DOI:10.16356/j.1005-2615.2025.03.008

基于强化学习的双臂空间机器人应急姿态控制

黎 丰[1,2],李 宁[1,2],邹怀武[1,2],靳永强[1,2],张崇峰[1,2]
(1.宇航空间机构全国重点实验室,上海 201108;2.上海宇航系统工程研究所,上海 201108)

摘要: 针对双臂空间在轨服务机器人在遭遇极端异常情况下飞轮、发动机故障等传统姿态控制失效问题,提出了一种基于强化学习的双臂空间机器人应急姿态控制算法。与传统姿态控制算法相比,本文仅通过飞行器所配置的两条机械臂进行有限的飞行器姿态恢复。通过搭建算法训练的物理环境,应用无模型的近端策略优化(Proximal policy optimization,PPO)算法进行姿态控制,结合在轨操作中机械臂运动学约束,设计奖励函数优化飞行器姿态控制精度。为验证上述策略有效性,在MuJoCo仿真环境中进行星体姿态恢复数值仿真,并针对不同星体质量、不同末端负载质量等工况进行算法适应性评估,结果表明该强化学习方法能满足飞行器进行有限姿态控制的需求,无需参数调节且具有一定鲁棒性。

关键词: 机器人;强化学习;双臂协同;姿态控制
中图分类号: TP242 **文献标志码:** A **文章编号:** 1005-2615(2025)03-0467-08

Attitude Control for Emergency Recovery Based on Reinforcement Learning Method for Dual-arm Space Robots

LI Feng[1,2], *LI Ning*[1,2], *ZOU Huaiwu*[1,2], *JIN Yongqiang*[1,2], *ZHANG Chongfeng*[1,2]
(1. National Key Laboratory of Aerospace Mechanism, Shanghai 201108, China; 2. Aerospace System Engineering Shanghai, Shanghai 201108, China)

Abstract: Aiming at the traditional attitude control failure in on-orbit service dual-arm space robots under extreme abnormal conditions such as flywheel and engine malfunctions, an emergency attitude control algorithm for dual-arm space robots based on reinforcement learning is proposed. This approach achieves limited attitude recovery of the spacecraft using only the two robotic arms configured on the spacecraft which differs from traditional attitude control algorithms. A physical environment for algorithm training is constructed and a model-free proximal policy optimization (PPO) algorithm is used for attitude control. By incorporating the kinematic constraints of manipulators movements during on-orbit operations, the reward function is designed to optimize the precision of spacecraft attitude control. To validate the effectiveness of the proposed strategy, numerical simulations of the space robot attitude recovery are conducted in the MuJoCo environment. The adaptability of the algorithm is evaluated under various conditions, including various masses of the base, various masses of the end. Results demonstrate that the reinforcement learning method is suitable for spacecraft limited attitude control and show a certain robustness without the need of parameter fine-tuning.

Key words: robot; reinforcement learning; dual-arm collaboration; attitude control

基金项目: 国家自然科学基金委与中国航天科技集团公司联合基金(U21B6002)。
收稿日期: 2025-02-28;**修订日期:** 2025-04-24
通信作者: 靳永强,男,研究员,E-mail:jinyong_qiang@126.com。

引用格式: 黎丰,李宁,邹怀武,等. 基于强化学习的双臂空间机器人应急姿态控制[J]. 南京航空航天大学学报(自然科学版),2025,57(3):467-474. LI Feng, LI Ning, ZOU Huaiwu, et al. Attitude control for emergency recovery based on reinforcement learning method for dual-arm space robots[J]. Journal of Nanjing University of Aeronautics & Astronautics (Natural Science Edition),2025, 57(3):467-474.

近年来在轨服务空间机器人执行任务种类呈现多样化,包括但不限于在轨补加、碎片清除和在轨装配,通常配置多自由度机械臂以完成目标抓取与转位、精细操作等具体动作。在轨服务飞行器作为卫星的服务方,与通常卫星一样,同样存在姿态确定与控制系统(Attitude determination and control system, ADCS),其性能直接影响在轨服务质量,是决定任务成败的关键[1]。ADCS一旦发生故障,会导致卫星高速旋转、遥测条件不稳定等情况发生。在以上异常情况下,首要目标是调整飞行器姿态,以保障太阳翼对日,从而恢复整星供电,建立星体-地表遥测通道,帮助地面人员获得卫星异常信息,便于处理异常问题[2]。空间机器人系统具有强非线性、多自由度耦合特性,其机械臂运动引起的扰动较大程度地影响了飞行器姿态,对飞行器姿态控制提出了较高的要求。

目前空间飞行器的姿态控制方法仍然是基于经典的比例-积分-微分(Proportional-integral-derivative, PID)调节方法,其控制受制于具体的航天器动力学参数,因此必须根据地面试验经验和在轨质量变化情况合理设置控制参数,以达到较好的控制效果。针对在轨工况存在的时变扰动,如质心漂移、发动机推力偏移等,传统方法在鲁棒性方面欠佳,为此邓博炜等[3]提出了一种基于干扰力矩补偿的姿态控制方法,通过扩展卡尔曼滤波对干扰力矩在线估计、补偿,以提高控制精度。陆晴等[4]针对姿轨协同控制中出现的随机扰动,引入二阶扩展状态观测器进行前馈补偿,结合比例-微分(Proportional-derivative, PD)反馈控制方法实现姿轨稳定控制。文献[5]提出鲁棒自适应神经网络来增强滑模控制效果,能在消耗更少能量的情况下实现高精度姿态跟踪。

以上基于现代控制理论的方法,往往需要获得具体的动力学模型,需要在原有的动力学模型参数基础上,对控制参数进行设计并对系统扰动、不确定性进行估计。近年来,基于深度神经网络的强化学习在高自由度机器人控制中表现优秀,已成为机器人控制的主流[6]。在飞行器姿态控制中,部分学者已对相关算法进行了相应研究[7-9]。郑鹤鸣等[10]基于Soft演员-评论家模型(Soft actor-critic, SAC)设计了姿态控制器以解决在轨加注过程中液体晃动、转移引起的组合体惯量变化带来的姿态随机扰动问题。Tammam等[11]采用双延迟深度确定性策略梯度算法(Twin delayed deep deterministic policy gradient, TD3)实现了飞行器间的交会和姿态控制,使飞行器能自主调整相对位姿,适应环境干扰和系统噪声。在飞行器发生姿态控制故障时,Peng等[12]提出一种基于强化学习的天线调整策略,控制天线稳定、规律地指向地球,保障遥测通道,使飞行器能在后续任务中重启姿态控制。

空间机器人相比于常规飞行器,其机械臂运动引起基座扰动,相互耦合增加两方的控制难度。传统的基于模型的规划方法如广义雅可比矩阵(Generalized Jacobian matrix, GJM)方法受限于动态奇异性问题和建模误差。面对该问题,地面研究已集中于空间机械臂的智能化解决方案[13]。对于机器人的轨迹规划[14-15]、动态目标跟踪[16]、目标抓捕[17]等在轨操作任务,部分学者采用强化学习来进行运动控制器设计,旨在考虑如何在基座扰动下提高控制性能。岳博晨等[18]基于双向长短期记忆神经网络(Bidirectional long short-term memory, Bi-LSTM)设计了一种自由漂浮空间机械臂控制方法,通过神经网络估计动力学模型的不确定性部分,结合自适应滑模控制器进行机械臂控制,提高了轨迹跟踪精度,但计算的复杂程度高,仅在二自由度机械臂上进行应用。Wang等[19]提出了一种基于近端策略优化(Proximal policy optimization, PPO)算法的多目标轨迹规划策略,通过引入基于泊松分布的动作集成方法(Action ensembles based on poisson distribution, AEP),显著提高了控制策略的精度和稳定性。在此基础上Wang等[20]提出了一种基于层次解耦优化(Hierarchical decoupling optimization, HDO)的无碰撞轨迹规划方法,通过引入双层架构,将任务分解为高级层的避障规划和低级层的位姿跟踪,显著降低了任务优化的复杂性。Cao等[21]提出了一种基于先验策略引导的双臂空间机器人的路径跟踪高效学习算法(Efficient learning-based path tracking, Efficient LPT),通过引入逆运动学作为先验策略,指导智能体的探索方向,提高了算法的收敛效率。

本文主要针对传统姿态控制失效下空间机器人姿态恢复问题,通过研究空间机器人独有的机械臂系统,实现飞行器有限的姿态控制。本文针对姿态控制问题设计相应的动作空间、状态空间和奖励函数,基于PPO算法搭建空间机器人强化学习训练环境。通过引入历史状态、随机化动力学参数,并改进PPO算法,提升控制算法的跟踪精度和鲁棒性。对初始姿态翻滚的飞行器进行控制算法验证,并设计了鲁棒性效果验证实验,结果表明相同网络控制参数能够一定程度上适应基座质量变化、末端负载质量变化等工况。

1　在轨空间机器人应急姿态控制场景

本文采用的双臂空间机器人如图1所示。由于帆板、天线等星上载荷对飞行器姿态影响小，因此不考虑星上载荷建模，将飞行器本体简化为蓝色正方体。飞行器所搭载的两条机械臂均为7自由度的库卡机械臂 iiwa-14，其默认动力学参数、关节运动范围、关节力矩范围等参数均按实际机械臂参数设置。两条机械臂均安装于黑色基座的几何中心，基座与星体安装面的几何中心在各个方向均存在偏移，两侧基座刚性连接于飞行器$\pm y$侧，且相对于正方体中心呈中心对称。在该场景中，空间机器人通过机械臂抓取本体上的小型飞行器并保持末端连接状态。飞行器质量为M，记其惯性系下姿态的欧拉角对应为$\boldsymbol{\Psi}$，三轴角速度为$\boldsymbol{\Omega}$，机械臂关节角为$\boldsymbol{q}$，关节角速度为$\dot{\boldsymbol{q}}$，关节力矩为$\boldsymbol{\tau}$。各臂末端负载质量为m。设置重力加速度为0，以等效在轨工况。

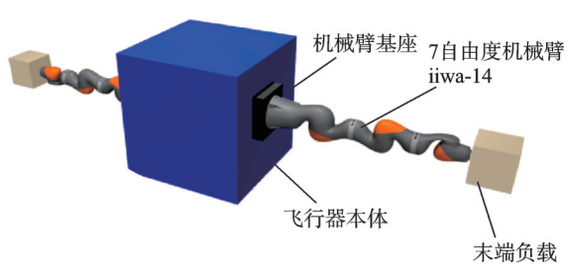

图1　双臂空间机器人模型

Fig.1　Dual-arm space robot model

该任务中智能体可视为执行动作的双臂机器人，其从环境中获得状态信息，并根据深度强化学习策略对环境进行自主探索。智能体在与环境不断交互探索中训练策略，多步训练后智能体通过训练结束的策略结合实时环境状态决策出最优动作。该任务中机器人初始姿态设置为随机值，训练智能体控制机械臂的各关节角速度，使左右机械臂在运动中与飞行器本体产生角动量交换，左右机械臂往复运动直至调节飞行器本体三轴姿态角至安全姿态。为增加角动量交换数值，提升姿态调整效率，机械臂保持末端抓持一定负载的状态进行姿态调整。

2　基于PPO的双臂空间机器人应急姿态控制算法

2.1　PPO算法原理

在强化学习中，智能体与环境交互可视为马尔可夫决策过程（Markov decision process，MDP），通常马尔可夫决策过程的数学模型$\langle S, A, P, R, \gamma\rangle$由以下各部分组成，包括环境状态$\boldsymbol{s}_t\in S$，智能体动作$\boldsymbol{a}_t\in A$，奖励函数$r_t\in R(\boldsymbol{s}_t,\boldsymbol{a}_t)$和状态转移方程$\boldsymbol{s}_{t+1}\sim P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$。下标$t$表示当前时间。$\gamma$表示折扣率，反映历史奖励的损失情况。智能体的策略函数为π，智能体通过环境交互从π中采样动作$\boldsymbol{a}_t$施加于环境，环境反馈的状态信息$\boldsymbol{s}_{t+1}$输入奖励函数R得到奖励值，输入智能体通过最大化该策略得到的期望回报，来不断更新状态价值函数$V^{\pi}(\boldsymbol{s}_t)$，更新策略函数参数，直至得到最优策略$\pi^*$。在深度强化学习中，采用深度神经网络拟合策略函数和价值函数，采用梯度下降方法最大化奖励函数，以适应复杂的动力学任务。

PPO算法由信任区域策略优化（Trust region policy optimization，TRPO）发展形成，是一种广泛应用于强化学习的on-policy策略优化算法，适用于优化大型非线性网络。相比TRPO，PPO不需要复杂的二次优化，同时样本利用率高，参数敏感度低。其基于Actor-Critic框架，采用Actor网络π_θ拟合策略函数，更新原则按策略梯度

$$\nabla_\theta L(\theta)=E_\theta[\nabla_\theta \ln\pi_\theta(\boldsymbol{s},\boldsymbol{a})A^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})] \tag{1}$$

式中A^{π_θ}为优势函数，表示状态$\boldsymbol{s}$下选择动作$\boldsymbol{a}$的价值与当前状态$\boldsymbol{s}$价值期望的差值。通过该梯度迭代更新参数θ。

采用Critic网络V_φ拟合价值函数，采用时序差分更新参数，更新梯度为

$$\nabla_\varphi L(\varphi)=-(r+\gamma V_\varphi(\boldsymbol{s}_{t+1})-V_\varphi(\boldsymbol{s}_t))\nabla_\varphi V_\varphi(\boldsymbol{s}_t) \tag{2}$$

本文采用改进的PPO-clip算法，将策略网络损失函数更改为

$$L(\theta)=-E_{\theta_{\text{old}}}\left[\min\left(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}A^{\pi_{\theta_{\text{old}}}},\text{clip}\left(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}},1-\epsilon,1+\epsilon\right)A^{\pi_{\theta_{\text{old}}}}\right)\right] \tag{3}$$

clip函数定义为$\text{clip}(x,y,z)=\max(\min(x,z),y)$，通过该函数限制策略函数的数值范围，保证新旧参数间变化有限，有利于网络进行参数优化。

2.2　状态空间与动作空间

在该任务中，状态空间引入上下文的状态，即状态空间设置包含前两步历史状态，如式(4)所示。

$$\boldsymbol{s}_t=[\boldsymbol{c}_{s_{t-2}},\boldsymbol{c}_{s_{t-1}},\boldsymbol{c}_{s_t}] \tag{4}$$

$$\boldsymbol{c}_{s_t}=[\boldsymbol{q}_t,\dot{\boldsymbol{q}}_t,\boldsymbol{\Psi}_t,\boldsymbol{\Psi}_{\text{d}},\boldsymbol{\Omega}_t,\boldsymbol{\Omega}_{\text{d}}] \tag{5}$$

式中下标d表示期望值，当前步输出状态为关节角度、关节角速度、基座姿态角、期望基座姿态角、基座角速度和期望基座角速度，维度共为40维，在增加前两步历史状态后，进入框架进行计算的状态维度为120维。为提高算法的收敛性，本文针对不同物理含义的状态量，根据其值变化范围，进行适当

裁剪,必要时进行归一化处理。

双臂空间机器人通过关节轨迹规划实现飞行器姿态恢复,具体由两条7自由度机械臂运动实现,由于角动量交换主要反映于机械臂在关节空间中的加减速,这里关节采用速度控制模式,因此动作空间为14维度的关节期望角速度。该关节期望角速度通过机械臂关节驱动层的PD控制进行跟踪,同时期望角速度根据机械臂运动能力进行裁剪。

2.3 奖励函数

作为强化学习任务的优化目标,奖励函数直接决定了智能体的决策倾向。本文优化目标是飞行器姿态恢复,因此奖励以当前姿态角与期望姿态角距离最小为目标。此外增加了额外奖励实现关节运动轨迹平滑、输出能量优化和避免碰撞干涉等。所有奖励值为负值,以实现姿态角更快收敛。以上4种奖励分别如下:

姿态期望奖励为

$$r_1=-\|\boldsymbol{\Psi}-\boldsymbol{\Psi}_{\mathrm{d}}\|^2-\ln(\|\boldsymbol{\Psi}-\boldsymbol{\Psi}_{\mathrm{d}}\|^2+\epsilon) \tag{6}$$

式中ϵ通常取较小值,为防止姿态收敛时$\|\boldsymbol{\Psi}-\boldsymbol{\Psi}_{\mathrm{d}}\|^2$数值太小引起该奖励函数求解过大。

运动平滑奖励为

$$r_2=-\|\dot{\boldsymbol{q}}-\dot{\boldsymbol{q}}_{\mathrm{d}}\|^2 \tag{7}$$

输出能量惩罚为

$$r_3=-\|\dot{\boldsymbol{q}}\|^2 \tag{8}$$

干涉惩罚为

$$r_4=\begin{cases}-1 & \text{碰撞发生}\\ 0 & \text{其他}\end{cases} \tag{9}$$

单步的总奖励值为$r=\sum_{i=1}^{4}k_i r_i$。类比于优化函数中的各惩罚项系数,通过调整各奖励函数前的权重参数k_i可实现智能体不同的控制效果。

2.4 网络框架与训练流程

当前任务通过智能体规划机械臂具体运动使飞行器姿态到达理想角度,因此该问题实质为轨迹规划问题,在强化学习框架下可等效为最大化奖励函数问题。在该运动规划中,需要满足机械臂运动学约束,包括关节角度限制、关节角速度限制、关节力矩限制和规划轨迹无碰撞,需要在物理引擎中满足动力学,具体表达关系为

$$\max J(\pi_\theta)=E\left[\sum_{t=0}^{\infty}\gamma^t r(t)\right] \tag{10}$$

$$\text{s.t.}\begin{cases} r=\mathrm{Reward}(\boldsymbol{\Psi},\boldsymbol{\Psi}_{\mathrm{d}},\dot{\boldsymbol{q}},\dot{\boldsymbol{q}}_{\mathrm{d}},\ \mathrm{collision})\\ \boldsymbol{q}_{\min}\leqslant\boldsymbol{q}\leqslant\boldsymbol{q}_{\max}\\ \dot{\boldsymbol{q}}_{\min}\leqslant\dot{\boldsymbol{q}}\leqslant\dot{\boldsymbol{q}}_{\max}\\ \boldsymbol{\tau}_{\min}\leqslant\boldsymbol{\tau}\leqslant\boldsymbol{\tau}_{\max}\\ \text{无碰撞}\end{cases}$$

针对上述优化问题,整个训练框架如图2所示,主要包括智能体、物理环境、奖励函数和关节电机驱动部分。物理环境将需要的状态输入至奖励函数中,奖励函数计算得到相应奖励值。智能体根据状态输入和奖励值,以期望回报最大为目标,优化各网络参数,再根据策略网络当前参数计算对应状态的期望关节角速度,输入PD控制器中计算关节力矩,施加于空间机器人,通过空间机器人正动力学再计算出状态值,循环上述过程直至奖励值平稳收敛,最终得到优化后的策略网络参数用于姿态控制。

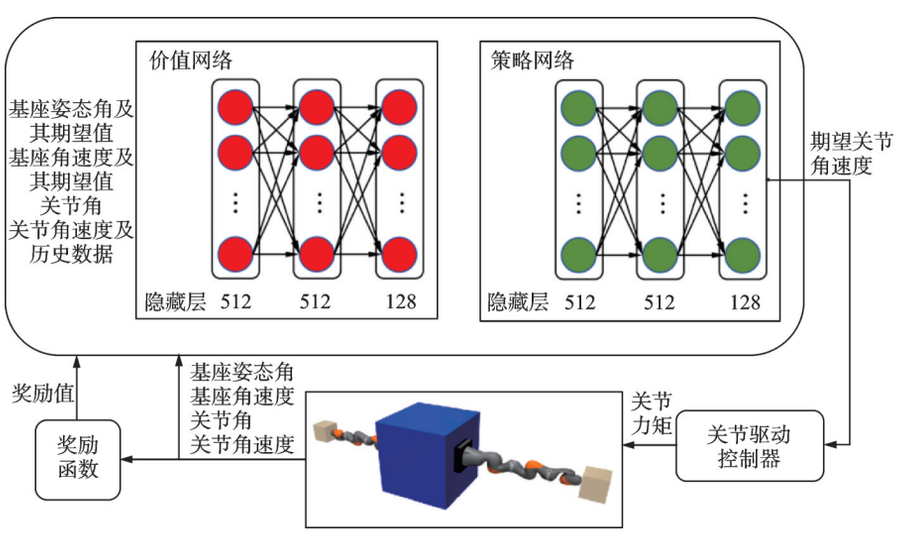

图2 网络训练框架

Fig.2 Network training framework

3 仿真结果与分析

3.1 仿真环境及训练过程说明

本文基于开源动力学物理引擎MuJoCo进行仿真环境搭建,仿真中机械臂控制频率为100 Hz,机械臂模型采用MuJoCo官方提供的xml模型,该描述文件中包括该机械臂真实的物理参数、碰撞几何包络,能充分反映该机械臂的动力学特性。本次

仿真设置4 096个仿真环境进行并行计算。

物理参数及训练超参数设置见表1。策略网络和价值网络均采用三层全连接层，各层间连接激活函数，增加网络非线性。

表1　环境参数设置

Table 1　Parameter settings of the environment

环境参数	数值
飞行器质量M/kg	300
末端负载质量m/kg	30
训练周期/s	20
策略网络结构	512,512,128
价值网络结构	512,512,128
优化器	Adam
学习率	5E−4
PPO 截断率	0.2
折扣率	0.99
训练步数	10 000
并行环境	4 096
初始姿态角范围/(°)	[−30,30]

根据当前飞行器所配置机械臂及末端负载质量，结合训练周期长度，本文设置飞行器初始姿态角度范围为±30°，旨在验证该方法具有应急情况下的有限姿态调节能力。在训练中采用20% 偏差随机初始化飞行器质量。在单次训练开始时，飞行器初始姿态角在±30°范围内随机取值，质量在[240 kg，360 kg]范围内随机取值。当单次训练超过20 s时，环境重置，继续进行下一轮训练。经过10 000轮训练，收敛过程如图3所示，平均奖励值最终趋于稳定。

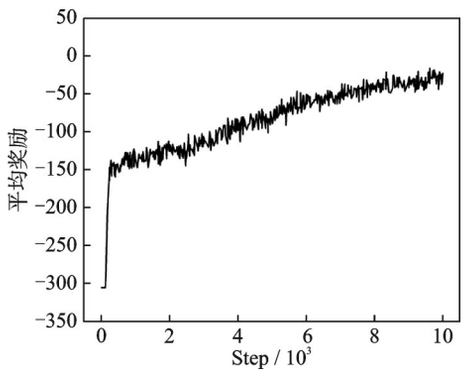

图3　强化学习训练曲线

Fig.3　Mean reward curve during reinforcement training learning

训练后的智能体针对存在初始姿态偏差的飞行器的控制过程如图4所示，给出两种随机初始姿态角下的控制结果。图4中黄色正方体为目标姿态，其三轴姿态角为0°，从左向右依次为姿态校正的过程。空间机器人通过基座两侧机械臂的往复运动，逐渐将蓝色基座与黄色立方体重合，飞行器恢复正常姿态角。

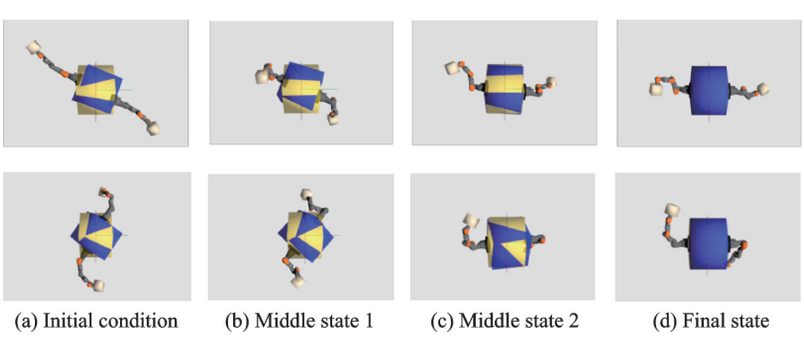

图4　不同初始姿态下飞行器姿态恢复控制过程

Fig.4　Recovery process from various initial conditions of orientation

3.2　仿真结果说明

3.2.1　飞行器应急姿态控制仿真

基于上述训练收敛的策略网络参数对飞行器姿态控制进行仿真实验，在初始姿态范围[−30°，30°]间随机取值5次，统计20 s时间段内三轴姿态变化情况，如图5所示。飞行器各轴姿态角在9 s左右控制至0°附近。

为了进一步验证控制效果，随机初始化100次仿真环境，统计各姿态角在控制开始后第20 s时基座姿态角控制结果，其均值如表2所示，各轴控制误差均小于1°，表明该方法的姿态控制精度可达到1°以内。

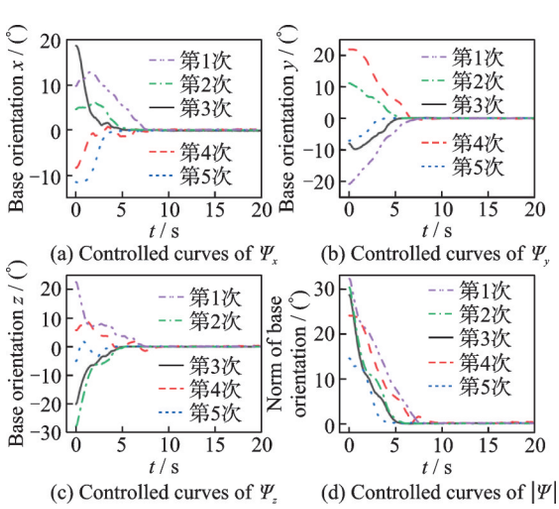

图5　空间机器人姿态恢复效果

Fig.5　Effects of space robot orientation recovery

表2 姿态角控制结果

Table 2 Results of orientation control

各轴姿态角	Ψ_x	Ψ_y	Ψ_z
误差值	0.77	0.54	0.52

3.2.2 鲁棒性测试仿真

为验证控制策略的鲁棒性,从以下3方面进行鲁棒性测试:设置较大初始姿态角;变化基座质量;变化末端负载质量。针对部分工况采用成功率判断控制性能,仿真场景针对100个空间机器人,初始基座角度在[-30°,-10°]∪[10°,30°]范围内随机采样,不改变网络参数,在60 s内采用相同控制器控制飞行器姿态。由于飞行器姿态角范围在5°内可视为姿控恢复正常,控制结束后飞行器三轴姿态角均小于5°视为控制成功,将成功机器人数目在总机器人数目中的占比作为该工况的成功率。

(1)实际训练环境中初始姿态角在±30°范围内随机取值,考虑恶劣情况,初始三轴姿态角均设置为+40°,对飞行器进行控制仿真,控制效果如图6所示。由于追求训练速度,训练时各工况的重置时间为20 s。针对+40°初始姿态工况,在20 s时飞行器控制未达到目标角度,但仍有下降趋势,最终在32 s稳定于0°附近。因此该模型在不改变参数的情况下,尽管控制速度缓慢,对初始姿态角大于训练设计范围的飞行器仍具有一定控制效果。

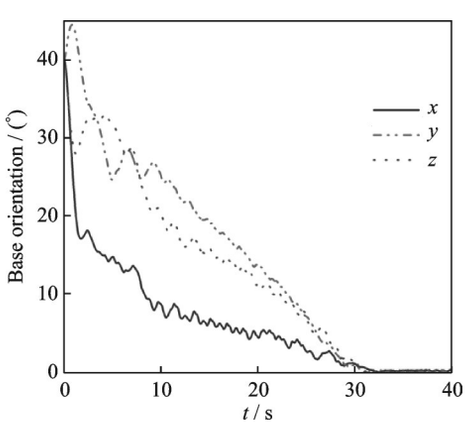

图6 初始姿态角40°偏差下姿态控制曲线

Fig.6 Attitude control curves with 40° initial orientation

(2)在轨执行任务过程中,涉及空间机器人姿轨控时会消耗飞行器燃料,直接影响飞行器质量特性,此外地面测量误差、在轨设备更换等也会导致卫星质量改变。为了验证本文算法在基座质量变化下的有效性,设置相关工况进行仿真分析。本文训练时基座的参考质量设置为300 kg,这里将基座质量在200~3 000 kg范围内变化,计算其姿态控制成功率,结果如图7所示。在300 kg附近成功率最高;当基座质量增加后,成功率降低;基座质量增加至1 200 kg,即训练环境设置值的4倍时,仍然保持90%以上的成功率;增加至2 500 kg后,成功率下降至70%以下。以上结果表明,该机械臂调控星体姿态的方法,对飞行器本体质量敏感程度低。在状态反馈中未引入飞行器质量特性,策略网络主要通过飞行器姿态角、角速度等偏差反馈来指导机械臂运动,使其运动产生的基座干扰力矩能控制基座朝向理想的姿态角运动。

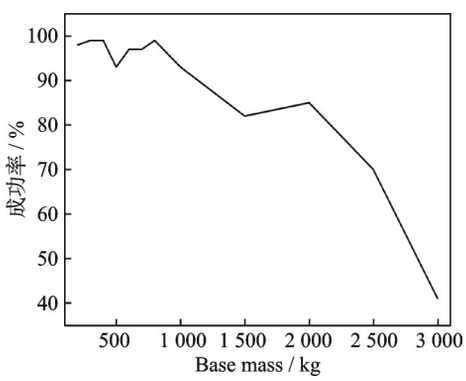

图7 不同基座质量下姿态控制成功率曲线

Fig.7 Curve of success rate with respect to various base masses

(3)机械臂在轨执行任务突发姿态控制故障时,无法保证末端负载与设计值相同。为了验证本文算法在不同末端负载下的有效性,设置相关工况进行仿真分析。设计末端负载质量在20~60 kg范围内变化,分别改变±y侧机械臂末端负载,仿真计算姿态控制成功率,结果如图8所示。在两侧负载均为20 kg时,控制效果较差,原因为当末端负载为20 kg时网络仍按30 kg负载所训练的策略控制机械臂运动,20 kg负载下角动量交换效果差,飞行器姿态改变小,直接导致成功率降低。两侧负载均在30 kg及其以上时,姿态恢复成功率在76%以上,整体变化趋势不明显。两侧负载质量不同时,姿态恢复成功率有下降趋势,尤其在单边负载为20 kg情况下,成功率明显下降。以上结果表明,末端负载对控制算法的有效性存在影响,末端负载越小,机械臂运动对基座扰动越不明显,无法有效控制飞行器恢复正常姿态。为解决该问题,可增加在轨机械臂末端负载质量辨识方法,通过训练多个不同末端负载对应的控制策略网络来应对负载变化。实际使用时先采用末端负载辨识,再选择对应负载质量的策略网络进行控制以达到最佳效果。

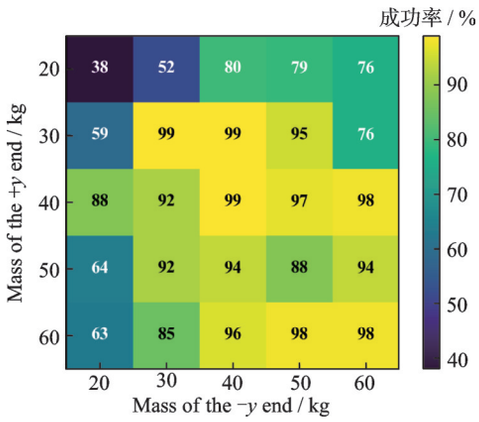

图 8　末端负载质量变化下的控制效果

Fig.8　Results of orientation control under various masses of end

4　结　　论

本文设计了基于强化学习的双臂空间机器人应急姿态控制算法，主要解决飞行器传统姿态控制失效时的姿态恢复问题，通过控制机械臂轨迹规划使其与飞行器产生角动量交换，实现飞行器有限的姿态恢复。本文基于 MuJoCo 物理引擎建立动力学模型，采用 PPO 算法搭建强化学习框架，对该任务进行了训练与仿真实验。根据任务特点设计动作空间、状态空间及奖励函数。训练完毕后采用多种工况分别验证了该算法的姿态恢复精度、控制鲁棒性。结果显示本文提出的控制策略，可实现一定程度的飞行器姿态恢复，且对于控制对象的动力学参数变化不敏感。未来可扩展该控制方案功能，使其能控制飞行器到达任意给定姿态角；针对末端负载较小情况，研究机械臂运动轨迹优化方法以提升姿态控制能力。

参考文献：

[1] OROZCO M L, GIRALDO B S. Attitude determination and control in small satellites: A review[J]. IEEE Journal on Miniaturization for Air and Space Systems, 2024, 5(3): 182-186.

[2] WANG Zhenhua, YANG Sen, ZHOU Zeya, et al. Review on attitude determination and control methods for satellite emergency recovery[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2024, 41(6): 675-688.

[3] 邓博炜，田源，王悦，等. 基于干扰力矩补偿的空间飞行器姿态控制方法[J]. 导弹与航天运载技术，2022(3)：76-81.
DENG Bowei, TIAN Yuan, WANG Yue, et al. Attitude control method of spacecraft based on compensation of disturbance torques[J]. Missiles and Space Vehicles, 2022(3): 76-81.

[4] 陆晴，陈筠力，张德新，等. 随机扰动下的航天器姿轨保持自抗扰控制[J]. 上海航天(中英文)，2023，40(4)：136-145.
LU Qing, CHEN Junli, ZHANG Dexin, et al. Active disturbance rejection control for spacecraft trajectory and attitude keeping under stochastic perturbation[J]. Aerospace Shanghai (Chinese & English), 2023, 40(4): 136-145.

[5] 李成洋，王 伟，耿宝魁，等. 考虑输入饱和的空间飞行器姿态神经鲁棒自适应滑模控制[J]. 宇航学报，2024，45(8)：1269-1280.
LI Chengyang, WANG Wei, GENG Baokui, et al. Neural robust adaptive sliding mode method for spacecraft attitude control with input saturation[J]. Journal of Astronautics, 2024, 45(8): 1269-1280.

[6] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. (2017-08-01). https://doi.org/10.48550/arXiv. 1707.0634728. 06347.

[7] 肖冰，张海朝. 航天器姿态稳定强化学习鲁棒最优控制方法[J]. 航空学报，2024，45(1)：628890.
XIAO Bing, ZHANG Haichao. Reinforcement learning robust optimal control for spacecraft attitude stabilization[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(1): 628890.

[8] 张瑞卿，钟睿，徐毅. 基于强化学习的航天器姿态控制器设计[J]. 上海航天(中英文)，2023，40(1)：80-85.
ZHANG Ruiqing, ZHONG Rui, XU Yi. Satellite attitude control based on reinforcement learning method [J]. Aerospace Shanghai (Chinese & English), 2023, 40(1): 80-85.

[9] 赵毓，郭继峰，颜鹏，等. 稀疏奖励下多航天器规避决策自学习仿真[J]. 系统仿真学报，2021，33(8)：1766-1774.
ZHAO Yu, GUO Jifeng, YAN Peng, et al. Self-learning-based multiple spacecraft evasion decision making simulation under sparse reward condition [J]. Journal of System Simulation, 2021, 33(8): 1766-1774.

[10] 郑鹤鸣，翟光，孙一勇. 面向在轨加注的组合体姿态 SAC 智能控制[J]. 宇航学报，2023，44(7)：1020-1033.
ZHENG Heming, ZHAI Guang, SUN Yiyong. SAC intelligent attitude control method for on-orbit refueling combination[J]. Journal of Astronautics, 2023, 44(7): 1020-1033.

[11] TAMMAM A, ZENATI A, AOUF N. Deep reinforcement learning for rendezvous and attitude control of CubeSat class satellite[C]//Proceedings of the 7th International Conference on Automation, Control and Robots (ICACR). Kuala Lumpur, Malaysia: [s.n.], 2023.

[12] PENG Hao, BAI Xiaoli. Reorient satellite antenna us-

ing reinforcement learning under unknown attitude failures[C]//Proceedings of AIAA SciTech Forum and Exposition. Orlando,United States: AIAA, 2023.

[13] 陈钢,高贤渊,赵治恺,等.空间机械臂智能规划与控制技术[J].南京航空航天大学学报,2022,54(1):1-16.
CHEN Gang, GAO Xianyuan, ZHAO Zhikai, et al. Review on intelligent planning and control technology of space manipulator[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2022, 54(1): 1-16.

[14] 张辉,左孝中,张伟,等.空间双臂机器人漂浮基座扰动最小化轨迹规划[J].电光与控制,2025,32(2):24-31.
ZHANG Hui, ZUO Xiaozhong, ZHANG Wei, et al. Trajectory planning for minimizing floating pedestal disturbance of spatial dual-arm robots[J]. Electronics Optics & Control, 2025, 32(2): 24-31.

[15] WANG Shengjie, CAO Yuxue, ZHENG Xiang, et al. An end-to-end trajectory planning strategy for free-floating space robots[C]//Proceedings of the 40th Chinese Control Conference. Shanghai, China: [s.n.], 2021: 26-28.

[16] 刘勇,李祥,蒋沛阳,等.基于DDPG-PID的机器人动态目标跟踪与避障控制研究[J].南京航空航天大学学报,2022,54(1):41-50.
LIU Yong, LI Xiang, JIANG Peiyang, et al. Research on robot dynamic target tracking and obstacle avoidance control based on DDPG-PID[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2022, 54(1): 41-50.

[17] 孙康,王耀兵,杜德嵩,等.基于深度强化学习算法的空间漂浮基机械臂抓捕控制策略[J].载人航天,2020,26(6):751-757.
SUN Kang, WANG Yaobing, DU Desong, et al. Capture control strategy of free-floating space manipulator based on deep reinforcement learning algorithm[J]. Manned Spaceflight, 2020, 26(6): 751-757.

[18] 岳博晨,贾世元,王一帆,等.基于Bi-LSTM网络的自由漂浮空间机械臂控制[J].空间控制技术与应用,2023,49(3):18-27.
YUE Bochen, JIA Shiyuan, WANG Yifan, et al. Free-floating space manipulator control based on Bi-LSTM networks[J]. Aerospace Control and Application, 2023, 49(3): 18-27.

[19] WANG Shengjie, ZHENG Xiang, CAO Yuxue, et al. A multi-target trajectory planning of a 6-DoF free-floating space robot via reinforcement learning [C]//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021.

[20] WANG Shengjie, CAO Yuxue, ZHENG Xiang, et al. Collision-free trajectory planning for a 6-DoF free-floating space robot via hierarchical decoupling optimization[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 4953-4960.

[21] CAO Yuxue, WANG Shengjie, ZHENG Xiang, et al. Reinforcement learning with prior policy guidance for motion planning of dual-arm free-floating space robot [J]. Aerospace Science and Technology, 2023, 136: 108098.

(编辑:陈珺)