): 17－30.
LI S L, ZHANG H W, ZHOU W J. Review of imitation learning methods and its application in robotics [J]. Computer Engineering and Applications, 2019, 55(4): 17－30.(in Chinese)

[19] ARGALL B D, CHERNOVA S, VELOSO M, et al. A survey of robot learning from demonstration [J]. Robotics and Autonomous Systems, 2009, 57

(5): 469－483.

[20] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. 2nd ed..

[21] PENG X B, ABBEEL P, LEVINE S, et al. DeepMimic: example-guided deep reinforcement learning of physics-based character skills [J]. ACM Transactions on Graphics, 37(4): 143.

[22] LIU R R, NAGEOTTE F, ZANNE P, et al. Deep reinforcement learning for the control of robotic manipulation: a focussed mini-review [J]. Robotics, 2021, 10(1): 22.

[23] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey [J]. IEEE Signal Processing Magazine, 2017, 34(6): 26－38.

[24] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529: 484－489.

[25] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518: 529－533.

[26] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning [J]. Nature, 2019, 575: 350－354.

[27] AFSAR M M, CRUMP T, FAR B. Reinforcement learning based recommender systems: a survey [J]. ACM Computing Surveys, 2022, 55(7): 1－38.

[28] LI D, ZHAO D B, ZHANG Q C, et al. Reinforcement learning and deep learning based lateral control for autonomous driving application notes [J]. IEEE Computational Intelligence Magazine, 2019, 14(2): 83－98.

[29] KUMAR V, HOELLER D, SUNDARALINGAM B, et al. Joint space control via deep reinforcement learning [C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2021: 3619－3626.

[30] LOBBEZOO A, QIAN Y J, KWON H J. Reinforcement learning for pick and place operations in robotics: a survey [J]. Robotics, 2021, 10(3): 105.

[31] SINGH B, KUMAR R, SINGH V P. Reinforcement learning in robotic applications: a comprehensive survey [J]. Artificial Intelligence Review, 2022, 55(2): 945－990.

[32] GU S X, HOLLY E, LILLICRAP T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates [C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2017: 3389－3396.

[33] BELTRAN-HERNANDEZ C C, PETIT D, RAMIREZ-ALPIZAR I G, et al. Learning force control for contact-rich manipulation tasks with rigid position-controlled robots [J]. IEEE Robotics and Automation Letters, 2020, 5(4): 5709－5716.

[34] ZHU X, KANG S, CHEN J. A Contact-Safe Reinforcement Learning Framework for Contact-Rich Robot Manipulation [C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2022.

[35] NIEMEYER G, SLOTINE J J E. A simple strategy for opening an unknown door [C]// IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 1997, 2: 1448－1453.

[36] LUTSCHER E, CHENG G. A set-point-generator for indirect-force-controlled manipulators operating unknown constrained mechanisms [C]// 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2012: 4072－4077.

[37] ZOLLNER R, ASFOUR T, DILLMANN R. Programming by demonstration: dual-arm manipulation tasks for humanoid robots [C]// 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2005: 479－484.

[38] CARRERA A, PALOMERAS N, HURTÓS N, et al. Learning multiple strategies to perform a valve turning with underwater currents using an I-AUV [C]// MTS/IEEE OCEANS 2015-Genova: Discovering Sustainable Ocean Energy for a New. New York, USA: IEEE, 2015.

[39] AMANHOUD W, KHORAMSHAHI M, BONNESOEUR M, et al. Force adaptation in contact tasks with dynamical systems [C]// 2020 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2020: 6841－6847.

[40] KOROPOULI V, LEE D, HIRCHE S. Learning interaction control policies by demonstration [C]// 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA:

IEEE, 2011: 344－349.

[41] LUO J L, SOLOWJOW E, WEN C T, et al. Reinforcement learning on variable impedance controller for high-precision robotic assembly [C]// 2019 International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2019: 3080－3087.

[42] APOLINARSKA A A, PACHER M, LI H, et al. Robotic assembly of timber joints using reinforcement learning [J]. Automation in Construction, 2021, 125: 103569.

[43] PETERS B S, ARMIJO P R, KRAUSE C, et al. Review of emerging surgical robotic technology [J]. Surgical Endoscopy, 2018, 32(4): 1636－1655.

[44] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning [C/OL]// NIPS Workshop on Deep Learning , 2013 (2013-12-19) [2022-6-19]. http://arxiv.org/abs/1312.5602.

[45] GU S X, LILLICRAP T, SUTSKEVER I, et al. Continuous deep Q-learning with model-based acceleration [C]// 33rd International Conference on Machine Learning. New York, USA: PMLR, 2016 (6): 2829－2838.

[46] HAARNOJA T, TANG H R, ABBEEL P, et al. Reinforcement learning with deep energy-based policies [C]// 34th International Conference on Machine Learning. New York, USA: PMLR, 2017 (3): 1352－1361.

[47] Kakade S M. A natural policy gradient [C]// 4th International Conference on Neural Information Processing Systems. Cambridge, UK: MIT Press, 2001.

[48] THEODOROU E, BUCHLI J, SCHAAL S. A generalized path integral control approach to reinforcement learning [J]. Journal of Machine Learning Research, 2010, 11: 3137－3181.

[49] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization [C]// 32nd International Conference on Machine Learning. New York, USA: PMLR, 2015, 37: 1889－1897.

[50] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB/OL]. (2017-7-20)[2022-6-19]. https://doi.org/10.48550/arXiv.1707.06347.

[51] SEJDINOVIC D, STRATHMANN H, GARCIA M, et al. Guided policy search [C]// 31st international conference on machine learning. New York, USA: PMLR, 2014

[52] CHEBOTAR Y, KALAKRISHNAN M, YAHYA A, et al. Path integral guided policy search [C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2017: 3381－3388.

[53] KONDA V, TSITSIKLIS J. Actor-critic algorithms. Advances in Neural Information Processing Systems 12, 1999. Cambridge, UK: MIT Press, 2000

[54] LILLICRAP T P, HUNT J J, Pritzel A, et al. Continuous control with deep reinforcement learning [EB/OL]. (2015-9-9) [2022-6-19]. https://doi.org/10.48550/arXiv.1509.02971.

[55] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor [C]// 35th International Conference on Machine Learning. 2018, 80: 1861－1870.

[56] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications [EB/OL]. (2018-12-13)[2022-6-19]. https://doi.org/10.48550/arXiv.1812.05905.

[57] FUJIMOTO S, HOOF H V, MEGER D. Addressing function approximation error in actor-critic methods [C]// 35th International Conference on Machine Learning. 2018, 80: 1587－1596.

[58] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C]// 33rd International Conference on Machine Learning. New York, USA: PMLR, 2016: 1928－1937.

[59] KALAKRISHNAN M, RIGHETTI L, PASTOR P, et al. Learning force control policies for compliant manipulation [C]// 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2011: 4639－4644.

[60] LEVINE S, WAGENER N, ABBEEL P. Learning contact-rich manipulation skills with guided policy search [C]// 2015 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: 2015.

[61] YAHYA A, LI A, KALAKRISHNAN M, et al. Collective robot reinforcement learning with distributed asynchronous guided policy search [C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 79－86.

[62] BAEK D, HWANG M, KIM H, et al. Path Planning for Automation of Surgery Robot based on Probabilistic Roadmap and Reinforcement Learning [C]// 2018 15th International Conference on Ubiquitous Robots (UR). Honolulu, HI, USA. IEEE, 2018: 342－347.

[63] TODOROV E, EREZ T, TASSA Y. Mujoco: A physics engine for model-based control [C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2022.

[64] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. Openai gym. [EB/OL]. (2016-6-5)[2022-6-19]. https://doi.org/10.48550/arXiv.1606.01540.

[65] 刘乃龙, 刘钊铭, 崔龙. 基于深度强化学习的仿真机器人轴孔装配研究 [J]. 计算机仿真, 2019, 36(12): 296－301.
LIU N L, LIU Z M, CUI L. Deep reinforcement learning based robotic assembly in simulation [J]. Computer Simulation, 2019, 36(12): 296－301.(in Chinese)

[66] 朱子璐,刘永奎,张霖,等. 基于深度强化学习的机器人轴孔装配策略仿真研究[J/OL].系统仿真学报[2022-6-19]. https://doi.org/10.16182/j.issn100-4731x.joss.23-0518.
ZHU Z L, LIU Y K, ZHANG L, et al. Simulation of robotic peg-in-hole assembly strategy with deep reinforcement learning [J/OL]. Journal of System Simulation [2022-6-19]. https://doi.org/10.16182/j.issn1004731x.joss.23-0518.

[67] MAKOVIYCHUK V, WAWRZYNIAK L, GUO Y, et al. Isaac gym: high performance gpu-based physics simulation for robot learning [EB/OL]. (2021-8-24)[2022-6-19]. https://doi.org/10.48550/arXiv.2108.10470.

[68] RICHTER F, OROSCO R, YIP M C. Open-sourced reinforcement learning environments for surgical robotics [EB/OL]. (2019-3-5)[2022-6-19]. https://doi.org/10.48550/arXiv.1903.02090.

[69] TAGLIABUE E, PORE A, DALL'ALBA D, et al. Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery [C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2020: 3261－3266.

[70] XU J Q, LI B, LU B, et al. SurRoL: an open-source reinforcement learning centered and dVRK compatible platform for surgical robot learning [C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2021: 1821 - 1828.

[71] LUO J L, SOLOWJOW E, WEN C T, et al. Deep reinforcement learning for robotic assembly of mixed deformable and rigid objects [C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2018: 2062－2069.

[72] BOGDANOVIC M, KHADIV M, RIGHETTI L. Learning variable impedance control for contact sensitive tasks [J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6129－6136.

[73] BUCHLI J, STULP F, THEODOROU E, et al. Learning variable impedance control [J]. International Journal of Robotics Research, 2011, 30(7): 820－833.

[74] ABU-DAKKA F J, SAVERIANO M. Variable impedance control and learning a review [J]. Frontiers in Robotics and AI, 2020, 7: 590681.

[75] ZHAO Z R, XIAO J C, WANG S P, et al. Model-based reinforcement learning variable impedance control for 1-DOF PAM elbow exokeleton [C]// 2021 China Automation Congress (CAC). [S.l.]: IEEE, 2021: 2369－2374.

[76] BELOUSOV B, WIBRANEK B, SCHNEIDER J, et al. Robotic architectural assembly with tactile skills: simulation and optimization [J]. Automation in Construction, 2022, 133: 104006.

[77] DING Z H, TSAI Y Y, LEE W W, et al. Sim-to-real transfer for robotic manipulation with tactile sensory [C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2021: 6778 - 6785.

[78] NEMEC B, ŽLAJPAH L, UDE A. Door opening by joining reinforcement learning and intelligent control [C]// 2017 18th International Conference on Advanced Robotics (ICAR). New York, USA: IEEE, 2017: 222－228.

[79] SINGH A, YANG L, HARTIKAINEN K, et al. End-to-end robotic reinforcement learning without reward engineering [EB/OL]. (2019-4-16)[2022-6-19]. https://doi.org/10.48550/arXiv.1904.07854.

[80] ZHU H, GUPTA A, RAJESWARAN A, et al. Dexterous manipulation with deep reinforcement learning: efficient, general, and low-cost [C]//

2019 International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2019: 3651－3657.

[81] RAJESWARAN A, KUMAR V, GUPTA A, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations [EB/OL]. (2019-4-16)[2022-6-19]. https://doi.org/10.48550/arXiv.1709.10087.

[82] GENG Y R, AN B S, GENG H R, et al. End-to-end affordance learning for robotic manipulation [C]// 2023 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2023: 5880－5886.

[83] 袁庆霓,齐建友,虞宏建.基于深度强化学习的机械臂视觉伺服智能控制[J/OL].计算机集成制造系统[2022-6-19]. http://kns.cnki.net/kcms/detail/11.5946.TP.20230607.1330.012.html.
YUAN Q N, QI J Y, YU H J. Visual servo intelligent control method for robot arms based on deep reinforcement learning [J/OL]. Computer Integrated Manufacturing Systems [2022-6-19]. http://kns.cnki.net/kcms/detail/11.5946.TP.20230607.1330.012.html.

[84] CAULI N, VICENTE P, KIM J, et al. Autonomous table-cleaning from kinesthetic demonstrations using Deep Learning [C]// 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). New York, USA: IEEE, 2018: 26－32.

[85] DEVIN C, ABBEEL P, DARRELL T, et al. Deep object-centric representations for generalizable robot learning [C]// 2018 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2018: 7111－7118.

[86] LEW T, SINGH S, PRATS M, et al. Robotic table wiping via reinforcement learning and whole-body trajectory optimization [C]// 2023 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: 2023: 7184－7190.

[87] IBARZ J, TAN J, FINN C, et al. How to train your robot with deep reinforcement learning: lessons we have learned [J]. International Journal of Robotics Research, 2021, 40(4-5): 698－721.

[88] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: a survey [J]. International Journal of Robotics Research, 2013, 32(11): 1238－1274.

[89] TANG T, LIN H C, ZHAO Y, et al. Teach industrial robots peg-hole-insertion by human demonstration [C]// 2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). Piscataway, USA: IEEE, 2016: 488－494.

[90] EHLERS D, SUOMALAINEN M, LUNDELL J, et al. Imitating human search strategies for assembly [C]// 2019 International Conference on Robotics and Automation (ICRA). New York, USA: 2019: 7821－7827.

[91] ARDUENGO M, COLOMÉA, LOBO-PRAT J, et al. Gaussian-process-based robot learning from demonstration [J/OL]. Journal of Ambient Intelligence and Humanized Computing, 2020 [2022-12-31].https://doi.org/10.1007/s12652－023－04551－7.

[92] KUMAR A, SINGH A, EBERT F, et al. Pre-training for robots: offline RL enables learning new tasks from a handful of trials [EB/OL]. (2022-10-11) [2022-12-1]. https://doi.org/10.48550/arXiv.2210.05178.

[93] NAIR A, MCGREW B, ANDRYCHOWICZ M, et al. Overcoming exploration in reinforcement learning with demonstrations [C]//2018 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2018: 6292－6299.

[94] NG A Y, RUSSELL S. Algorithms for inverse reinforcement learning [C]// 7th International Conference on Machine Learning. New York, USA: PMLR, 2000: 663－670.

[95] ARORA S, DOSHI P. A survey of inverse reinforcement learning: challenges, methods and progress [J]. Artificial Intelligence, 2021, 297: 103500.

[96] ZHANG X, SUN L T, KUANG Z A, et al. Learning variable impedance control via inverse reinforcement learning for force-related tasks [J]. IEEE Robotics and Automation Letters, 2021, 6(2): 2225－2232.

[97] KHADER S A, YIN H, FALCO P, et al. Stability-guaranteed reinforcement learning for contact-rich manipulation [J]. IEEE Robotics and Automation Letters, 2021, 6(1): 1－8.

[98] JAIN A, KEMP C C. Pulling open novel doors and drawers with equilibrium point control [C]// 2009 9th IEEE-RAS International Conference on Humanoid Robots. [S.l.]: IEEE Computer Society, 2009: 498－505.

[99] JAIN A, KEMP C C. Pulling open doors and drawers: Coordinating an omni-directional base and a compliant arm with Equilibrium Point control [C]// 2010 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2010: 1807－1814.

[100] LI Y N, GANESH G, JARRASSÉN, et al. Force, impedance, and trajectory learning for contact tooling and haptic identification [J]. IEEE Transactions on Robotics, 2018, 34(5): 1170－1182.

[101] WANG C H, KUANG Z A, ZHANG X, et al. Safe online gain optimization for variable impedance control [EB/OL]. (2022-10-11)[2022-12-1]. https://doi.org/10.48550/arXiv.2111.01258.

[102] MARTÍN-MARTÍN R, LEE M A, GARDNER R, et al. Variable impedance control in end-effector space: an action space for reinforcement learning in contact-rich tasks [C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2019: 1010－1017.

[103] 陈婵娟, 赵飞飞, 李承, 等. 多传感器协助机器人精确装配 [J]. 机械设计与制造, 2020(3): 281－284.
CHEN C J, ZHAO F F, LI C, et al. Multi-sensor assisted robotic accurate assembly [J]. Machinery Design & Manufacture, 2020(3): 281－284. (in Chinese)

[104] DRESP-LANGLEY B, NAGEOTTE F, ZANNE P, et al. Correlating grip force signals from multiple sensors highlights prehensile control strategies in a complex task-user system [J]. Bioengineering, 2020, 7(4): 143.

[105] IEEE/RSJ international conference on intelligent robots and systems [C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2022: 1.

[106] LEE M A, ZHU Y K, SRINIVASAN K, et al. Making sense of vision and touch: self-supervised learning of multimodal representations for contact-rich tasks [C]// 2019 International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2019: 8943－8950.

[107] WU Z, LIAN W Z, UNHELKAR V, et al. Learning dense rewards for contact-rich manipulation tasks [C]// 2021 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2021: 6214－6221.

[108] ICHIWARA H, ITO H, YAMAMOTO K, et al. Contact-Rich Manipulation of a Flexible Object based on Deep Predictive Learning using Vision and Tactility [C]// 2022 International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2022: 5375－5381.

[109] RUPAM MAHMOOD A, KORENKEVYCH D, KOMER B J, et al. Setting up a reinforcement learning task with a real-world robot [C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2018: 4635－4640.

[110] KASPAR M, MUÑOZ OSORIO J D, BOCK J. Sim2Real transfer for reinforcement learning without dynamics randomization [C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, USA: IEEE, 2020: 4383－4388.

[111] ZHAO W S, QUERALTA J P, WESTERLUND T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey [C]// 2020 IEEE Symposium Series on Computational Intelligence (SSCI). Canberra, ACT, Australia. New York, USA: IEEE, 2020: 737－744.