Journal of System Simulation

Volume 36 | Issue 6 Article 14

6-28-2024

Simulation of Robotic Peg-in-hole Assembly Strategy Based on DRL

Zilu Zhu
School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China, zilu_zhu@163.com

Yongkui Liu
School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China, yongkuiliu@163.com

Lin Zhang
School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

Lihui Wang
Department of Production Engineering KTH Royal Institute of Technology, Stockholm 25175, Sweden

See next page for additional authors

Simulation of Robotic Peg-in-hole Assembly Strategy Based on DRL

Abstract

Abstract: Aiming at the existing peg-in-hole assembly method problems of dependence on accurate contact state models, difficulties in data acquisition, low sampling efficiency, and poor security, a simulation research method for robot peg-in-hole assembly strategy based on DRL is proposed. A simulation environment of robot peg-in-hole assembly based on ROS-Gazebo is built, and a method of gravity compensation for force/torque sensor based on a least square method is proposed. The reinforcement learning paradigm is employed to model the robot peg-in-hole assembly, and a method based on soft actor-critic(SAC) algorithm is proposed. The communication mechanism between the simulation environment and the deep reinforcement learning algorithm is established through ROS. Simulation experiments show that the proposed SAC algorithm enables robots to accomplish the peg-inhole assembly task autonomously and compliantly with good generalization ability.

Keywords

peg-in-hole assembly, DRL, compliance control, assembly strategy simulation, ROS-Gazebo simulation environment

Authors

Zilu Zhu, Yongkui Liu, Lin Zhang, Lihui Wang, and Tingyu Lin

Recommended Citation

Zhu Zilu, Liu Yongkui, Zhang Lin, et al. Simulation of Robotic Peg-in-hole Assembly Strategy Based on DRL[J]. Journal of System Simulation, 2024, 36(6): 1414-1424.

第 36 卷第 6 期 系统仿真学报© Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

基于深度强化学习的机器人轴孔装配策略仿真研究

朱子璐[1]，刘永奎[1*]，张霖[2]，王力翬[3]，林廷宇[4]

(1. 西安电子科技大学 机电工程学院，陕西 西安 710071；2. 北京航空航天大学 自动化科学与电气工程学院，北京 100191；
3. 瑞典皇家理工学院 生产工程系，斯德哥尔摩 25175；4. 北京仿真中心 北京市复杂产品先进制造系统工程技术研究中心，北京 100854)

摘要：针对现有轴孔装配方法存在的依赖于精确的接触状态模型、数据采集困难、采样效率低、安全性差等问题，提出了一种基于DRL的机器人轴孔装配策略仿真研究方法。*搭建了基于ROS-Gazebo机器人轴孔装配仿真环境，提出了基于最小二乘法对力/力矩传感器进行重力补偿的方法；基于RL的范式对轴孔装配问题建模，并提出了一种基于SAC(soft actor-critic)算法的机器人轴孔装配方法；通过ROS建立了仿真环境与深度强化学习算法的通信机制。*实验结果表明：该算法能够使机器人自主且柔顺地完成轴孔装配任务，并具有较好的泛化性。

关键词：轴孔装配；DRL；柔顺控制；装配策略仿真；ROS-Gazebo仿真环境

中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2024)06-1414-11

DOI: 10.16182/j.issn1004731x.joss.23-0518

引用格式: 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]. 系统仿真学报, 2024, 36(6): 1414-1424.

Reference format: Zhu Zilu, Liu Yongkui, Zhang Lin, et al. Simulation of Robotic Peg-in-hole Assembly Strategy Based on DRL[J]. Journal of System Simulation, 2024, 36(6): 1414-1424.

Simulation of Robotic Peg-in-hole Assembly Strategy Based on DRL

Zhu Zilu[1], *Liu Yongkui*[1*], *Zhang Lin*[2], *Wang Lihui*[3], *Lin Tingyu*[4]

(1. School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China; 2. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China; 3. Department of Production Engineering KTH Royal Institute of Technology, Stockholm 25175, Sweden; 4. Beijing Complex Product Advanced Manufacturing Engineering Research Center, Beijing Simulation Center, Beijing 100854, China)

Abstract: Aiming at the existing peg-in-hole assembly method problems of dependence on accurate contact state models, difficulties in data acquisition, low sampling efficiency, and poor security, a simulation research method for robot peg-in-hole assembly strategy based on DRL is proposed. *A simulation environment of robot peg-in-hole assembly based on ROS-Gazebo is built, and a method of gravity compensation for force/torque sensor based on a least square method is proposed. The reinforcement learning paradigm is employed to model the robot peg-in-hole assembly, and a method based on soft actor-critic(SAC) algorithm is proposed. The communication mechanism between the simulation environment and the deep reinforcement learning algorithm is established through ROS.* Simulation experiments show that the proposed SAC algorithm enables robots to accomplish the peg-in-hole assembly task autonomously and compliantly with good generalization ability.

Keywords: peg-in-hole assembly; DRL; compliance control; assembly strategy simulation; ROS-Gazebo simulation environment

收稿日期：2023-05-05 修回日期：2023-06-23

基金项目：国家自然科学基金(61973243)

第一作者：朱子璐(2001-)，女，本科生，研究方向为人机协作装配、机器人装配技能学习。E-mail：zilu_zhu@163.com

通讯作者：刘永奎(1981-)，男，副教授，博士，研究方向为机器人学习、数字孪生、云边协同制造。E-mail：yongkuiliu@163.com

0 引言

装配是生产制造中的典型环节，也是最耗费人力的环节。相比人力装配，机器人装配具有更高的效率和准确性，可以将劳动力从重复且繁重的工作中解放出来。然而，目前应用于装配作业的机器人大多需要建立精确的基于位置控制的模型，依靠大量的参数部署工作和繁琐复杂的离线编程[1]，功能单一且只能完成特定的任务，无法适应复杂多变的工作环境。

轴孔装配是工业装配中一项非常典型且重要的操作，是大多数装配作业的主要工作模式。由于轴孔装配过程中机器人与环境存在大量复杂的接触，微小的位姿偏差都可能导致轴与孔之间产生巨大的接触力，严重时甚至会造成机器人的损坏，传统的机器人位置控制方法不再适用。

轴孔装配一直是工业机器人领域的研究热点，国内外学者对机器人轴孔装配方法的研究主要集中在2个方向：基于传统控制的方法和基于学习的方法[2]。

基于传统控制的轴孔装配方法涉及被动柔顺和主动柔顺2类。在被动柔顺控制中，文献[3]设计了一个具备远程柔顺中心的柔性手腕(remote compliance center, RCC)，在装配过程中当轴与孔之间存在位姿偏差时产生弹性形变，被动地调整装配中心位置。在主动柔顺控制中，传统的方法往往基于模型，主要针对轴孔装配中的动力学进行理论分析建模[4]，通过建立模型来设计控制策略，调整机器人的位姿，并根据实际装配过程中视觉[5]或力觉信息[6]的反馈进行装配。文献[7]在传统力反馈装配策略中加入了倾斜螺旋的搜孔方式，基于导纳控制策略实现了机器人在位置偏差的情况下对轴孔的主动柔顺装配。文献[8]分别从静态和动态环境下平面和空间角度对接触状态进行分析，以此来确定控制参数，实现轴孔装配中的最小接触力控制。文献[9]将轴孔装配中的卡阻状态细化为径向卡阻和双向卡阻，分别建立了两种卡阻状态下的几何模型和接触力模型，获得理论装配轨迹。

随着人工智能技术的迅速发展，机器学习算法被引入轴孔装配领域，成为一个热门的研究方向。目前，基于学习的轴孔装配方法研究可以分为模仿学习和强化学习两类。模仿学习是一种基于专家示教重建期望策略的方法[10]。文献[11]采用高斯混合模型对示教数据进行拟合，通过从人工示教中学习装配技能完成轴孔装配任务。文献[12]设计了一个基于力相关技能的统计模型，其中关键参数从示教数据中学习。文献[13]通过示教来牵引机器人完成轴孔装配任务，用传感器收集人类在装配过程中施加的力和校正速度，直接从示教数据学习导纳控制器的增益。RL是一种探索式的学习方法，智能体以试错的方式与环境不断交互，通过最大化累积奖励学习最优的策略。文献[14]使用DQN算法使机器人通过自学习完成装配。文献[15]基于Actor-Critic算法在仿真环境中通过预训练模型进行核心技能学习，在真实环境中再次训练以适应具体的装配环境，实现了机器人自主轴孔装配。文献[16]提出一种基于神经网络的轨迹规划器，通过RL生成和优化装配任务轨迹。

然而，当前机器人轴孔装配还存在几个方面的问题和挑战：① 传统的方法依赖于精确的接触状态模型，受限于固定的控制策略，无法适应复杂的装配环境。② 基于模仿学习的方法存在专家示教数据分布不均、大量数据采集困难等问题。③ 基于RL的方法在策略的训练效率、装配时间、装配成功率等方面还有待提升，在真实环境中训练时的安全难以保障。此外，工业场景下的装配环境日益趋于不确定和非结构化，要求装配策略具备更高的安全性、更强的鲁棒性和更好的适应性。④ 使用RL算法对机器人轴孔装配策略进行训练时，需要机器人与装配环境不断交互试错，而在物理环境中存在采样效率低、安全性差等问题[17]。

本文基于ROS下的物理仿真平台Gazebo搭建机器人轴孔装配仿真环境，在UR5机器人末端腕关节安装力/力矩传感器插件，并提出了基于最小

DOI: 10.16182/j.issn1004731x.joss.23-0518

第 36 卷第 6 期 系统仿真学报 Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

二乘法对其进行重力补偿，按照RL的范式对该轴孔装配问题进行建模，并提出了一种基于SAC (soft actor-critic)算法的机器人轴孔装配方法，通过ROS建立仿真环境与DRL的通信机制。

1 机器人轴孔装配策略仿真研究

1.1 机器人轴孔装配仿真系统设计

1.1.1 仿真系统架构

在物理装配环境中，场景覆盖度和样本数量有限，难以满足统计学习模型的训练需求。因此，需要建立仿真模型，模拟物理环境下的各种装配场景和极端工况，为统计学习模型生成大量的训练数据。统计学习模型基于这些数据进行训练与验证。将从仿真模型中采集的大量数据组织为各类信息，对相关的信息进行训练和拟合，进行深度的数据挖掘。基于仿真实验，学习最佳控制策略。在仿真模型中对策略进行测试，验证统计学习模型的可行性和有效性。本文建立的仿真模型为机器人轴孔装配仿真系统，统计学习模型为基于RL的轴孔装配问题模型。

机器人轴孔装配仿真系统在机器人操作系统ROS中部署，仿真系统的整体架构如图1所示。

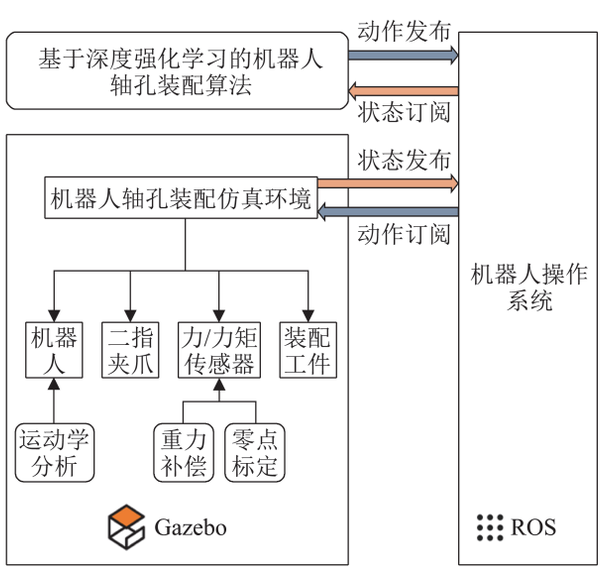


图1 仿真系统架构

Fig. 1 Simulation system architecture

基于ROS下的物理仿真平台Gazebo搭建机器人柔顺轴孔装配仿真环境。通过机器人运动学分析对仿真环境中的机器人进行运动控制，基于最小二乘法对力/力矩传感器进行重力补偿和零点标定，使仿真环境满足机器人柔顺轴孔装配策略的仿真需求。

通过ROS建立仿真环境与基于DRL的机器人柔顺轴孔装配算法模块之间的通信机制。算法模块订阅仿真环境发布的状态信息，根据状态和当前的策略选择合适的动作并发布；仿真环境订阅到动作信息，通过机器人运动学求解控制机器人执行对应的动作。

在机器人轴孔装配仿真系统中，DRL算法模块作为控制器输出控制信息，仿真环境中的机器人和夹爪作为执行机构，装配工件作为被控对象，力/力矩传感器作为检测装置感知环境接触力并反馈，构成一个完整的机器人轴孔装配仿真系统，支持对机器人轴孔装配策略进行研究、训练和验证。

1.1.2 仿真环境搭建

由于视觉传感器在装配中可能存在遮挡问题，难以准确预估轴在孔中的状态。因此，使用力/力矩传感器获取装配过程中的环境接触力，通过对装配过程中接触力的控制实现柔顺轴孔装配。

在机器人操作系统ROS下的物理仿真平台Gazebo搭建机器人轴孔装配仿真环境，包括UR5机器人、OnRobot RG2二指夹爪、轴孔装配工件、装配平台，在UR5机器人末端腕关节安装力/力矩传感器插件。搭建的仿真环境如图2所示。

Gazebo仿真平台提供强大的物理引擎，可以实现机器人的运动学、动力学仿真，同时还支持传感器数据及噪音的仿真。使用Gazebo搭建的机器人轴孔装配仿真环境可以满足机器人轴孔装配策略的仿真需求。此外，在仿真环境中基于最小二乘法对力/力矩传感器进行重力补偿，为装配过程提供准确的受力感知[18]，保证了仿真数据的真实性和有效性，实现了对物理环境高保真的模拟。

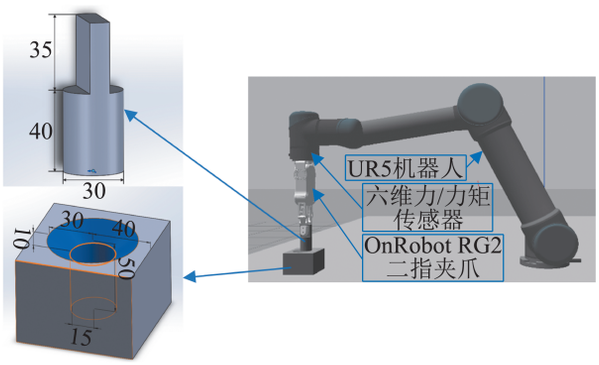


图2　轴孔装配仿真环境
Fig. 2　Simulation environment of peg-in-hole assembly

1.1.3 通信机制建立

机器人轴孔装配仿真系统基于ROS的分布式框架进行开发，各进程以节点的形式独立运行，使用了ROS标准的IO接口，实现机器人的感知、决策、控制等算法模块间点对点的松耦合连接，使仿真系统能更好地组织运行。Gazebo的接口为ROS提供了很好的支持，通过ROS建立了轴孔装配仿真环境和DRL算法的通信机制，提高了对仿真环境中数据的采样和处理效率。

仿真系统的通信机制如图3所示，首先通过ROS的Topic通信获取仿真环境中力和力矩信息，通过Service通信获取机器人位姿信息，作为状态输入DRL算法模块。DRL算法根据当前状态选择合适的动作，通过逆运动学求解出6个关节角度。通过ROS的Action通信控制仿真环境中的UR5机器人抓取装配工件运动，并获取下一个力/力矩传感器数据和机器人位姿信息，作为状态输入DRL模块。以此循环训练，直到算法收敛，探索到最优的装配策略。

1.2 机器人轴孔装配策略研究

DRL是一种探索式的学习方法，智能体借助深度学习的特征提取能力和强化学习的策略学习能力，通过环境给出的奖惩进行学习，依靠自身的经验不断探索环境，最终找到正确的策略以适应外部多变的环境。将强化学习应用于机器人轴孔装配中，无需分析接触状态，机器人通过不断探索装配环境直接学习装配策略。深度学习提高了装配策略的泛化性，使其能更好地适用于不确定、非结构化的装配环境[19]。

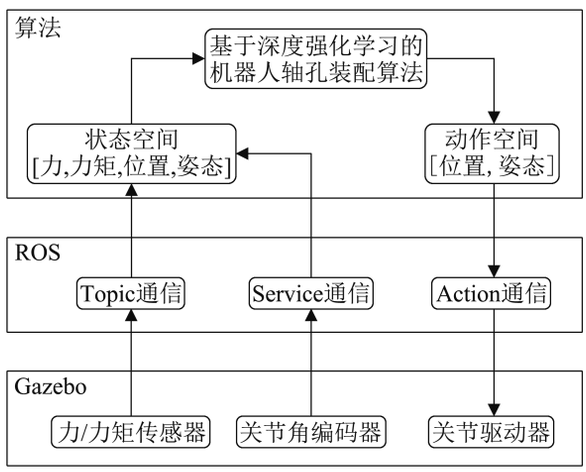


图3　仿真系统通信机制
Fig. 3　Communication mechanism of simulation system

1.2.1 机器人轴孔装配问题建模

马尔可夫决策过程是强化学习理论分析的基础，由四元组(S,A,P,R)表示马尔可夫决策过程模型。对机器人柔顺轴孔装配过程建立数学模型，状态转移概率分布为$P=p(s_{t+1}|s_t,a)$，表示机器人在当前状态s_t下执行动作a后，转移到下一个状态s_{t+1}的概率分布。机器人状态空间为S，动作空间为A，奖励函数为R。

(1) 状态空间

状态空间的设计需要准确且全面地描述智能体所处的环境、动作带来的变化，以及智能体与任务之间的关系。对于轴孔装配，需要考虑装配过程中的力、力矩、轴的位置、姿态，因此状态空间定义为

$$S=[F_x,F_y,F_z,T_x,T_y,T_z,P_x,P_y,P_z,O_x,O_y,O_z] \quad (1)$$

式中：F、T为力/力矩传感器测得的力、力矩，单位为N、N·m；P、O为轴的位置、姿态，单位为m、rad。

(2) 动作空间

动作空间用来描述智能体在训练环境中可执行的动作。在轴孔装配环境中，将动作空间定义为

$$A=[P_x,P_y,P_z,O_x,O_y,O_z] \quad (2)$$

第 36 卷第 6 期 系统仿真学报 Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

式中：P_x、P_y、P_z为机器人末端在x、y、z轴方向的平移，范围为[−0.001, 0.001]，单位为m；O_x、O_y、O_z为机器人末端绕x、y、z轴的旋转，范围为[−0.01, 0.01]，单位为rad。本文将轴向孔方向的探索范围扩大10倍，通过优化动作空间的权重加强对装配动作的引导。在本文的装配场景中，装配孔位于轴的正下方，因此，将z轴的平移范围定义为[−0.01, 0.001]，单位为m。

(3) 奖励函数

智能体根据奖励进行策略的优化，奖励函数决定了所设计算法的收敛程度和收敛速度。在轴孔装配问题中，不仅要考虑是否完成装配任务，还要考虑装配过程中接触力的大小，因此，将奖励函数定义为

$$R_{\mathrm{FT}}=\left\| \frac{F}{F_{\mathrm{threshold1}}},\frac{T}{T_{\mathrm{threshold1}}} \right\| \tag{3}$$

$$R_{\mathrm{PO}}=\left\| (P-P_{\mathrm{target}}),(O-O_{\mathrm{target}}) \right\| \tag{4}$$

$$R_{\mathrm{done}}=\begin{cases} D, & \text{未成功} \\ 2\times D, & \text{成功} \end{cases} \tag{5}$$

$$R=-\delta_1 R_{\mathrm{FT}}-\delta_2 R_{\mathrm{PO}}+\delta_3 R_{\mathrm{done}} \tag{6}$$

式中：R_{FT}为装配过程中的接触力、接触力矩的大小；$F_{\mathrm{threshold1}}$、$T_{\mathrm{threshold1}}$为力、力矩的阈值1，超出阈值时惩罚加大但不终止回合训练；R_{PO}为装配过程中轴的位姿与目标位姿的差距；R_{done}为触发终止条件时的奖励，与装配深度D成正比，终止时成功完成装配的奖励是未成功完成装配的2倍；R为奖励函数；δ_1、δ_2、δ_3为奖励函数各项的加权系数。

(4) 终止条件

考虑到轴孔装配过程中的安全问题，同时为了避免不必要的探索，提高训练效率，除了成功完成装配外，还设置了终止条件。

1) 接触力/力矩过大：设置了$F_{\mathrm{threshold2}}$、$T_{\mathrm{threshold2}}$，表示力、力矩的阈值2，阈值2是装配中的最大阈值，训练过程中超出最大阈值时，终止本回合训练；

2) 超出装配范围：在孔位上方定义一个长方体的区域作为装配范围，当训练过程中轴离开此区域时终止本回合训练；

3) 姿态误差过大：训练过程中轴的姿态误差过大时，说明探索方向错误，终止本回合训练。

1.2.2 基于SAC的机器人轴孔装配算法

在针对轴孔装配问题的研究中，常用的DRL算法有DQN算法[14]、PPO算法[1]、深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法[4]、孪生延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)算法[20]等。对上述算法进行对比分析后，选择了PPO和DDPG算法对轴孔装配过程进行训练，并对学习到的策略进行了测试。通过对测试结果的分析，结合随机性策略稳定性高和离线策略方法数据利用率高的优点，基于SAC算法对机器人柔顺轴孔装配策略进行研究。

SAC算法是基于最大熵思想的深度强化学习算法，在连续动作空间任务中有着出色的控制性能；使用随机性策略替换了DDPG中的确定性策略，并通过引入最大熵学习目标极大地改善了离线策略方法的探索性能；在训练稳定性、样本利用效率、策略探索能力等方面有突出表现，适用于机器人技能学习[21]。SAC算法在优化策略以获取更高累积奖励的同时，也会最大化策略的熵，其最优策略定义为

$$\pi^*=\max_{\pi} E\left[\sum_t \gamma^t (r(s_t,a_t)+\alpha H(\pi(\cdot|s_t)))\right] \tag{7}$$

式中：$H(\pi(\cdot|s_t))=E[-\log\pi(\cdot|s_t)]$，$\pi$为目前已找到的最大累计奖励的策略；$\alpha$为熵正则化系数；$H(\pi(\cdot|s_t))$为熵值。

引入熵的思想可以让策略尽可能随机，智能体更充分地探索状态空间，避免策略过早落入局部最优点，并且可以探索到多个可行方案来完成指定任务，提高策略的抗干扰能力。

SAC算法是一种off-policy算法，采用了经验回放机制消除数据的相关性，重复使用历史数据，提高了样本利用率；使用双Q网络结构，避免Q值过度拟合，保证了训练的稳定性。SAC算法框架如图4所示。

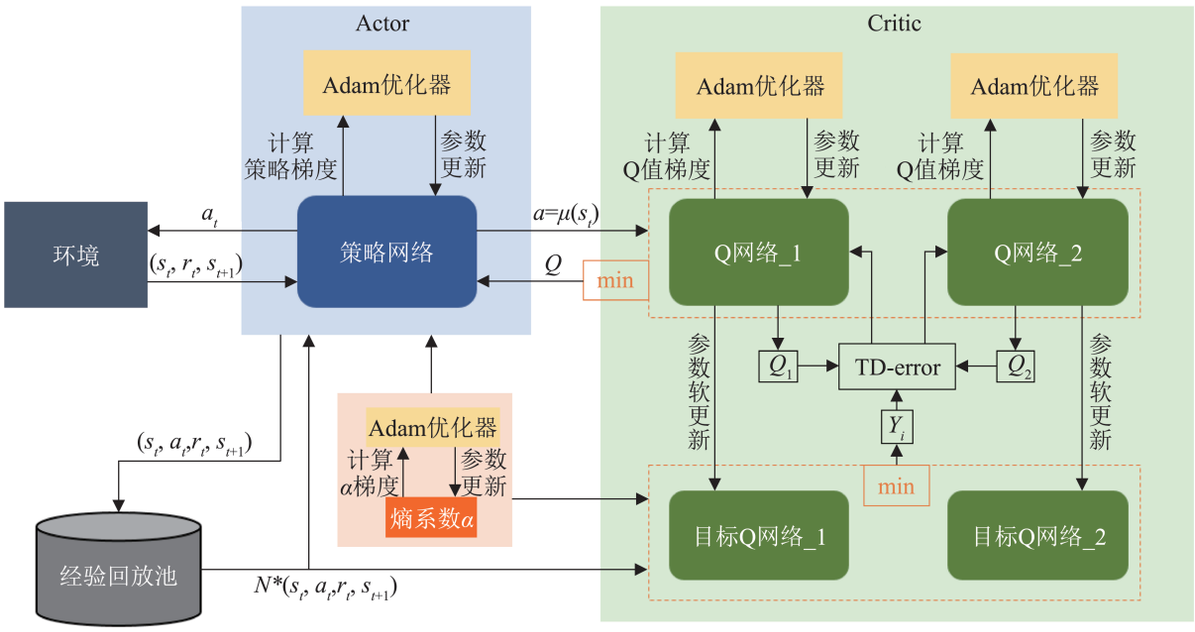


图4　SAC算法框架
Fig. 4　Framework of SAC algorithm

令$Q_\phi(s,a)$表示Q值函数，$\pi_\theta(a|s)$表示策略函数。Q值函数通过最小化柔性Bellman残差进行学习：

$$J_Q(\phi)=E[(Q(s_t,a_t)-r(s_t,a_t)-\gamma E_{s_{t+1}}[V_{\tilde{\phi}}(s_{t+1})])^2] \tag{8}$$

式中：$V_{\tilde{\phi}}(s)=E_{\pi_\theta}[Q_{\tilde{\phi}}(s,a)-\alpha\log\pi_\theta(a|s)]$，$Q_{\tilde{\phi}}$为目标Q值网络，其参数$\tilde{\phi}$由Q值网络的参数$\phi$通过Polyak平均的方法进行软更新。

$$\tilde{\phi}\leftarrow\tau\tilde{\phi}+(1-\tau)\phi \tag{9}$$

由于软更新参数$\tau\ll1$，目标网络的参数更新平缓，提高了学习过程的稳定性。

策略函数通过最小化式(10)的KL散度得到。

$$J_\pi(\theta)=E_{s\sim D}[E_{a\sim\pi_\theta}[\alpha\log\pi_\theta(a|s)-Q_\phi(s,a)]] \tag{10}$$

熵权衡系数α通过自动熵调节方法[22]进行更新：

$$J(\alpha)=E_{a\sim\pi_\theta}[-\alpha\log\pi_\theta(a|s)-\alpha\kappa] \tag{11}$$

式中：κ为目标熵的超参数。

使用基于SAC的机器人轴孔装配算法在搭建的仿真环境中对轴孔装配任务进行训练，具体实现过程：

(1) 机器人抓取装配工件后运动至初始位姿，使用随机网络参数分别初始化策略网络、2个Q值网络，并对应复制相同的参数初始化2个目标Q值网络，初始化经验回放池。

(2) 机器人在每个训练回合进行一次轴孔装配的尝试，每一步先获取当前的环境状态s_t，根据当前策略网络选择动作$a_t=\pi_\theta(s_t)$，机器人执行动作a_t，获得奖励r_t，环境状态由s_t变为s_{t+1}，将(s_t,a_t,r_t,s_{t+1})存入经验回放池中，训练回合结束后机器人返回初始位姿。

(3) 重复执行步骤(2)，当经验回放池中存放的数据量大于批量batch_size时，开始在每一步更新神经网络参数。

首先从经验回放池中采样N个元组(s_i,a_i,r_i,s_{i+1})，用目标Q值网络计算Y_i，通过最小化柔性Bellman残差更新Q值网络，基于最小化KL(Kullback-Leibler)散度更新策略网络，使用重参数化技巧保证梯度反向传播，可以直接使用梯度下降算法对策略网络的参数进行优化，最后通过Polyak平均对2个目标Q值网络进行参数软更新。

(4) 在每一步更新完神经网络参数后，通过自动熵调节方法更新熵正则化系数α，寻找策略在探索(最大化熵)与利用(最大化Q值)之间的平衡。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/14
DOI: 10.16182/j.issn1004731x.joss.23-0518

第 36 卷第 6 期 系统仿真学报 Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

(5) 重复训练回合，不断优化神经网络的参数，直到算法收敛，学习的策略达到预期的性能水平。

2 机器人轴孔装配策略仿真验证

为了验证本文算法的有效性和泛化性，在机器人轴孔装配仿真环境中对算法进行了训练，探索最优的轴孔装配策略。设置了多组实验对策略进行仿真验证，从仿真过程中接触力和力矩、装配成功率、装配时间等指标对策略进行评估。

2.1 策略仿真训练

本文算法基于Ubuntu 18.04操作系统的Python3和Anaconda3进行开发，使用了TensorFlow2.0深度学习框架和TensorLayer建立和训练神经网络，计算机CPU配置为Intel(R) Core(TM) i7-10700 CPU @ 2.90 GHz，GPU配置为NVIDIA GeForce RTX 2070 SUPER，内存为32 GB。

训练过程中，5个神经网络均使用了4层全连接网络，每一层的神经元个数设置为512，训练回合为1 500，每回合最大探索步数为100，模型的超参数设置如表1所示。

表1 模型超参数
Table 1 Hyper-parameters of model

参数名称	参数值
经验回放池尺寸	10^6
批量尺寸	256
奖励折扣因子	0.99
软更新因子	0.01
学习率	3×10^{-4}
目标熵	−2

使用本文算法在搭建的仿真环境中对机器人轴孔装配策略进行训练。奖励函数中，力和力矩的阈值1取值分别为$F_{\text{threshold1}}=50$、$T_{\text{threshold1}}=5$；奖励函数中的各项加权系数取值分别为$\delta_1=0.5$、$\delta_2=1$、$\delta_3=1\,000$。

终止条件参数：力的最大阈值$F_{\text{threshold2}}=80$ N，力矩的最大阈值$T_{\text{threshold2}}=8$ N·m；装配范围为装配孔上平面以孔位圆心为中心，长0.1 m、宽0.1 m、高0.013 m的长方体；轴姿态的最大翻滚范围为0.11 rad，最大俯仰范围为0.07 rad；当装配深度达到有效装配深度的±0.001 m时认为成功完成装配。训练采用的轴孔工件间隙为0，装配孔的深度设置为4 cm，孔口有倒角，轴无初始位置误差，完成一次训练的平均时长为6.5 h。

机器人在轴孔装配仿真环境中不断探索，以学习最优的装配策略。训练过程中回合奖励曲线如图5所示。当装配成功时获得的回合奖励较高，而装配失败时获得的回合奖励较低，经过1 500回合的训练，装配成功率不断提升，回合奖励逐渐收敛。

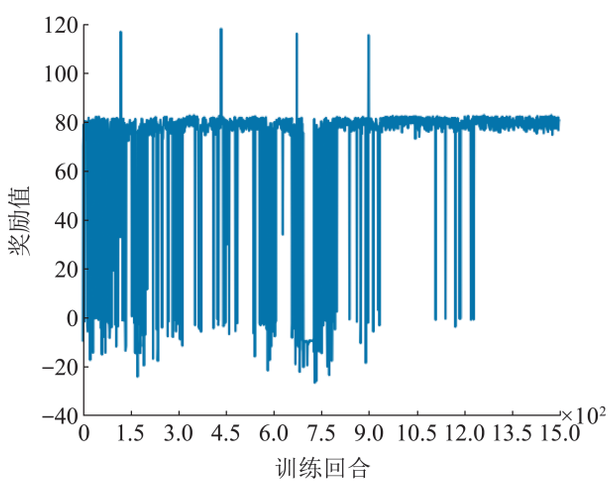


图5 训练时的回合奖励
Fig. 5 Reward of each episode during training

训练过程中的熵正则化系数α的变化曲线如图6所示。α的初始值为1，此时机器人可执行的动作随机性最大，随着训练回合的增加，为了完成轴孔装配任务获取最大化回报，机器人在与环境的交互中不断学习，α迅速减小至0.1左右，此时可执行的动作趋于确定。为了提高策略的抗干扰能力，α开始增大，鼓励机器人采取随机性更高的动作，通过不同的运动轨迹完成轴孔装配任务。但随机性过高的动作可能导致机器人无法完成装配任务，因此，α不断波动，在成功完成装配的前提下尽可能采取高度随机的动作，寻找在最大化回报策略中随机性最高的策略，最终使学习到的策略达到探索与利用之间的平衡。

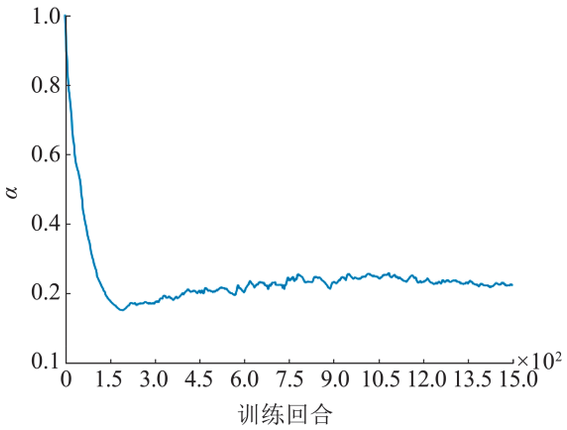


图6　训练时的α曲线
Fig. 6　Curve of α value during training

2.2 策略仿真验证

在搭建的轴孔装配仿真环境中对训练好的模型进行测试，验证策略的有效性和泛化性。仿真实验的参数设置与训练时相同。

为了验证装配策略的有效性，统计了机器人轴孔装配仿真过程中的接触力和力矩。在一个装配回合中，训练前后装配过程中力和力矩数据如图7所示。训练初始阶段，成功完成一次轴孔装配的过程中，力的绝对值最大可达40 N，力矩的绝对值最大可达6 N·m；使用提出的算法对机器人装配过程进行训练后，力稳定在2 N以下，力矩稳定在0.3 N·m以下。结果表明，经过训练后装配过程中的接触力大大减少，实现了柔顺轴孔装配的目的。

为了验证装配策略的泛化性，绘制了多个轴孔装配工件模型并导入Gazebo仿真环境中，从轴孔间的间隙、轴的初始位置误差、孔口有无倒角等角度设置了多组仿真实验，每组实验测试100个回合，统计了仿真过程中装配时间、装配成功率的均值，实验结果如表2所示，其中，A为轴孔间隙，A0表示轴孔间隙为0，A1表示轴孔间隙为0.4 mm；B为初始位置误差，B0表示初始位置误差为0，B1表示初始位置x、y、z坐标的误差在[−0.5 , 0.5]均匀分布中随机采样，单位为mm；C为孔口倒角，C0表示孔口无倒角，C1表示孔口有倒角。具体模型设置如图8所示。

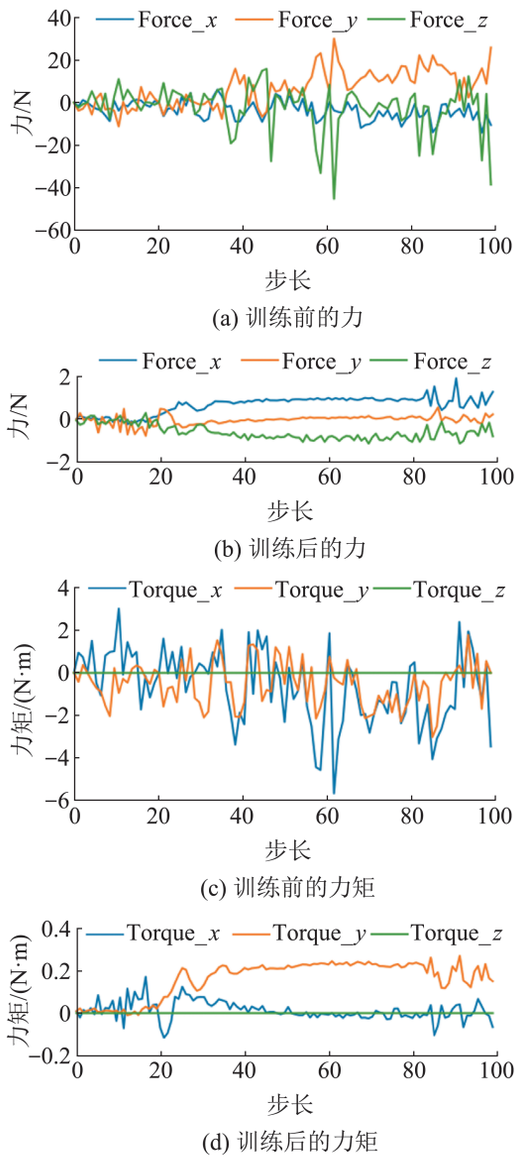
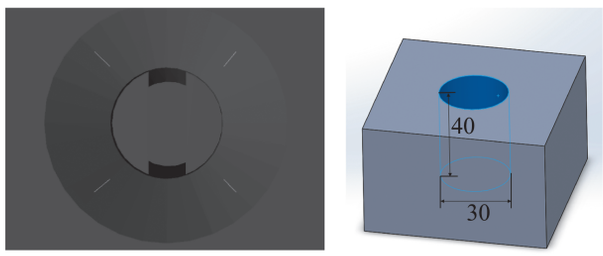


图7　装配过程的力和力矩
Fig. 7　Force and torque during assembly

表2　实验结果
Table 2　Experimental results

实验设置	装配时间/s	装配成功率/%
A0+B0+C0	6.27	100
A0+B0+C1	6.09	99
A0+B1+C0	6.33	100
A0+B1+C1	7.94	98
A1+B0+C0	5.88	100
A1+B0+C1	5.41	100
A1+B1+C0	6.01	100
A1+B1+C1	7.61	99

第 36 卷第 6 期 系统仿真学报 Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024



(a) 轴孔间隙为0.4 mm (b) 孔口无倒角

图8 仿真实验模型设置

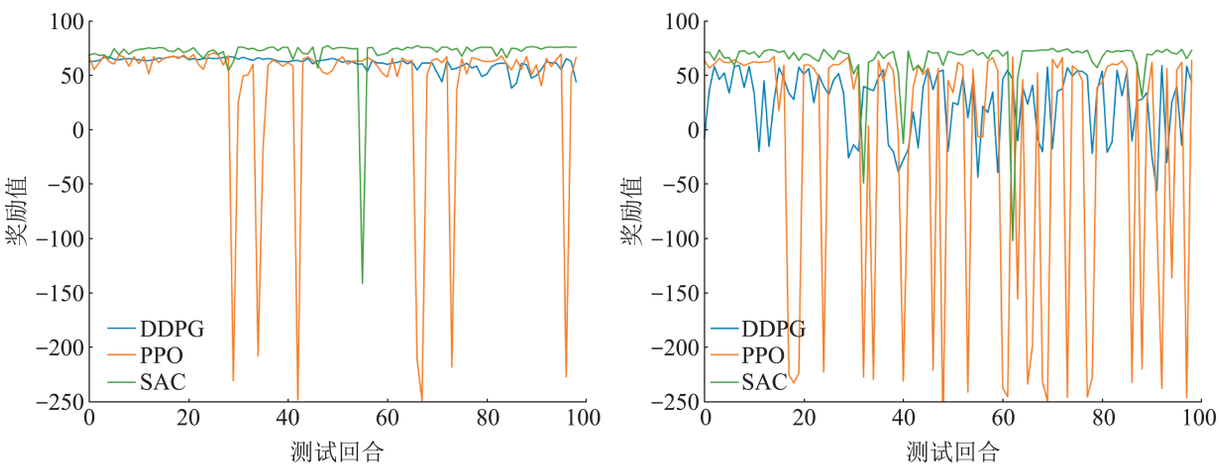
Fig. 8 Setting of simulation experiment model

仿真结果表明，训练的装配策略显著提高了轴孔装配的成功率和装配速度，且适用于不同的装配场景。在无初始位置误差的情况下，装配成功率达99%以上，初始位置误差在±0.5 mm的情况下，装配成功率达98%以上。装配过程中的接触力稳定在5 N以下，力矩稳定在0.5 N·m以下，达到了较好的柔顺装配效果。

此外，每个测试回合结束后只是控制机器人回到初始位姿。因此，在测试过程中装配轴的抓取位姿也会随着测试回合的增加而产生累积误差，但装配性能未受到影响，仍能安全稳定地完成装配任务。

2.3 策略对比分析

在仿真环境中，分别基于DDPG算法和PPO算法对机器人轴孔装配过程进行训练。状态空间、动作空间、奖励函数、终止条件与1.2.1节定义一致，训练场景与实验参数设置与2.1、2.2节相同。对训练好的模型进行测试。每组实验测试100个回合，测试过程的回合奖励如图9所示。测试结果如表3所示。



(a) 无初始位置误差 (b) 有初始位置误差

图9 测试过程的回合奖励

Fig. 9 Reward of each episode during testing

表3 各算法测试结果

Table 3 Test results of each algorithm

算法	训练时间/h	实验设置	装配成功率/%	装配时间/s	最大接触力/N	最大接触力矩/(N·m)
SAC	6.5	B0	99	6.09	5	0.5
		B1	98	7.94	5	0.7
DDPG	5.8	B0	100	3.25	50	7.0
		B1	74	3.56	65	8.0
PPO	4.3	B0	93	11.79	5	0.8
		B1	75	14.02	30	3.5

DDPG 算法使用了确定性策略，无需对动作概率分布进行采样，直接采用最大概率的动作，极大地提高了装配速度和装配成功率。但是学习到的策略陷入了局部最优解，只考虑到装配效率，并未实现柔顺装配的目的，装配过程中的接触力和力矩较大；同时，确定性策略导致算法对环境的探索能力差，固定的动作无法应对多样化的环境，当加入环境噪声后，装配成功率下降 26%，算法的鲁棒性较差。

PPO 算法是一种在线策略方法，探索能力较强，且通过限制策略的更新幅度保证模型的收敛性和稳定性；训练中能够在相对较少的迭代次数下收敛到较好的策略，具有相对较高的装配成功率，且对装配过程中的接触力和力矩实现了有效控制。但是学习到的策略装配效率较差，平均装配时间为 DDPG 算法的约 4 倍、SAC 算法的约 2 倍；加入环境噪声后，策略的性能下降，装配成功率下降 18%，装配过程中的最大接触力和力矩大幅增加，无法满足柔顺控制的需求。

SAC 算法是一种使用随机性策略的离线策略方法，结合了 DDPG 和 PPO 算法的优点，具有较强的策略探索能力和数据利用效率，解决了强化学习对超参数敏感的问题；通过最大熵思想鼓励多样性策略的探索，使机器人可以在相同的输入状态下通过不同的动作完成轴孔装配任务。SAC 算法在训练过程中需要寻求策略在探索和利用之间的平衡，使训练时间相对较长，但学习到的策略具有更强的鲁棒性和泛化性，能够有效应对扰动。加入环境噪声后，装配成功率仍可达 98%，且装配过程中具有较好的柔顺性，对接触力和力矩实现了有效控制，体现出本文算法的有效性和优越性。

3 结论

综上所述，本文提出了一种基于 DRL 的机器人轴孔装配策略仿真研究方法。基于 ROS-Gazebo 设计并搭建了一个机器人轴孔装配仿真系统，在仿真环境中提出了一个基于最小二乘法对力/力矩传感器进行重力补偿的方法；提出了一种基于深度强化学习 SAC 的机器人轴孔装配方法；基于 ROS 建立了各算法模块与仿真环境之间的通信机制，实现了对基于 DRL 的轴孔装配训练过程的仿真。通过优化动作空间的权重加强对轴孔装配动作的引导，考虑装配过程中的力和力矩信息，通过仿真探索最优的机器人轴孔装配策略，并设置多组实验对策略进行仿真验证和对比分析。仿真结果表明，搭建的机器人轴孔装配仿真系统保证了装配策略训练过程的安全性，提高了策略的训练效率，同时学习到的装配策略能够有效抑制控制精度误差造成的影响，可泛化至不同的装配场景，提高了装配效率，保证了装配过程的安全性和稳定性。

下一步研究将引入视觉信息，通过多传感器信息融合提高机器人对装配环境的感知能力，优化算法流程和状态空间设计，进一步提高训练过程中策略的学习效率，验证算法在高精度轴孔装配问题中的可行性和有效性。

参考文献：

[1] 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 Nailong, Liu Zhaoming, Cui Long. Deep Reinforcement Learning Based Robotic Assembly in Simulation[J]. Computer Simulation, 2019, 36(12): 296-301.

[2] Jiang Jingang, Huang Zhiyuan, Bi Zhuming, et al. State-of-the-art Control Strategies for Robotic PiH Assembly[J]. Robotics and Computer-Integrated Manufacturing, 2020, 65: 101894.

[3] Whitney D E. Quasi-static Assembly of Compliantly Supported Rigid Parts[J]. Journal of Dynamic Systems, Measurement, and Control, 1982, 104(1): 65-77.

[4] Xu Jing, Hou Zhimin, Wang Wei, et al. Feedback Deep Deterministic Policy Gradient with Fuzzy Reward for Robotic Multiple Peg-in-hole Assembly Tasks[J]. IEEE Transactions on Industrial Informatics, 2019, 15(3): 1658-1667.

DOI: 10.16182/j.issn1004731x.joss.23-0518

[5] 魏明明, 傅卫平, 蒋家婷, 等. 操作机器人轴孔装配的行为动力学控制策略[J]. 机械工程学报, 2015, 51(5): 14-21.
Wei Mingming, Fu Weiping, Jiang Jiating, et al. Dynamics of Behavior Control Strategy in Peg-in-hole Assembly Task of Manipulator[J]. Journal of Mechanical Engineering, 2015, 51(5): 14-21.

[6] 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 Chanjuan, Zhao Feifei, Li Cheng, et al. Multi-sensor Assisted Robotic Accurate Assembly[J]. Machinery Design & Manufacture, 2020(3): 281-284.

[7] 薛亚东, 陈庆盈, 尹建平. 基于力反馈的轴孔柔顺装配策略[J]. 自动化与仪器仪表, 2021(4): 152-155, 163.
Xue Yadong, Chen Qingying, Yin Jianping. Peg-in-hole Compliant Assembly Strategy Based on Force Feedback[J]. Automation & Instrumentation, 2021(4): 152-155, 163.

[8] Shirinzadeh B, Zhong Yongmin, Tilakaratna P D W, et al. A Hybrid Contact State Analysis Methodology for Robotic-based Adjustment of Cylindrical Pair[J]. The International Journal of Advanced Manufacturing Technology, 2011, 52(1): 329-342.

[9] 潘柏松, 颜天野, 胡鑫达, 等. 基于几何约束与隐马尔可夫链模型的轴孔装配策略[J]. 计算机集成制造系统, 2022, 28(12): 3766-3776.
Pan Baisong, Yan Tianye, Hu Xinda, et al. Peg-in-hole Assembly Strategy Based on Geometric Constraint and Hidden Markov Model[J]. Computer Integrated Manufacturing Systems, 2022, 28(12): 3766-3776.

[10] 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 Shuailong, Zhang Huiwen, Zhou Weijia. Review of Imitation Learning Methods and Its Application in Robotics[J]. Computer Engineering and Applications, 2019, 55(4): 17-30.

[11] Song Jingzhou, Chen Qingle, Li Zhendong. A Peg-in-hole Robot Assembly System Based on Gauss Mixture Model[J]. Robotics and Computer-Integrated Manufacturing, 2021, 67: 101996.

[12] Gao Xiao, Ling Jie, Xiao Xiaohui, et al. Learning Force-relevant Skills from Human Demonstration[J]. Complexity, 2019, 2019: 5262859.

[13] Tang Te, Lin H C, Zhao Yu, et al. Teach Industrial Robots Peg-hole-insertion by Human Demonstration[C]//2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). Piscataway, NJ, USA: IEEE, 2016: 488-494.

[14] Li Fengming, Jiang Qi, Zhang Sisi, et al. Robot Skill Acquisition in Assembly Process Using Deep Reinforcement Learning[J]. Neurocomputing, 2019, 345: 92-102.

[15] 王竣禾, 姜勇. 基于深度强化学习的动态装配算法[J]. 智能系统学报, 2023, 18(1): 2-11.
Wang Junhe, Jiang Yong. Dynamic Assembly Algorithm Based on Deep Reinforcement Learning[J]. CAAI Transactions on Intelligent Systems, 2023, 18(1): 2-11.

[16] Young-Loul Kim, Kuk-Hyun Ahn, Jae-Bok Song. Reinforcement Learning Based on Movement Primitives for Contact Tasks[J]. Robotics and Computer-Integrated Manufacturing, 2020, 62: 101863.

[17] 徐德, 秦方博. 机器人自动轴孔装配研究进展[J]. 智能科学与技术学报, 2022, 4(2): 200-211.
Xu De, Qin Fangbo. Research Development on Automated Robotic Peg-in-hole Assembly[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4 (2): 200-211.

[18] 黄玲涛, 王彬, 倪水, 等. 基于力传感器重力补偿的机器人柔顺控制研究[J]. 农业机械学报, 2020, 51(3): 386-393.
Huang Lingtao, Wang Bin, Ni Shui, et al. Robotic Compliant Control Based on Force Sensor Gravity Compensation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(3): 386-393.

[19] Deng Yuelin, Hou Zhimin, Yang Wenhao, et al. Sample-efficiency, Stability and Generalization Analysis for Deep Reinforcement Learning on Robotic Peg-in-hole Assembly[C]//International Conference on Intelligent Robotics and Applications. Cham: Springer International Publishing, 2021: 393-403.

[20] Feng Xiaoxin, Shi Tian, Li Weibing, et al. Reinforcement Learning-based Impedance Learning for Robot Admittance Control in Industrial Assembly[C]//2022 International Conference on Advanced Robotics and Mechatronics (ICARM). Piscataway, NJ, USA: IEEE, 2022: 1092-1097.

[21] Haarnoja T, Zhou A, Abbeel P, et al. Soft Actor-critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor[C]//Proceedings of the 35th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2018: 1861-1870.

[22] Haarnoja T, Zhou A, Hartikainen K, et al. Soft Actor-critic Algorithms and Applications[EB/OL]. (2019-01-29) [2023-04-26]. https://arxiv.org/abs/1812.05905.