

基于改进 TD3 算法的机械臂智能规划方法研究

张强, 文闻, 周晓东, 刘维惠, 初晓昱

(北京控制工程研究所, 北京 100190)

摘要: 针对某卫星搭载的4自由度机械臂轨迹规划问题, 提出了一种基于改进的双延迟深度确定性策略梯度(TD3)算法的智能规划方法。该方法采用分阶段训练策略, 在预训练阶段, 采用了目标位置引导联合TD3算法进行轨迹优化的混合规划策略, 训练结束后规划算法能够在机械臂关节空间对任意起点、终点进行速度轨迹的自主规划。这种目标引导机制减少了训练时不必要的探索, 在一定程度上解决了高维动作空间中学习效率低下的问题。在二次训练阶段, 首先通过示教获得一条无碰撞的安全参考轨迹, 然后在训练过程中不断对这条轨迹进行模仿, 使得最终算法输出的轨迹具备避障能力。

关键词: 避障规划; 目标引导; 轨迹示教; 双重训练

中图分类号: TP183

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202225

Research on the manipulator intelligent trajectory planning method based on the improved TD3 algorithm

ZHANG Qiang, WEN Wen, ZHOU Xiaodong, LIU Weihui, CHU Xiaoyu

Beijing Institute of Control Engineering, Beijing 100190, China

Abstract: An intelligent trajectory planning and obstacle avoidance method based on the improved twin delayed deep deterministic policy gradient algorithm (TD3) was proposed to solve the trajectory planning problem for a 4-DOF manipulator mounted on a satellite. The training strategy had 2 periods. In the pre-training stage, the target position was always guided combining with the output of the strategy network to optimize the trajectory. After the pre-training, the algorithm can autonomously output the velocity trajectory while the initial position and the target were specified randomly in the joint space of the manipulator. This target-guided mechanism decreased the unnecessary explorations and improved the learning efficiency in high dimensional action space. In the second training stage, a collision-free safety reference trajectory was firstly obtained by demonstration, and then this trajectory was constantly learned during the training process until the final output trajectory has the ability to avoid obstacles.

Key words: obstacle avoidance planning, target-guide, trajectory demonstration, double training

0 引言

随着机器人技术和空间控制技术的不断进步, 空间机械臂在卫星、空间站有了广泛的应用。近年来逐渐推广的完全采用电推进系统的卫星平台也越来越多地使用机械臂来完成电推力器的指向和位

置调节: 通过控制机械臂实施大范围运动, 使得机械臂末端推力器按照指定的方向喷出工质, 从而达到调整卫星位置、姿态的目的。相比传统的2自由度矢量调节机构, 机械臂拥有更大的运动范围, 能够更有效地完成位保、变轨和动量卸载等任务。国内外通常也称这类机械臂为可展开式推力矢量调

收稿日期: 2022-02-16; 修回日期: 2022-04-11

通信作者: 文闻, wenwenbice@163.com

基金项目: 科技创新 2030—“新一代人工智能”重大项目 (No.2018AAA0103004)

Foundation Item: The National Key Research and Development Program of China (No.2018AAA0103004)

节机构。目前在轨的诺斯罗普-格鲁曼公司的 MEV-1 卫星、欧洲通信卫星公司的 Eutelsat-172B 卫星和 SES-12 卫星^[1]都携带了类似的机械臂。但是,此类机械臂的自由度一般大于或等于 3,在卫星表面布满诸多载荷的环境中进行运动时,此类机械臂面临与传统用于捕获与操作任务的机械臂类似的避障规划问题。

对于空间机械臂规划问题,传统的快速扩展随机树(rapidly exploring random tree, RRT)算法^[2]、PRM(probabilistic roadmaps)算法^[3]以及二者衍生出的一系列改进算法已经可以满足当前的在轨需求。但是,随着人工智能技术的不断进步,未来空间技术的一个发展趋势是不断提高任务规划与控制决策过程的自主智能化水平。随着智能卫星、智能操控技术的不断发展,将智能控制算法尤其是深度强化学习算法用于机械臂控制也是当前的一个热门研究方向^[4]。本文主要针对某卫星平台搭载的 4 自由度机械臂开展智能规划与避障算法研究,从而实现机械臂的精准指向和安全避障功能,提高卫星推力调节控制的智能化水平。

1 算法介绍

1.1 算法原理分析

在空间指向任务中,4 自由度机械臂需要规划出一条从当前位置到目标指向位置的无碰撞路径,最终使得机械臂末端携带的推力器的推力中心点移动到指令要求的推力方向上,同时推力器产生的羽流方向与指令中的推力方向一致。卫星携带两条机械臂进行推力矢量调节示意图如图 1 所示。

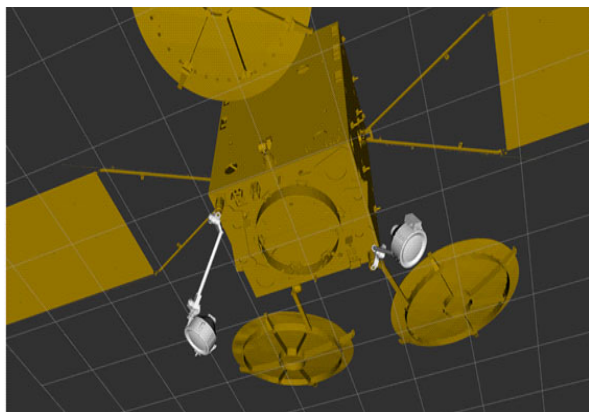


图1 卫星携带两条机械臂进行推力矢量调节示意图

从图 1 可以看出,卫星表面装有各类载荷,为了保证运动安全,需要在机械臂规划时考虑避障问

题。近几年来,随着深度学习技术的发展,许多基于深度强化学习(deep reinforcement learning, DRL)原理的机械臂智能控制方法^[5]被提出,如 Gu S X 等人^[6]利用异步离线策略共享更新机制,实现了在多个真实机械臂上的复杂技能学习;Zhu Y K 等人^[7]训练了端到端的视觉运动策略,完成了多个视觉伺服机械臂抓取规划任务;Sangiovanni B 等人^[8]提出了一种将深度强化学习算法和传统规划算法结合的混合机械臂规划算法,实现了点到点规划功能。上述研究成果表明,针对一些典型的工业机械臂或协作机械臂,已经具备将智能算法从仿真环境移植到真实机械臂上的能力,并可以取得不错的效果,但是在空间应用领域,考虑到算法的复杂度和硬件计算水平的限制,大部分基于 DRL 原理的空间机械臂规划控制算法还停留在仿真阶段。

常见的用于机械臂规划控制的 DRL 方法大致可分为基于值函数和基于策略梯度两大类^[4-5]。简单问题采用离散动作空间、基于值函数的方法就可以达到不错的效果。将这类方法由离散动作空间扩展到连续动作空间得到的归一化优势函数(normalized advantage function, NAF)^[6]算法目前已经在 MuJoCo 仿真环境下的一系列操作类任务和机器人控制类任务中进行了验证。结果表明,该算法不但可以在虚拟环境中解决机械臂对动态障碍物的避障问题^[9],还在真实机械臂上实现了算法迁移,例如可以使机械臂学会开门动作^[6]。对于简单的强化学习问题,基于值函数的方法可以在离散的动作空间使用探索机制尽可能地将“状态-动作”枚举出来,然后解决问题,但是在复杂的机械臂控制与规划领域通常面临状态连续(关节角度)、动作连续(控制力矩)、运动自由度高的问题,使得难以采用基于值函数的方法解决机械臂的控制规划问题。一方面动作空间的连续导致很难评估所有动作价值并选择价值最大的动作,另一方面也无法将所有的动作逐一尝试出来。基于策略梯度的方法则能够直接对策略进行优化,无须进行最优动作的选择,更适用于包含连续动作和状态空间问题的求解,因此在面对机械臂规划问题时,研究人员大多采用策略梯度算法进行求解。

基于策略梯度的方法中的策略又可分为随机策略和确定策略。对于随机策略方法,在给定状态和参数的情况下,输出的动作服从某一个概率分布;而对于确定策略方法,给定状态和参数,输出

的动作是确定的。目前,在机械臂控制领域,成功应用的随机策略方法有置信域策略优化(trust region policy optimization, TRPO)^[10]、近端策略优化(proximal policy optimization, PPO)^[11]等,但这类方法在面临高维动作空间时需要大量采样才能进行动作评估,此外选择深度神经网络表示策略时求解计算量较大。因此,对于机械臂这类控制对象而言,随机策略并不是一个很好的选择。

对于机械臂实际规划控制问题通常使用基于 AC(actor-critic)框架的各类算法来解决,即包含了策略网络、Q 网络以及确定策略的算法。该类算法在高维连续任务中表现良好,因此被大量用于机械臂规划任务。这类算法集成了值函数估计和策略搜索,常用的算法包括异步优势动作评价(asynchronous advantage actor-critic, A3C)^[12]、确定性策略梯度(deterministic policy gradient, DPG)^[13]、深度确定性策略梯度(deep deterministic policy gradient, DDPG)^[14]、分布式深度确定性策略梯度(distributed distributional deep deterministic policy gradient, D4PG)^[15]、双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)^[16]等。AC 框架主要提供了一种动作与评价同时进行的并行结构,可以同时并行执行多个训练任务,并提供了训练经验共享机制,这大大提高了计算效率、样本利用率和计算资源使用率。大量研究人员在算法结构、网络更新机制、经验回放机制上不断优化创新,使用于连续动作的深度强化学习算法的性能越来越强。

目前,对于同时具有连续动作和状态空间的问题,DDPG 是一种相对简单、兼容性较好的算法,它是 Lillicrap T P 等人^[14]将简单 Q 学习算法扩展到 DQN(deep Q-network)算法的思路,是对 DPG 算法进行改进的一种算法,能够较好地解决连续动作空间上的强化学习问题。DDPG 算法的优势在于实现了连续动作空间控制,策略输出即动作(如机械臂的关节控制扭矩、速度、位置等连续变量)。相对于随机策略梯度方法,在动作空间维数很大时,确定性策略梯度方法不需要采集很多样本就能对策略进行评估,因此,DDPG 算法近年来在机器人控制领域取得了不错的效果。根据 DeepMind 公司的研究成果^[17],在大多数机器人控制任务中(如机器人站立、行走、机械臂抓取等),DDPG 算法在训练收敛速度方面要优于 A3C 算法,同样,在一些标准的

机器人控制任务中,DDPG 算法也要优于 PPO 和 TRPO 算法。因此,近年来,在机械臂规划领域,DDPG 算法也有了非常广泛的应用^[18-19]。尤其是 Qureshi A H 等人^[18]提出了一种基于神经网络的新型通用避障规划算法框架,通过对工作空间中的点云信息进行编码来获得障碍物信息,从而实现快速避障。但是,DDPG 算法也存在一些不足,一个典型的问题是会产生 Q 函数值过估计,从而影响策略网络的更新。

本文在研究时注意到,在针对某一具体问题时,新提出的、针对某种算法的改进算法往往具备更好的性能,例如 DDPG 算法的改进算法——TD3 算法和 D4PG 算法,它们在一些机器人控制任务中的表现就优于标准的 DDPG 算法。TD3 算法是一种无模型、在线、异策略的强化学习方法,相对于 DDPG 算法有 3 点改进:一是学习两个 Q 函数,并且每次以较小的那个 Q 值进行网络参数的更新;二是更新策略网络的频率比更新 Q 网络的频率低;三是给目标策略网络引入了平滑机制,在产生动作时加入噪声,使得在学习过程中尽可能获得更准确的 Q 值。因此,TD3 算法相比 DDPG 算法在理论上更具备鲁棒性,例如,Min C H 等人^[20]将经验回放(hindsight experience replay, HER)与 TD3 算法结合,通过设置稀疏奖励来学习机械臂控制策略,获得了不错的效果。标准 TD3 算法的流程如下。

算法 1 TD3 算法流程

- 1) 初始化在线 Q 网络 Q_{θ_1} 、 Q_{θ_2} 和在线策略网络 μ , 其参数分别为 θ_1^Q 、 θ_2^Q 和 θ^μ
- 2) 初始化目标网络 Q'_{θ_1} 、 Q'_{θ_2} 和 μ' , 神经网络的初始参数与 Q_{θ_1} 、 Q_{θ_2} 、 μ 一致
- 3) 初始化回放缓冲区 R
- 4) **for** 回合数=1: M **do**
- 5) 每一回合开始前,定义当前回合噪声 N_t , 开始探索观测环境的状态 s_t
- 6) **for** $t=1:T$ **do**
- 7) 根据当前的策略和动作噪声 ε 选择一个动作:

$$a_t = \text{clip}(\mu(s_t | \theta^\mu) + \varepsilon)$$
- 8) 执行动作 a_t 并观测回报 r_t 及新的环境状态 s_{t+1}

在 R 中存储一个状态转移过程

- (s_i, a_i, r_i, s_{i+1})
- 9) 从 R 中采集一个包含 N 个状态转移过程 $\{(s_i, a_i, r_i, s_{i+1}) | i=1, 2, \dots, N\}$ 的小样本集

$$a' = \text{clip}(\mu(s_i | \theta^\mu) + \text{clip}(\varepsilon))$$

$$y = r + \gamma \min_{i=1,2} Q'_{\theta_i}(s_{i+1}, a')$$
 - 10) 更新在线 Q 网络参数

$$\theta_i \leftarrow \arg \min_{\theta_i} \frac{1}{N} \sum (y - Q_{\theta_i})^2 \quad (0 < \gamma < 1,$$

$$\gamma \text{ 为折扣因子})$$
 - 11) if 满足延迟更新条件
 - 12) 更新在线策略网络:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q_{\theta_i}(s, a) \big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s | \theta^\mu) \big|_{s_i}$$
 - 13) 按如下原则更新目标网络:

$$\theta_i^{Q'} \leftarrow \tau \theta_i^Q + (1 - \tau) \theta_i^{Q'}$$

$$\theta^{\mu'} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'}$$
 - 14) end if
 - 15) end for
 - 16) end for

本文在研究机械臂智能规划算法时,综合考虑了算法解决通用问题的能力和训练时的收敛速度,最终利用 TD3 算法能够解决连续状态、动作空间问题的优势和优于 DDPG 算法的鲁棒性,提出了一种用于 4 自由度机械臂的智能规划方法。在 TD3 算法框架的基础上,增加了目标引导功能,并采用了双重训练机制。首先在仿真环境中进行预训练以实现机械臂关节空间的点到点任意规划,然后根据示教出的安全轨迹进行二次训练,通过模仿示教轨迹实现机械臂的避障功能。

1.2 目标引导与 TD3 算法的联合框架

对于机械臂而言,其动作空间维度较高,直接使用 TD3 算法会使得训练时机械臂运动随机性太大,难以保证算法的稳定性。针对这种问题,Zhong J 等人^[19]给出了一个很好的思路,即在训练初期利用机械臂逆运动学方法计算出一条参考轨迹,将其作为专家指导。这种混合式的策略在机械臂智能控制中比较常见,例如 Xu J 等人^[21]在解决轴孔智能装配问题时,将阻抗控制算法作为专家控制策略以获得基本的“轴-孔”装配轨迹,然后采用 DDPG 算法学习并修正该轨迹,从而获得智能柔顺装配策略。因此,本文同样采用了“专家指导+智能算法优化”的联合规划思想,这里“专家指导”指的是在规划

的全过程中,根据目标和当前关节角的差值计算出一条参考轨迹,从而降低神经网络在探索初期的盲目性,而“智能算法优化”指的是将 TD3 算法的输出作为轨迹的修正量,进而显著提高局部轨迹的优化效果。本文将关节速度 $\mathbf{v}$ 作为智能规划算法的最终输出,这是因为尽管动作空间连续,但 TD3 算法每步输出的动作未必连续,如果将关节位置作为轨迹控制量,可能会造成最终输出轨迹存在大量阶梯式的跳变。在本文中,目标引导的速度轨迹 $\mathbf{v}_{\text{ref}}$ 和 TD3 算法输出的速度修正量 $\mathbf{v}_a$ 之间通过一定的权重 w_1 、 w_2 进行分配,如式(1)所示。本文设计的增加了目标引导功能的 TD3 算法框架如图 2 所示,包含了完整的 TD3 算法和目标引导速度生成模块。

$$\mathbf{v} = w_1 \mathbf{v}_{\text{ref}} + w_2 \mathbf{v}_a \quad (1)$$

1.3 预训练及二次训练策略

按照图 2 的算法框架进行训练后,理论上可以实现 4 自由度机械臂在关节空间点到点规划功能。由于在训练过程没有提供障碍物信息,策略网络输出的轨迹不具备避障效果,甚至先后两次独立的训练完成后,对于相同的起点和终点,策略网络仅仅能够实现终点到达功能,规划出来的轨迹中间点可能完全不同,因为在关节空间不考虑约束的情况下,起点和终点之间可能存在无数条轨迹。

在各类机械臂避障规划算法中,如果不考虑利用传感器对障碍进行测量感知,然后进行实时规划,那么通过人类示教获得一条无碰撞的安全路径供机械臂参考学习是一种有效的方案。例如,Zhu ZY 等人^[22]研究了从人类示教演示中学习动作来提高机械臂操控的自主性,刘维惠等人^[23]利用动态动作基元模型学习与示教轨迹形状类似的新轨迹,并使用 K 近邻(K nearest neighbor, KNN)算法实现对机械臂轨迹的局部修正,而 Abdo N 等人^[24]提出了一种通过少量示教即可学习操作技能的方法,并给出了该方法的前提条件和实施效果。

受上述研究的启发,本文将规划算法的训练分为两个阶段:第一阶段(预训练)对关节空间点到点规划策略进行学习,使得机械臂能够在任意给出的起点与终点之间规划出一条连续的路径;第二阶段(二次训练)主要针对有障碍环境进行训练,首先示教出一条无碰撞的安全轨迹,然后对这条安全轨迹进行学习,最终实现避障。

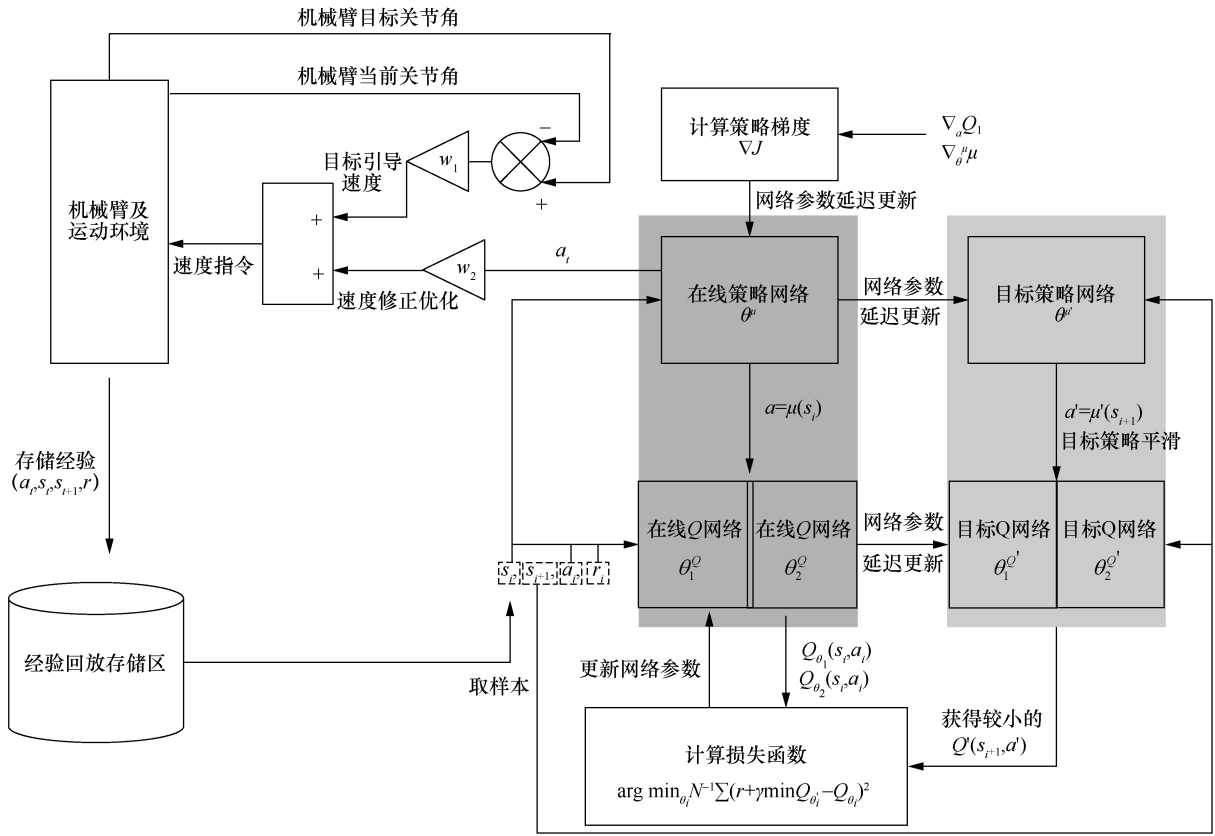


图 2 增加了目标引导功能的 TD3 算法框架

在二次训练中，首先通过人工示教获得一系列安全路径点，然后利用多项式插值生成一条光滑密集的轨迹，并将其作为下一步训练的参考。接下来，将预训练获得的所有策略网络和 Q 网络参数作为二次训练前网络参数的初值，同时将示教轨迹起点和终点提供给规划算法进行反复迭代训练。训练中，设置了合适的奖励函数以鼓励模仿示教轨迹的动作输出，从而驱使策略网络最终输出的轨迹与示教插值生成的轨迹尽可能接近。这样经过二次训练，策略网络就可以规划出一条安全无碰撞的路径。这种双重训练思想和实际机械臂使用情况更加接近，即当新任务、新场景出现时，先前训练的经验结果并没有被完全舍弃，而是在二次训练中作为一个基本策略被不断优化，直至规划算法能够满足新任务、新场景的需求。同时，由于不需要实时测量障碍物信息并建模，仅通过示教轨迹描述障碍物的原理也降低了算法的复杂度。

2 智能规划算法设计

2.1 神经网络结构设计

基于确定策略梯度原理设计机械臂智能避障

规划算法，本质上就是在训练结束后，将训练好的策略网络作为一个独立的神经网络控制器应用到机械臂的规划过程中。在本文中，为了满足通用的机械臂规划控制需求，将该神经网络的输入定义为机械臂自身能够通过传感器测量的物理量及可计算获得的变量，包括关节当前位置、目标位置、当前位置和目标位置在关节空间的距离，神经网络输出为机械臂关节速度修正量。本文以 4 自由度机械臂为研究对象，策略网络动作输出的维度为 4，这里将当前关节角向量和目标关节角向量差值的二范数定义为两组关节角之间的距离。

图 3 和图 4 分别为本文设计的 Q 网络结构和策略网络结构，算法中涉及的 4 个 Q 网络结构完全一致，2 个策略网络的结构也完全一致，每个网络均包含 3 个线性的全连接隐层，其中目标网络的结构和在线网络结构完全一致。未来若针对不同自由度的机械臂训练智能规划算法，仅调整神经网络的规模即可。

Q 网络结构如图 3 所示，输入层维数为需要观测的状态量的个数加机械臂输出动作个数，输入内容包含关节当前位置、目标位置、二者距离以及策略网络输出的速度修正量。针对本文连续动作空间

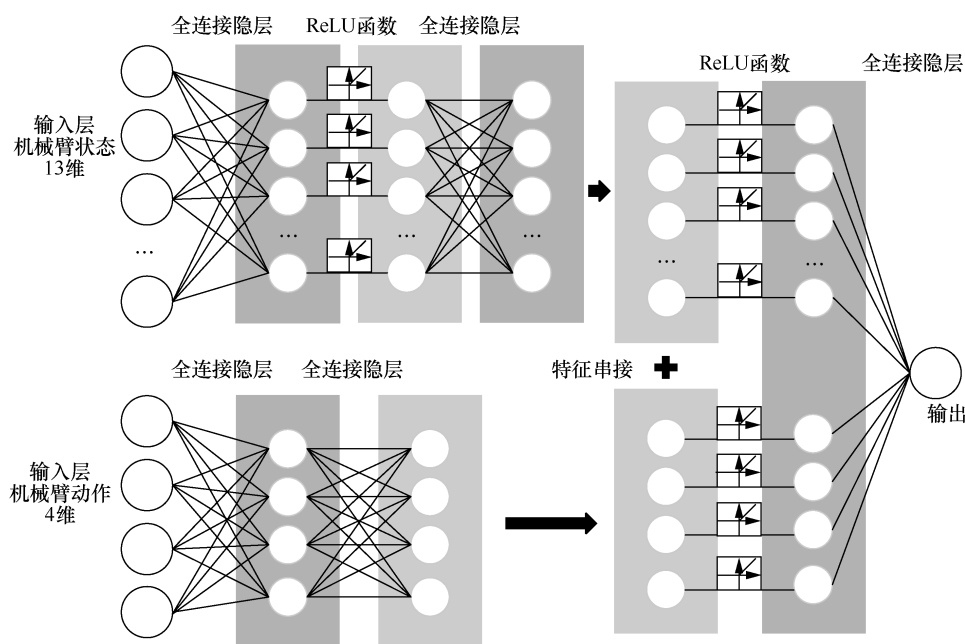


图3 Q网络结构

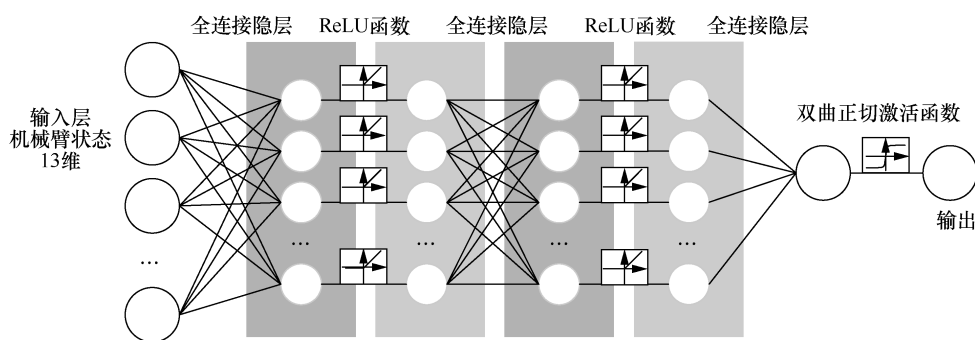


图4 策略网络结构

控制问题,在常见的用于机械臂规划的策略网络^[14,19]中,神经元数量通常为100~300。按照本文的设计,需要对神经网络进行两次训练,为了保证神经网络具有足够的泛化能力,同时综合考虑了状态空间和动作空间的维度,将各层的神经元数量分别设为300、200、200,经验证可以取得不错的效果。Q网络输出层维数为1,代表“状态-动作”的价值。各隐层的激活函数采用ReLU函数,输出层不设置激活函数。

策略网络结构如图4所示,输入层维数仅包含关节当前位置、目标位置以及二者距离,各层的神经元数量设置为200,输出层维数为4,输出为关节速度的修正量。各隐层的激活函数采用ReLU函数,输出层采用双曲正切激活函数。

2.2 奖励函数设计

(1) 预训练时奖励函数设置

预训练的主要目标是使得策略网络学会在关

节空间规划出一条连续无突变的轨迹。本文中,奖励函数的设计对于训练结果具有至关重要的作用,相对简单的奖励函数也可以获得很好的学习效果^[25]。本文在预训练阶段设计奖励函数时,主要思想是鼓励机械臂从当前位置不断趋近于目标位置,奖励函数 r 的形式如式(2)所示:

$$r = -\|\mathbf{a}\|_2 + r_d + r_t \quad (2)$$

$$r_d = -0.5d^2 \quad (3)$$

式(2)中, $\mathbf{a}$ 为策略网络输出动作向量。式(2)的第1项限制了策略网络产生的速度修正量不会过大,防止生成的关节空间轨迹产生突变;第2项 r_d 用来奖励机械臂靠近目标关节角的行为,即越远离目标,惩罚越大。 d 表示当前关节角和目标关节角之间的距离,只有当 $|d| \leq 0.1$ 时,认为机械臂到达目标,此时会获得一次性奖励 r_t ,本文将一次性奖

励值设置为 5 000。

在训练过程中, 达到以下任意一种状态时, 回合结束:

- ① $|d| \leq 0.1$, 即当前关节角与目标关节角的距离小于给定阈值时, 认为机械臂到达目标;
- ② 4 个关节角度超出给定阈值;
- ③ 单个回合中迭代的次数到达 1 000 次。

(2) 二次训练时奖励函数设置

经过预处理, 智能规划算法学会了点到点运动, 但是, 生成的轨迹并不能保证在运动过程中实现避障。因此, 本文在二次训练时需要设计奖励函数, 使得策略网络学会生成与示教轨迹近似的新轨迹。二次训练时的奖励函数针对学习单条轨迹做了以下修改:

$$r = -\|a\|_2 + r_d + r_t + r_{\text{ref}} \quad (4)$$

$$r_{\text{ref}} = -0.5(\|q'_i - q_i^{\text{ref}}\|_2)^2, \quad i \leq N \quad (5)$$

$$r_{\text{ref}} = -0.5(\|q'_i - q_N^{\text{ref}}\|_2)^2, \quad i > N \quad (6)$$

式(4)中的 r_{ref} 为二次训练规划出的路径点和示教产生的路径点之间的偏差。式(5)中, $i=1,2,3,\dots,N$ 为示教轨迹中路径点的编号, N 为示教轨迹包含路径点的个数, 变量 q'_i 为二次训练时的第 i 个轨迹点, q_i^{ref} 为参考轨迹中对应的第 i 个轨迹点, q_N^{ref} 为参考轨迹的最后一个轨迹点。这里需要注意的是, 预训练后对于不同起点和终点, 规划算法最终输出的轨迹点个数是不确定的, 当 $i \leq N$ 时, r_{ref} 按照式(5)计算奖励; 而当 $i > N$ 时, r_{ref} 按照式(6)计算奖励。

3 算法仿真验证

本文搭建了如图 5 所示的仿真环境, 用于智能规划算法的训练, 利用 MATLAB 软件的神经网络工具箱和强化学习工具箱实现了 TD3 算法中策略网络和 Q 网络的构建。机械臂运动学模型依托 Simulink 工具开发, 主要提供机械臂关节位置积分器、关节约束条件等功能。按照双重训练思想, 首先进行关节空间点到点的预训练。仿真时步长设为 0.01 s, 而机械臂单条轨迹最大持续时间为 10 s。预训练时, 在机械臂关节空间随机选取起点和终点, 考虑到机械臂各关节极限为 $\pm 180^\circ$, 这里将策略网络的输出幅值限定在 $\pm 1 \text{ rad/s}$, 从而保证机械臂在 10 s 内具有足够的速度到达任意终点, 不会出现因关节速度不足永远无法达到目标点的情况。同时, 算法

限制了目标引导速度在最终速度指令中的比重, 将式(1)中的 w_1 和 w_2 分别设置为 0.01 和 1, 防止算法仅依靠目标引导速度实现规划。此外, 无论是预训练还是二次训练, 图 5 中的学习框架均保持不变, 预训练获得的各个网络参数直接作为二次训练时网络的初值。

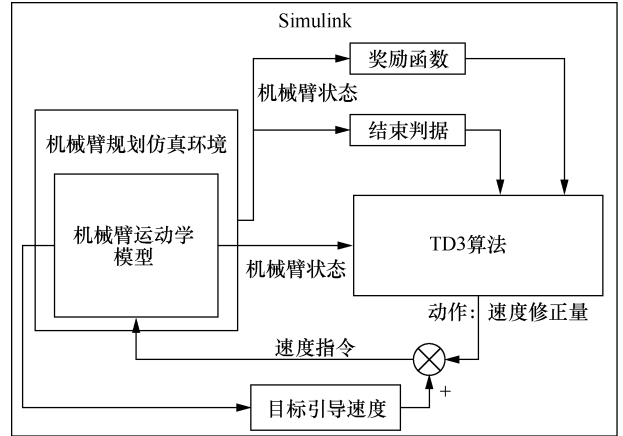


图 5 仿真环境及学习框架

图 6 给出了预训练过程的收益, 在预训练期间, 基于目标引导策略在每一步均给出了趋向目标位置的参考速度, 同时, 本文也按照图 2 的框架尝试取消目标引导速度进行训练。从图 6 可以看出, 本文的目标引导机制能够显著加速训练收敛。

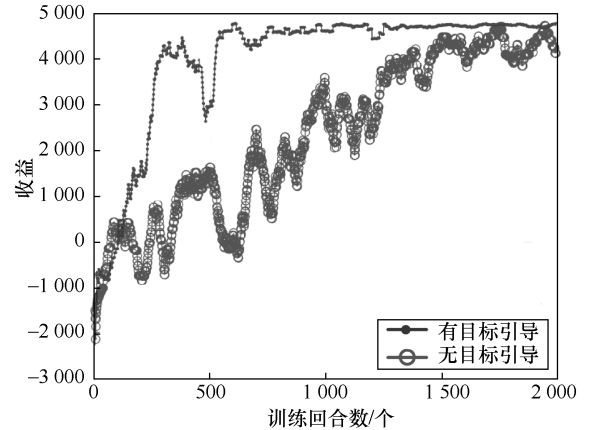


图 6 存在和取消目标引导速度的预训练收益

预训练完成后, 智能算法能够保证机械臂对于任意起点和终点均能规划出一条路径, 随机给出一组起点关节角 $q_{\text{start}} = [1.146 \ -0.055 \ 1.007 \ -0.193]$ 和终点关节角 $q_{\text{goal}} = [0.411 \ 0.976 \ 1.303 \ -0.273]$ (单位: rad)。预训练后单次随机规划结果如图 7 所示。从图 7 可以看出, 轨迹最终到达了目标, 同时在速度空间进行轨迹规划能够保证轨迹的光滑性。

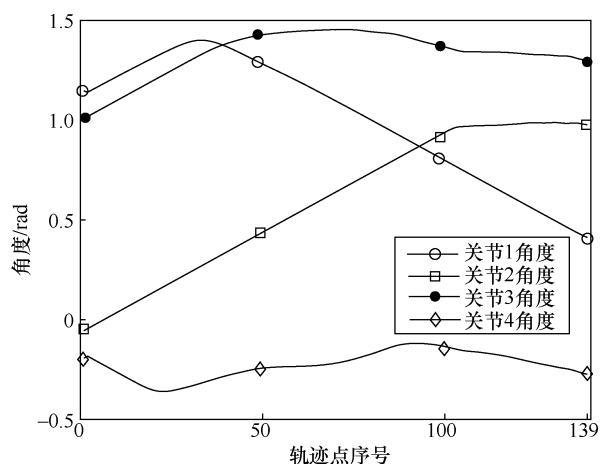


图7 预训练后单次随机规划结果

二次训练时的收益如图8所示。从图8可以看出,预训练后,经过约500个回合的训练后收益逐渐稳定,此时认为已经完成了对示教无碰撞轨迹的学习。图9~图12分别给出了相同起点和终点条件下,关节1~关节4的预训练后轨迹、示教轨迹和二次训练后轨迹。

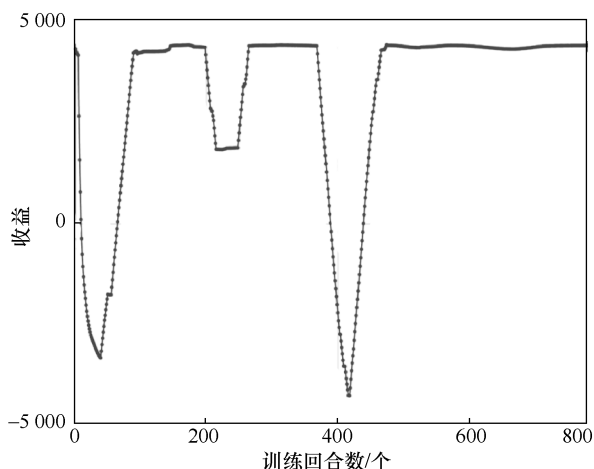


图8 二次训练时的收益

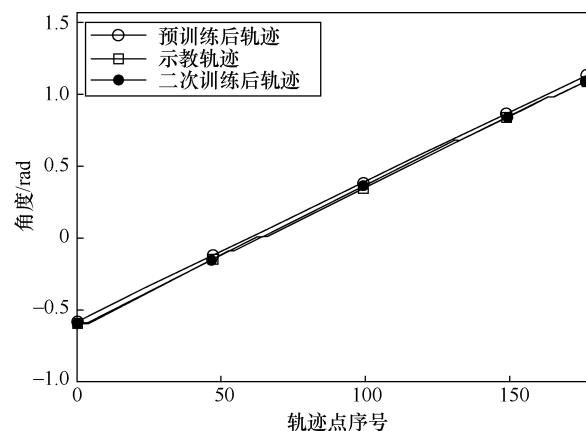


图9 关节1的预训练后轨迹、示教轨迹及二次训练后轨迹

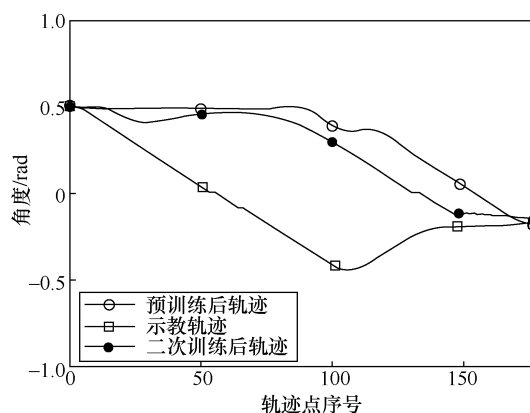


图10 关节2的预训练后轨迹、示教轨迹及二次训练后轨迹

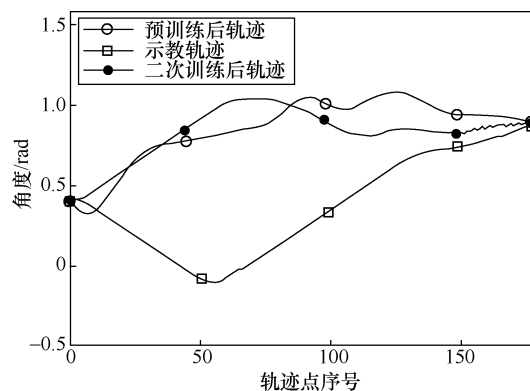


图11 关节3的预训练后轨迹、示教轨迹及二次训练后轨迹

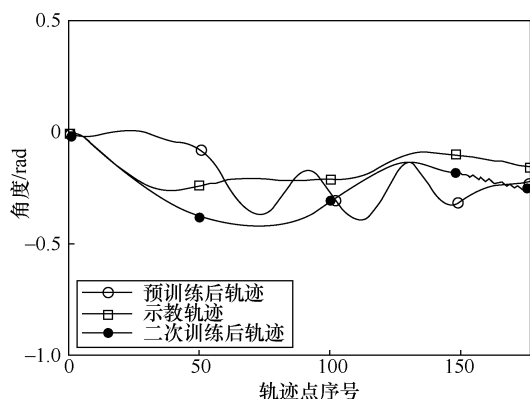


图12 关节4的预训练后轨迹、示教轨迹及二次训练后轨迹

可以看出,预训练后算法输出的关节2、3、4轨迹与示教轨迹有较大区别,而经过二次训练后,算法输出的轨迹更加接近示教轨迹。同时也可以看出,二次训练通过修改奖励函数对轨迹进行优化的策略并不是要求算法最终实现对示教参考轨迹的完全模仿,而是通过不断调整网络参数尽可能找到一条长度和示教轨迹接近且各个轨迹点也尽可能接近示教轨迹的新轨迹。

图13在笛卡儿空间绘制出机械臂末端推力器中心点的示教轨迹、预训练后轨迹和二次训练后轨

迹,可以看出,预训练后机械臂末端推力器的轨迹(蓝色)与卫星本体非常接近,在运动过程中臂杆会与卫星本体发生碰撞,而示教轨迹(红色)能够使机械臂末端、臂杆远离卫星本体,经过二次训练后,机械臂自主规划的新轨迹(绿色)与示教轨迹较接近,均远离卫星本体。因此,可以认为通过对网络参数进行调整和优化,二次训练能够较好地模仿示教轨迹,实现了期望的避障效果。同时也可以看出,二次训练后策略网络输出轨迹和示教轨迹之间仍然存在一些偏差,这与二次训练的奖励函数设置有关,即奖励函数优化的目标是使得策略网络输出轨迹和示教轨迹的整体偏差最小。

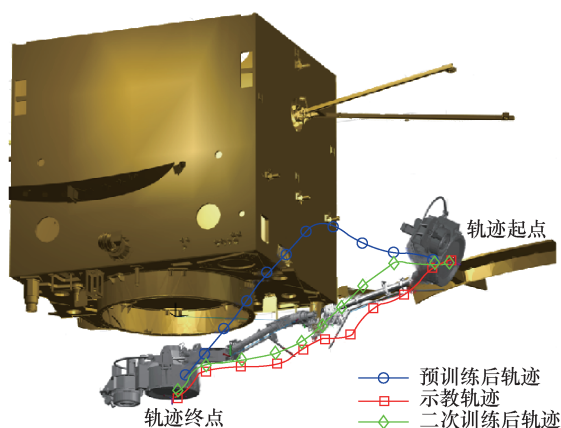


图13 笛卡儿空间中的机械臂末端中心点轨迹

4 结束语

本文提出了一种机械臂智能规划与避障算法,在 TD3 算法框架的基础上,引入了目标全程引导策略,使得策略网络能够快速实现点到点的关节速度级自主规划。在此基础上,引入双重训练机制,以示教出的安全轨迹为参考,通过修改奖励函数,以预训练获得的网络参数为网络初值,在二次训练中对参考轨迹进行模仿学习,从而使得机械臂最终能够规划出一条与参考轨迹接近的安全路径。仿真结果表明,这种不依赖机械臂运动学模型和障碍几何信息的智能规划方法达到了预期的效果。下一步的工作将主要研究如何进一步提高参考轨迹学习的精度,尤其是在与障碍物较近的位置如何保证轨迹安全。未来,随着各类高性能计算芯片、智能控制算法逐渐完成空间搭载试验和测试,卫星控制计算机的智能计算水平将会比当前阶段有很大提升,本文提出的方法也将有机会实现在轨测试和部署。

参考文献:

- [1] 胡照,王敏,袁俊刚. 国外全电推进卫星平台的发展及启示[J]. 航天器环境工程, 2015, 32(5): 566-570.
- [2] HU Z, WANG M, YUAN J G. A review of the development of all-electric propulsion platform in the world[J]. Spacecraft Environment Engineering, 2015, 32(5): 566-570.
- [3] DONALD B, LYNCH K, RUS D. Algorithmic and computational robotics[M]. [S.l.]: A K Peters/CRC Press, 2001.
- [4] KAVRAKI L E, SVESTKA P, LATOMBE J C, et al. Probabilistic roadmaps for path planning in high-dimensional configuration spaces[J]. IEEE Transactions on Robotics and Automation, 1996, 12(4): 566-580.
- [5] AMARJYOTI S. Deep reinforcement learning for robotic manipulation-the state of the art[J]. arXiv preprint, 2017, arXiv:1701.08878.
- [6] 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.
- [7] DUO N X, LYU 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.
- [8] GU S X, HOLLY E, LILLICRAP T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]// Proceedings of 2017 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2017: 3389-3396.
- [9] ZHU Y K, WANG Z Y, MEREL J, et al. Reinforcement and imitation learning for diverse visuomotor skills[J]. arXiv preprint, 2018, arXiv:1802.09564.
- [10] SANGIOVANNI B, INCREMONA G P, PIASTRA M, et al. Self-configuring robot path planning with obstacle avoidance via deep reinforcement learning[J]. IEEE Control Systems Letters, 2021, 5(2): 397-402.
- [11] SANGIOVANNI B, RENDINIELLO A, INCREMONA G P, et al. Deep reinforcement learning for collision avoidance of robotic manipulators[C]//Proceedings of 2018 European Control Conference. Piscataway: IEEE Press, 2018.
- [12] KURUTACH T, CLAVERA I, DUAN Y, et al. Model-ensemble trust-region policy optimization[J]. arXiv preprint, 2018, arXiv:1802.10592.
- [13] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint, 2017, arXiv:1707.06347.
- [14] ZHANG Y Z, CLAVERA I, TSAI B, et al. Asynchronous methods for model-based reinforcement learning[J]. arXiv preprint, 2019, arXiv:1910.12453.
- [15] SILVER D, LEVER G, HEES N, et al. Deterministic policy gradient algorithms[C]//Proceedings of the 31st International Conference on Machine Learning. [S.l.:s.n.], 2014: 387-395.
- [16] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint, 2015, arXiv:

1509.02971.

- [15] BARTH-MARON G, HOFFMAN M W, BUDDEN D, et al. Distributed distributional deterministic policy gradients[J]. arXiv preprint, 2018, arXiv:1804.08617.
- [16] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[J]. arXiv preprint, 2018, arXiv:1802.09477.
- [17] TASSA Y, DORON Y, MULDAL A, et al. DeepMind control suite[J]. arXiv preprint, 2018, arXiv:1801.00690.
- [18] QURESHI A H, SIMEONOV A, BENY M J, et al. Motion planning networks[C]//Proceedings of 2019 International Conference on Robotics and Automation. Piscataway: IEEE Press, 2019: 2118-2124.
- [19] ZHONG J, WANG T, CHENG L L. Collision-free path planning for welding manipulator via hybrid algorithm of deep reinforcement learning and inverse kinematics[J]. Complex & Intelligent Systems, 2021: 1-14.
- [20] MIN C H, SONG J B. End-to-end robot manipulation using demonstration-guided goal strategies[C]//Proceedings of 2019 16th International Conference on Ubiquitous Robots. Piscataway: IEEE Press, 2019: 159-164.
- [21] XU J, HOU Z M, WANG W, et al. Feedback deep deterministic policy gradient with fuzzy reward for robotic multiple peg-in-hole assembly tasks[C]//Proceedings of the IEEE Transactions on Industrial Informatics. Piscataway: IEEE Press, 2019: 1658-1667.
- [22] ZHU Z Y, HU H S. Robot learning from demonstration in robotic assembly: a survey[J]. Robotics, 2018, 7(2): 17.
- [23] 刘维惠, 陈殿生, 张立志. 利用示教学习的移动机械臂轨迹避障算法[J]. 哈尔滨工程大学学报, 2018, 39(9): 1546-1553.
LIU W H, CHEN D S, ZHANG L Z. Learning from demonstration based obstacle avoidance algorithm to plan the trajectory of a mobile manipulator[J]. Journal of Harbin Engineering University, 2018, 39(9): 1546-1553.
- [24] ABDO N, KRETZSCHMAR H, SPINELLO L, et al. Learning manipulation actions from a few demonstrations[C]//Proceedings of 2013 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2013: 1268-1275.
- [25] HEESS N, TB D, SRIRAM S, et al. Emergence of locomotion behaviours in rich environments[J]. arXiv preprint, 2017, arXiv:1707.02286.

[作者简介]



张强 (1978—), 男, 博士, 北京控制工程研究所机电技术事业部部长, 中国航天科技集团有限公司科技委惯性专业组成员、研究员, 主要研究方向为航天器各类驱动机构、执行机构、空间操控机械臂等。



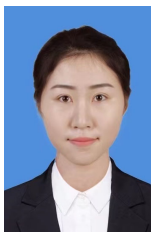
文闻 (1985—), 男, 博士, 北京控制工程研究所高级工程师, 主要研究方向为空间机械臂运动规划与智能操控技术。



周晓东 (1983—), 男, 博士, 北京控制工程研究所高级工程师, 主要研究方向为仿生机器人、机器人交互控制技术。



刘维惠 (1988—), 女, 博士, 北京控制工程研究所高级工程师, 主要研究方向为空间机械臂智能交互技术。



初晓昱 (1992—), 女, 博士, 北京控制工程研究所工程师, 主要研究方向为空间机械臂智能规划、地外行星探测机器人控制技术。