深空探测学报(中英文)
Journal of Deep Space Exploration

“嫦娥五号”月面采样机械臂路径规划

胡晓东，张　宽，谢　圆，张　辉，卢　皓，刘传凯，陈　翔，赵焕洲，谢剑锋

Path Planning of Lunar Surface Sampling Manipulator for Chang'E-5 Mission

HU Xiaodong, ZHANG Kuan, XIE Yuan, ZHANG Hui, LU Hao, LIU Chuankai, CHEN Xiang, ZHAO Huanzhou, and XIE Jianfeng

在线阅读 View online: https://doi.org/10.15982/j.issn.2096-9287.2021.20210095

“嫦娥五号”月面采样机械臂路径规划

胡晓东，张 宽，谢 圆，张 辉，卢 皓，刘传凯，陈 翔，赵焕洲，谢剑锋

（北京航天飞行控制中心，北京 100094）

摘 要： 针对“嫦娥五号”月面采样任务中采样机械臂的精准控制问题，提出了一种基于深度强化学习的路径规划方法。通过设计深度强化学习算法的多约束奖赏函数，规划了满足安全性、快速性、可达性3个约束的运动路径，实现了采样机械臂的精准控制。在满足任务安全性的提前下，缩短了天地之间的交互时间，机械臂控制效果平稳。在轨实验结果表明，该方法具有较高的准确性和鲁棒性，可为后续的深空探测在轨遥操作采样任务提供借鉴。

关键词： 月面采样；机械臂；路径规划；深度强化学习

中图分类号： V525 **文献标识码：** A **文章编号：** 2096-9287(2021)06-0564-08

DOI:10.15982/j.issn.2096-9287.2021.20210095

引用格式： 胡晓东，张宽，谢圆，等. “嫦娥五号”月面采样机械臂路径规划[J]. 深空探测学报（中英文），2021，8（6）：564-571.

Reference format: HU X D，ZHANG K，XIE Y，et al. Path planning of lunar surface sampling manipulator for Chang'E-5 mission[J]. Journal of Deep Space Exploration，2021，8（6）：564-571.

引 言

2021年12月1日23时11分，“嫦娥五号”着陆器和上升器组合体安全着陆于月面正面风暴洋东北部西经51.8°、北纬43.1°位置，月面初始化后约19 h完成了月面采样封装任务，通过钻取和表层采样两种方式共获取1 731 g月球样品[1]。

空间运动机械臂采取“大范围+精调”运动相结合的控制方式。大范围运动过程需综合考虑安全性、可达性、平稳性，从而实现空间机械臂的远距离移动；精调运动过程需要地面上注规划运动策略，机械臂微调到达精确的目标位置。目前，精调控制采用“视觉定位+专家决策”的方式实现，视觉定位计算出机械臂的当前位置以及与目标位置的偏差，专家结合以往控制经验对机械臂的运动路径进行现场决策，该方式主要面临的难题：①受细长机械臂柔性形变和关节间隙误差影响，机械臂会产生较大的控制偏差，为满足月面采样控制过程的毫米级精度要求[2-3]，需多次精调到达目标位置；②天地协同工作频繁、过程复杂，机械臂大范围运动到位后，地面首先进行视觉定位，专家基于定位结果给出精调方向和移动距离，难以适应未来“大时延”的深空采样任务；③因不确定的精调量，采样任务需提前准备机械臂运动参考坐标系下各轴向不同步长的运动指令，任务中计算出调整控制量后采用精调指令组合方式实施控制，但运动路径并不是最优的路径。

人工智能研究的对象侧重于相对复杂的控制环境，以及控制模型不确定性等情况[4-6]。强化学习作为人工智能中机器学习的重要组成部分，机器人通过传感器完成与环境之间的交互，以最大化奖励作为优化目标，通过策略函数获取最优策略，既可以解决合作目标的路径规划问题，又可以实现非合作目标的路径规划。

“嫦娥五号”月表取样采样过程中，有触月、采样、放样、抓罐和放罐5个步骤，使用四自由度细长柔性机械臂采集月面土壤样本并将其转移至初级密封罐中。机械臂的大范围运动至精调初始位置点后，采用精调方式控制机械臂到达目标位置。因细长机械臂的柔性形变特性和机械臂关节间存在间隙误差，使得机械臂末端实际抵达位置与预期位置存在一定的偏差[2]。判断机械臂末端采样器到达预定采样位置，是实现机械臂精准控制并保证能够采到样品的关键。

针对机械臂的细长、柔性特点使得难以通过开环控制实现精确采样的问题，本文提出了一种基于深度强化学习的月面采样机械臂路径规划方法，针对采样

收稿日期：2021-09-30 修回日期：2021-11-12

基金项目：国家自然科学基金青年科学基金资助项目（62003025）；国家自然科学基金资助项目（61972020）

机械臂任务环节多、约束条件复杂、工作环境恶劣等难题，构建了基于深度强化学习——深度Q网络（Deep Q Network，DQN）的路径规划方法，最后结合仿真实验和“嫦娥五号”月面在轨放样精调过程对方法的有效性进行了验证说明。

1　深度强化学习方法

1.1　强化学习

强化学习（Reinforcement Learning，RL）作为一种重要的机器学习方法，原理如图1所示[7]。基于马尔科夫决策过程，强化学习不需要对环境提前建模，通过试错机制求解最优策略，适应于求解未知环境下的机械臂路径规划问题。Maeda等[8]将人工神经网络与强化学习结合，在特定任务中实现了机器人的路径规划。Park等[9]将强化学习与传统随机路图（Probabilistic Road Maps，PRM）算法结合，提升了PRM算法处理动态环境的能力，实现了机械臂的实时高效路径规划。Lange等[10]利用卷积神经网络的非线性拟合能力，使其与强化学习算法结合，成功实现了移动机器人的路径规划。

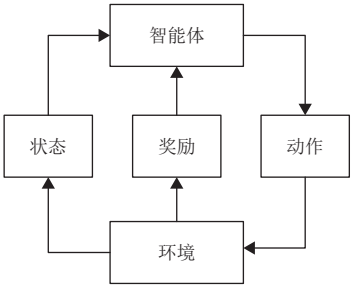


图 1　强化学习的原理

Fig. 1　Principles of reinforcement learning

1.2　深度学习

深度学习（Deep Learning，DL）[11]采用深度神经网络（Deep Neural Network，DNN）[12]作为算法的网络架构，适应于解决机械臂路径规划中的图像处理问题。深度学习模拟人脑的思考和训练方式，将大量的数据特征储存在DNN参数中，通过训练不断改变网络权值参数，其中，应用最广泛的框架是卷积神经网络（Convolutional Neural Network，CNN）[13]。卷积神经网络解决了传统方法网络参数多、训练难度大的问题。典型的卷积神经网络的结构包括相互交替的特征提取层与特征映射层、若干全连接层和分类器等，图2是一个典型的4层卷积神经网络。

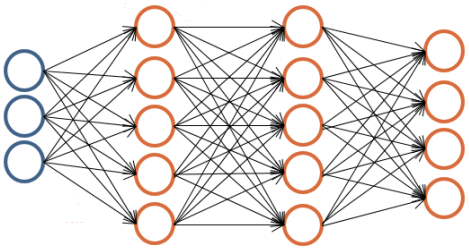


图 2　卷积神经网络结构

Fig. 2　Structure of convolutional neural network

1.3　深度强化学习

深度强化学习算法综合了深度学习的非线性拟合能力和强化学习的自学习特性，综合考虑多因素约束条件，适应于多维度输入的机械臂路径规划问题，提高了控制机械臂的智能性。2013年，谷歌在NIPS会议上发表了DQN算法[14]，揭开了深度强化学习的序幕，DQN利用4幅图像作为网络的输入，采用从图像中直接学习的策略来优化智能体的策略，图3为DQN算法的原理。

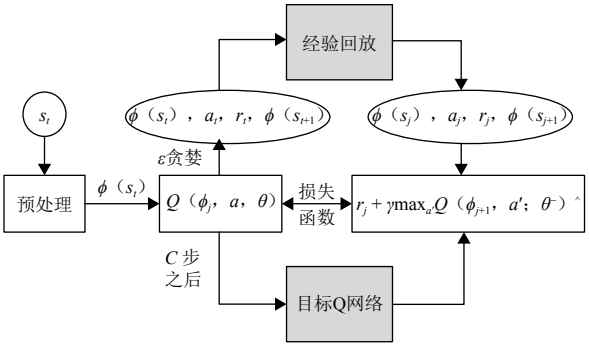


图 3　DQN算法的原理

Fig. 3　Principle of deep Q network algorithm

2015年，DeepMind发表了改进的DQN算法[15]，提出了经验回放（Experience Replay）和目标Q网络（Target Q Network）两大改进技术。Lei等[16]采用改进的DQN算法，实现了移动机器人在未知环境下的路径规划。Zhang等[17]利用DQN算法的自学习能力，解决了机械臂捕获目标的路径规划问题，并将该算法应用到真实场景下三关节机械臂的运动规划。为进一步提高DQN算法的准确率，Hasselt等[18]提出了Double Q-Network，通过2个Q网络的交替工作来降低犯错的概率。Schaul等[19]提出了Prioritized Replay机制，基于优先级的回放机制变相增加了训练的样本，减少了样本之间的干扰。Wang等[20]提出了Dueling Network，在网络内部将Q值分解为两部分：一部分是与动作无关的标量；另一部分是与动作有关的Advantage值。上述算法解决了离散空间的控制问题，但很难适应于连续高

维的控制难题。

2 深度强化学习的机械臂路径规划方法

2.1 月面采样机械臂模型

为解决采样任务环节多、约束条件多、工作环境恶劣等难题[21]，“嫦娥五号”采样机械臂采用“肩2+肘1+腕1”的4自由度构型设计方案[2]，机械臂连杆坐标系在展开状态下的定义如图4所示，D-H参数如表1所示。

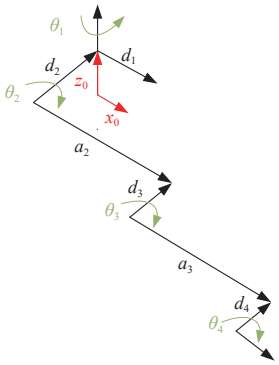


图 4 机械臂连杆坐标系定义

Fig. 4 Definition of manipulator in linkage coordinate system

表 1 机械臂D-H参数

Table 1 The D-H parameters of the manipulator

i	θ/（°）	α_{i-1}/（°）	a_{i-1}/mm	d/mm
1	θ_1	90	0	101.0
2	θ_2	0	0	85.5
3	θ_3	0	1 970	96.0
4	θ_4	0	1 770	93.0

月面采样机械臂正向运动学描述的是机械臂关节空间到末端笛卡尔空间的映射关系，基于代数法的连杆齐次变换矩阵为

$$ {}_{i}^{i+1}\boldsymbol{T}=\begin{bmatrix} c\theta_i & -\sin\theta_i & 0 & a_{i-1} \\ \sin\theta_i\cos\alpha_{i-1} & \cos\theta_i c\alpha_{i-1} & -\sin\alpha_{i-1} & -\sin\alpha_{i-1}d_i \\ \sin\theta_i\sin\alpha_{i-1} & \cos\theta_i s\alpha_{i-1} & -\cos\alpha_{i-1} & \cos\alpha_{i-1}d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1} $$

将表1中各行D-H参数代入齐次变换矩阵可得：

$$ {}_{0}^{1}\boldsymbol{T}=\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 & 0 \\ 0 & 0 & -1 & -d_1 \\ \sin\theta_1 & \cos\theta_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2} $$

$$ {}_{1}^{2}\boldsymbol{T}=\begin{bmatrix} \cos\theta_2 & -\sin\theta_2 & 0 & 0 \\ \sin\theta_2 & \cos\theta_2 & 0 & 0 \\ 0 & 0 & -1 & d_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3} $$

$$ {}_{2}^{3}\boldsymbol{T}=\begin{bmatrix} \cos\theta_3 & -\sin\theta_3 & 0 & a_2 \\ \sin\theta_3 & \cos\theta_3 & 0 & 0 \\ 0 & 0 & -1 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4} $$

$$ {}_{3}^{4}\boldsymbol{T}=\begin{bmatrix} \cos\theta_4 & -\sin\theta_4 & 0 & a_3 \\ \sin\theta_4 & \cos\theta_4 & 0 & 0 \\ 0 & 0 & -1 & d_4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5} $$

2.2 多约束奖励函数

奖励函数作为深度强化学习算法的评价准则，决定了采样机械臂运动策略函数的更新方向，本文选取安全性、快速性、可达性作为奖励函数的约束条件，设计多约束奖励函数。以奖励函数为评价准则，空间机械臂通过与环境的交互调整网络的权值，使目标函数达到最优。多约束奖励函数流程如图5所示。

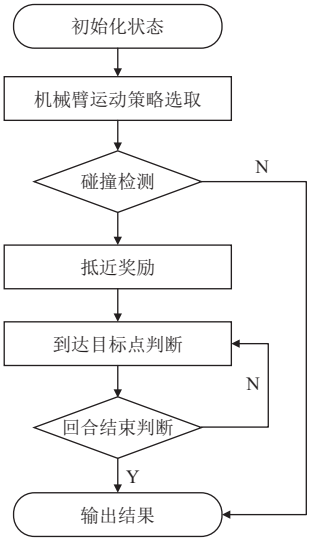


图 5 多约束奖励函数流程

Fig. 5 Process of multi constraint reward function

输入：当前状态S_t，终点坐标C_g，安全性计算方法f_1（θ）、时效性计算方法f_2（θ）、可达性计算方法f_3（θ）；

输出：当前状态对应的奖励R_t，回合结束符$Done$（$Done = True$表示回合结束，$Done = False$表示回合尚未结束）；

Step1（初始化状态）：根据当前状态S_t读取采样机械臂各关节角度和末端位姿；

Step2（机械臂运动策略选取）：机械臂从策略库中随机选取运动方向和移动距离，机械臂由当前位置p_t移动至下一位置p_{t+1}；

Step3（碰撞检测）：计算采样机械臂各连杆及末端与障碍物的距离，根据f_1（θ）的计算结果，若发生碰撞，给予反向奖励−100，即$R_t = R_t - 100$；

Step4（抵近奖励）：根据f_2（θ）的计算结果，

若机械臂最新位置p_{t+1}比上一位置p_t更接近目标点，则更新奖励$R_t = R_t + 0.1\times\| p_{t+1} - p_t \|$；

Step5（到达目标点判断）：若机械臂到达目标点，则更新到达目标点的判断符，即$On_goal = On_goal+1$；

Step6（回合结束判断）：为了消除到达目标点的抖动问题，当采样机械臂到达目标点并持续一定时间，即$On_goal > 50$时，判定为回合结束，令$Done = True$，并给予正向奖励+100，即$R_t = R_t + 100$；

Step7（输出结果）：输出当前状态对应的奖励R_t和回合结束符$Done$。

2.3　基于DQN的机械臂路径规划方法

本文基于建立的采样机械臂的运动学模型，将DQN应用于机械臂路径规划，采用奖励函数确定机械臂的策略更新方向，通过训练过程更新网络参数权重，机械臂路径规划算法如图6所示。

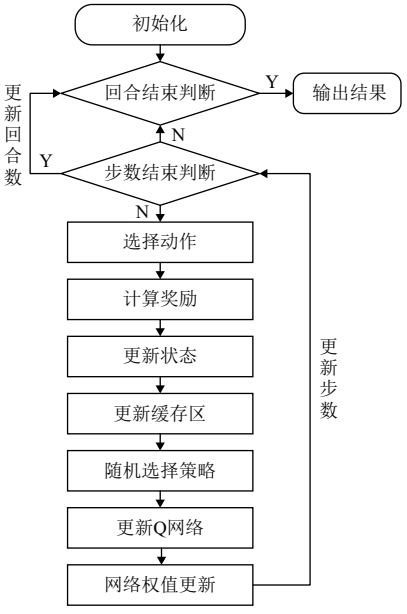


图 6　基于DQN的机械臂路径规划算法

Fig. 6　Path planning algorithm of manipulator based on DQN

Step1（初始化）：随机初始化深度神经网络Q的参数权重ω，给超参数α、λ赋值，清空缓存区M；

Step2（回合结束判断）：判断当前回合标志符是否为最大回合数？若是，则结束训练；若否，则继续Step3～Step10；

Step3（步数结束判断）：判断当前步数标志符是否为最大步数？若是，则结束当前回合；若否，则继续执行4～10；

Step4（选择动作）：根据状态选择并执行相应的动作，为了探索潜在的更优策略，选择动作时遵循的策略为贪婪算法[22]；

Step5（计算奖励）：调用多约束算法函数获得相应的奖励R_t；

Step6（更新状态）：更新状态S_t到S_{t+1}；

Step7（更新缓存区）：将前期经验放入缓存区$\boldsymbol{M}$；

Step8（随机选择策略）：从缓存区M中随机选择n组策略；

Step9（更新Q网络）：根据所选策略更新Q网络，$Q(S_t, A_t) = Q(S_t, A_t) + \alpha(\lambda\max_a Q(S_{t+1}, a) - Q(S_t, A_t) + R_{t+1})$；

Step10（网络权值更新）：根据Q网络训练的损失函数反向传播更新网络权重。

3　实验验证与分析

3.1　月面数字仿真实验与分析

为了验证深度强化学习方法的有效性，本文在MATLAB平台下开展了仿真实验。硬件平台的处理器型号是Intel I9-9880H，主频为2.3 GHz，显卡为GTX 1 060（GPU）。

DQN训练过程的回合数为5万个，每个回合的最大训练步数为500步，每个回合训练的终止条件为机械臂到达目标点或达到最大步数。训练过程中每个回合的步数变化曲线如图7所示。对图7结果进行分析可知：①经过深度强化学习训练，机械臂成功到达目标点，移动路径无碰撞；②随着训练次数的增加，机械臂单回合所需步数逐渐减小，表明机器人向最优路径靠拢；③由于机械臂初始位置具有随机性，最终收敛曲线存在一定的波动。

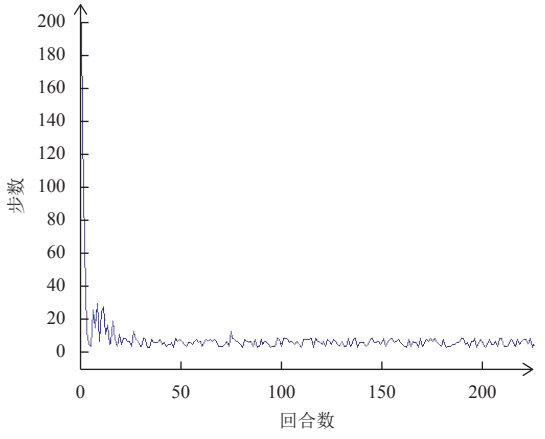


图 7　单回合所需步数

Fig. 7　Steps required for a single episode

为更好展示单回合的收敛情况，本文绘制了训练过程中单回合奖励的变化曲线，选取了第5 000/10 000/30 000/50 000回合作为代表，展示了训练过程的收敛情

况，仿真结果如图8所示。在训练的起始阶段（如第5 000回合），机器人处于初始探索阶段，奖励处于很低的水平；随着训练回合的增加（如第1万回合），奖励开始增大，但波动很大；当训练回合足够大时（如第3万回合和第5万回合），单回合奖励曲线已经收敛，表明机器人找到了最优路径。

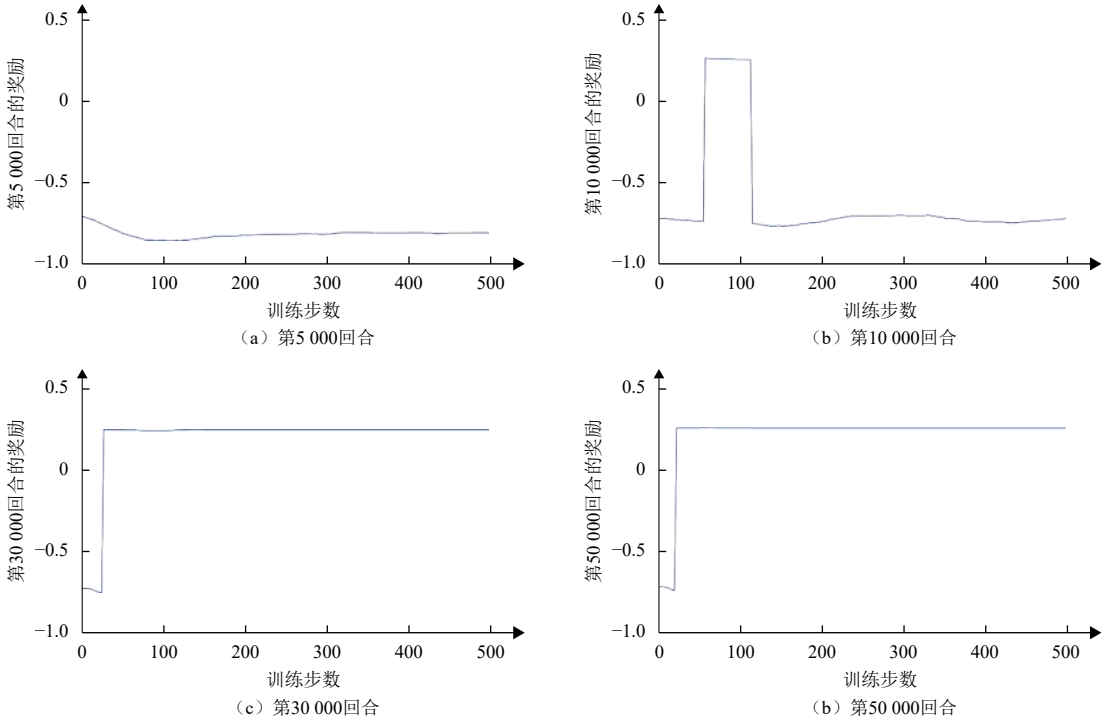


图 8 单回合奖励曲线

Fig. 8 Single-round reward curve

3.2 月面采样精调实验与分析

2020年12月2日，“嫦娥五号”表取采样过程共进行了12次放样，该采取过程采用“视觉定位+专家决策”的方式，视觉定位阶段对下传的近摄像机图像进行椭圆轮廓特征提取，特征自动提取完毕后，采用视觉定位方法计算出机械臂的位姿以及与放样目标点的距离，椭圆轮廓特征提取的图像如图9所示。

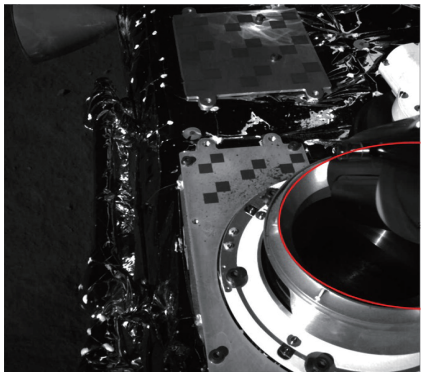


图 9 放样过程椭圆特征提取图像

Fig. 9 Ellipse feature extraction image in lofting process

专家决策阶段由地面专家根据地面验证试验的经验，选取合理可行的精调方向和精调距离，地面通过上注事先生成的单方向运动指令实现机械臂的精准控制，如果单次控制结果不满足放样2 mm的误差精度要求，则重复视觉定位和专家决策过程，“嫦娥五号”任务第一次放样精调过程的移动路径如图10所示。

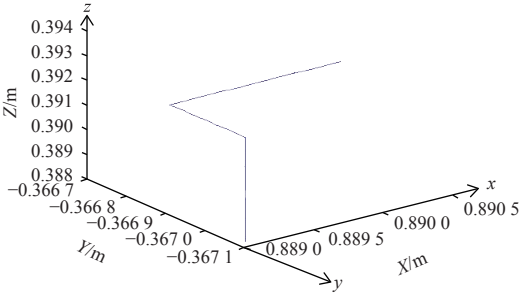
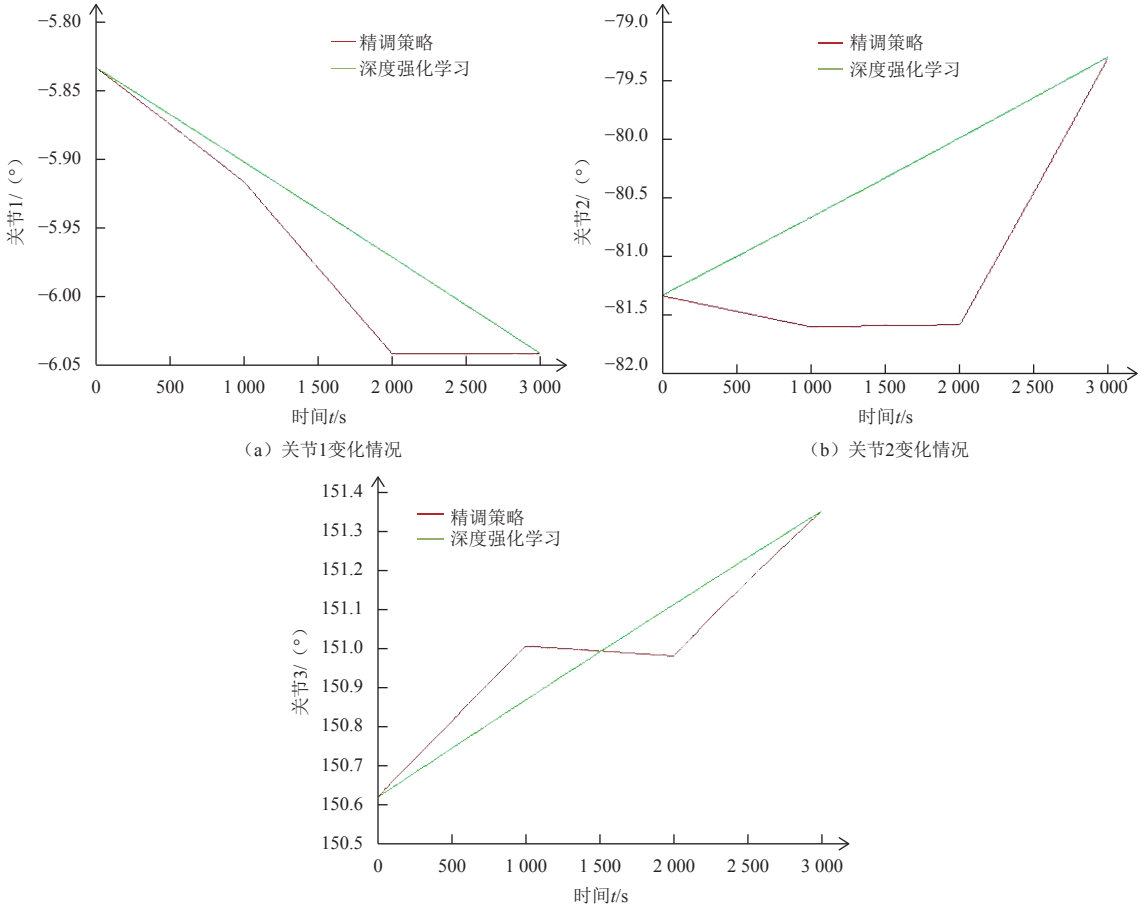


图 10 第一次放样精调过程

Fig. 10 Fine adjustment of the first lofting process

第一次放样精调过程先X、Y方向运动、后Z方向运动的方式实施，精调的过程共耗时20 min，其中机械臂的运动时长3 min，3次视觉定位时长5 min，专家决策及天地交互共耗时12 min。

本文采用“嫦娥五号”任务第一次放样精调的实际数据对训练后的模型进行了验证，实际在轨路径与本算法规划路径下的关节角度变化情况如图11所示。



（a）关节1变化情况　（b）关节2变化情况　（c）关节3变化情况

图 11　机械臂关节变化情况

Fig. 11　Chang’E-5 of joint angle of manipulator

仿真实验结果表明，在安全性方面，在轨路径与规划路径都没有与“嫦娥五号”探测器本体发生干涉，满足任务实施的安全性要求；在运动时间方面，基于深度强化学习的方法共耗时5 min，其中机械臂运动时长2 min，地面复核最终位置时长3 min，相比实际在轨耗时20 min缩短了15 min，极大地缩短了天地之间的交互时间；在机械臂控制效果方面，于在轨实际路径相比，规划路径的关节角度变化更平滑，控制效果更加平稳。

4　结　论

为解决月面无人采样任务中机械臂精准控制难度大、采样时间受限、天地协同复杂的难题，本文提出了基于深度强化学习的机械臂路径规划方法，机械臂根据任务实施过程中的安全性、快速性、可达性准则，自主规划移动路径，并判断控制目标是否满足任务指标要求。通过将该算法规划路径与实际在轨数据路径做比较，结果表明本算法在满足任务安全性的前提下，机械臂控制效果更加平稳，解决了柔性细长采样机械臂难以精准建模导致控制存在系统偏差的难题，实现了采样过程机械臂的精准控制，可以为后续月球及行星探测机械臂远程遥操作采样任务提供借鉴。

参　考　文　献

[1] 王琼，侯军，刘然，等. 我国首次月面采样返回任务综述[J]. 中国航天，2021(3)：34-39.

[2] 马如奇，姜清水，刘宾，等. 月球采样机械臂系统设计及试验验证[J]. 宇航学报，2018，39(12)：5-12.
MA R Q，JIANG Q S，LIU B，et al. Design and verification of a lunar sampling manipulator system[J]. Journal of Astronautics，2018，39(12)：5-12.

[3] 唐玲，梁常春，王耀兵，等. 基于柔性补偿的行星表面采样机械臂控制策略研究[J]. 机械工程学报，2017，53(11)：97-103.
TANG L，LIANG C C，WANG Y B，et al. Research on flexible compensation control strategy for planetary surface sampling manipulator[J]. Journal of Mechanical Engineering，2017，53(11)：97-103.

[4] NAKANISHI H，YOSHIDA K. Impedance control for free-flying space robots -basic equations and applications[C]//International

Conference on Intelligent Robots and Systems. [S. l]：IEEE，2006.

[5] SCHIELE A，HIRZINGER G. A new generation of ergonomic exoskeletons-the high-performance X-Arm-2 for space robotics telepresence[C]//International Conference on Intelligent Robots and Systems. [S. l]：IEEE，2011.

[6] NANOS K，PAPADOPOULOS E. On the use of free-floating space robots in the presence of angular momentum[J]. Intelligent Service Robotics，2011，4(1)：3-15.

[7] SUTTON R S，BARTO A G. Introduction to reinforcement learning[M]. Cambridge：MIT press，1998.

[8] MAEDA Y，WATANABE T，MORIYAMA Y. View-based programming with reinforcement learning for robotic manipulation[C]// IEEE International Symposium on Assembly and Manufacturing. [S. l]：IEEE，2011.

[9] PARK J J，KIM J H，SONG J B. Path planning for a robot manipulator based on probabilistic roadmap and reinforcement learning[J]. International Journal of Control Automation & Systems，2007，5(6)：674-680.

[10] LANGE S，RIEDMILLER M，VOIGTLANDER A. Autonomous reinforcement learning on raw visual input data in a real world application[C]//International Joint Conference on Neural Networks. [S. l]：IEEE，2012.

[11] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436.

[12] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. [S. l]：Curran Associates Incorperation，2012.

[13] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2015，39(6)：1137-1149.

[14] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Playing atari with deep reinforcement learning[J/OL]. (2021-10-9). https://arxiv.org/abs/1312.5602.

[15] OSTAFEW C J，SCHOELLIG A P，BARFOOT T D. Learning-based nonlinear model predictive control to improve vision-based mobile robot path-tracking in challenging outdoor environments[C]//IEEE International Conference on Robotics and Automation. [S. l]：IEEE，2016.

[16] LEI T，MING L. A robot exploration strategy based on Q-learning network[C]//IEEE International Conference on Real-Time Computing and Robotics. [S. l]：IEEE，2016.

[17] ZHANG F Y，LEITNER J，MILFORD M，et al. Towards vision-based deep reinforcement learning for robotic motion control[C]//proceedings of Australasian Conference on Robotics and Automation(ACRA). Australasian：IEEE，2015.

[18] HASSELT H V，GUEZ A，SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence Computer Science. Phoenix，Arizona，USA：AIAA，2016.

[19] SCGAUL T，QUAN J，ANTONOGLOU I，et al. Prioritized Experience Replay[EB/OL]. (2015-11-18). https://www.semanticscholar.org/paper/Prioritized-Experience-Replay-Schaul-Quan/c6170fa90d3b2efede5a2e1660cb23e1c824f2ca?p2df.

[20] WANG Z，SCHAUL T，HESSEL M，et al. Dueling network architectures for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York，USA：JMLR.，2015.

[21] 裴照宇，任俊杰，彭兢，等. “嫦娥五号”任务总体方案权衡设计[J]. 深空探测学报(中英文)，2021，8(3)：215-226.
PEI Z Y，REN J J，PENG J，et al. Overall scheme trade-off design of Chang’E-5 mission[J]. Journal of Deep Space Exploration，2021，8(3)：215-226.

[22] GOMES E R，KOWALCZYK R. Dynamic analysis of multiagent Q-learning with ε-greedy exploration[C]//International Conference on Machine Learning. [S. l]：ACM，2009.

作者简介：

胡晓东(1993–)，男，工程师，主要研究方向：航天器测控总体设计。

通讯地址：北京市海淀区北清路26号院5130信箱(100094)

电话：(010)66363119

E-mail：huxiaodong2037@163.com

谢剑锋(1972–)，男，研究员，主要研究方向：航天测控总体、轨道控制。**本文通讯作者**。

通讯地址：北京5130信箱104号(100094)

电话：(010)66363008

E-mail：jianfengxie@126.com

Path Planning of Lunar Surface Sampling Manipulator for Chang'E-5 Mission

HU Xiaodong，ZHANG Kuan，XIE Yuan，ZHANG Hui，LU Hao，LIU Chuankai，CHEN Xiang，
ZHAO Huanzhou，XIE Jianfeng

（Beijing Aerospace Control Center， Beijing 100094，China）

Abstract：Aiming at the problem of precise control of the sampling manipulator in the lunar surface sampling mission of "Chang'E-5"，a path planning method based on deep reinforcement learning is proposed. By designing the multi-constraint reward function of the deep reinforcement learning algorithm，a motion path that satisfies the three constraints of safety，speed and reachability is planned. The precise control of the sampling robotic arm is realized. Under the advance of meeting the task safety，the interaction time between heaven and earth is greatly shortened，and the control effect of the manipulator is more stable. Experimental results show that this method has high accuracy and robustness，and can provide reference for subsequent on orbit sampling tasks.

Keywords：Lunar surface sampling；manipulator；path planning；deep reinforcement learning.

Highlights：

- A path planning method of lunar surface sampling manipulator based on deep reinforcement learning is proposed.
- The control problem of slender and flexible manipulator is solved.
- The deep reinforcement learning control method has high accuracy and robustness.
- The method improves the efficiency of on orbit mission implementation

[责任编辑：杨晓燕，英文审校：宋利辉]