doi: 10.7690/bgzdh.2024.06.022

基于深度强化学习的机械臂动态目标抓取方法

张　轩[1]，卢惠民[2]，任君凯[2]，莫新民[1]，肖浩然[2]，张伟杰[1]，杨　璇[1]
（1. 西北机电工程研究所人体增强技术创新中心，陕西　咸阳 712099；
2. 国防科技大学智能科学学院，长沙　410073）

摘要：针对现有机械臂动态目标抓取方法轨迹规划困难、实时性不足、难以实现六自由度抓取等问题，提出一种基于深度强化学习（deep reinforcement learning，DRL）的机械臂动态目标抓取方法。进行马尔可夫决策过程（Markov decision process，MDP）建模，设计状态空间、动作空间以及奖励函数，实现机械臂对动态目标的六自由度抓取。基于 Pybullet 构建机械臂动态目标抓取仿真试验环境，对该方法进行训练，将训练得到的策略在新颖场景进行测试，并与经典规划控制的动态目标抓取方法进行对比。仿真结果表明：该方法能实现机械臂对动态目标的六自由度抓取，在抓取成功率和速度上具有优势。

关键词：动态目标抓取；马尔科夫；轨迹规划；深度强化学习；六自由度抓取
中图分类号：TP241　　**文献标志码：**A

A Dynamic Target Grasping Method for Manipulator Based on Deep Reinforcement Learning

Zhang Xuan[1], Lu Huimin[2], Ren Junkai[2], Mo Xinmin[1], Xiao Haoran[2], Zhang Weijie[1], Yang Xuan[1]
(1. *Human Enhancement Technology Innovation Center*, *Northwest Institute of Mechanical & Electrical Engineering*, *Xianyang* 712099, *China*; 2. *College of Intelligence Science and Technology*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: Aiming at the problems of trajectory planning difficulty, insufficient real-time performance and difficulty in realizing six-degree-of-freedom grasping of existing manipulator dynamic target grasping methods, a manipulator dynamic target grasping method based on deep reinforcement learning (DRL) is proposed. The Markov decision process (MDP) is modeled, and the state space, action space and reward function are designed to realize the six-degree-of-freedom grasping of the dynamic target by the manipulator. Based on Pybullet, the dynamic target grasping simulation test environment of manipulator is constructed, and the method is trained. The trained strategy is tested in a novel scene, and compared with the dynamic target grasping method of classical planning control. The simulation results show that the method can realize the six-degree-of-freedom grasping of the dynamic target by the manipulator, and has advantages in grasping success rate and speed.

Keywords: dynamic target grasping; Markov; trajectory planning; deep reinforcement learning; six-degree-of-freedom grasping

0　引言

机械臂动态目标抓取技术在工业生产线上应用广泛，是机械臂领域的研究热点。基于经典规划控制的动态目标抓取方法[1]需要对机械臂进行精确的运动学、动力学建模，建模过程繁琐耗时，结果易受模型误差影响。

近些年，随着机器学习技术的快速发展，基于深度强化学习（DRL）[2]的机器人控制方法引起了研究人员的关注。DRL 方法通过与环境进行大量交互，从中学习抽象特征和复杂的运动规律，使机械臂不依赖精准建模就可实现自主抓取操作。文献[3]首次使用深度强化学习算法实现了对动态目标的抓取，文献[4]采用对抗强化学习的方法，在机械臂学习抓取的同时，让目标物学习如何逃脱；但以上工作只考虑物体的抓取位置而不考虑抓取方向，没有实现更符合人类习惯的六自由度抓取。

基于此，为提升动态目标抓取过程的效率与智能化水平，笔者将深度强化学习算法与动态目标抓取任务相结合，对机械臂动态目标抓取过程中的轨迹规划问题进行马尔可夫决策过程（MDP）[5]建模，设计一种兼顾六自由度抓取位姿控制与安全的奖励函数，有效解决动态目标抓取过程中的机械臂轨迹规划问题。

收稿日期：2024-02-15；**修回日期：**2024-03-27
第一作者：张　轩（1998—），男，陕西人，硕士。

1 方法设计

笔者设计的机械臂动态目标抓取方法主要包含抓取位姿获取、机械臂轨迹规划过程的 MDP 建模和动态目标抓取策略训练 3 部分。整个方法在一个时间步内的流程如图 1 所示，智能体根据当前时间步的状态与奖励输出下一步动作，执行动作后返回新的环境信息作为状态向量并计算奖励奖励，智能体据此输出新的动作。

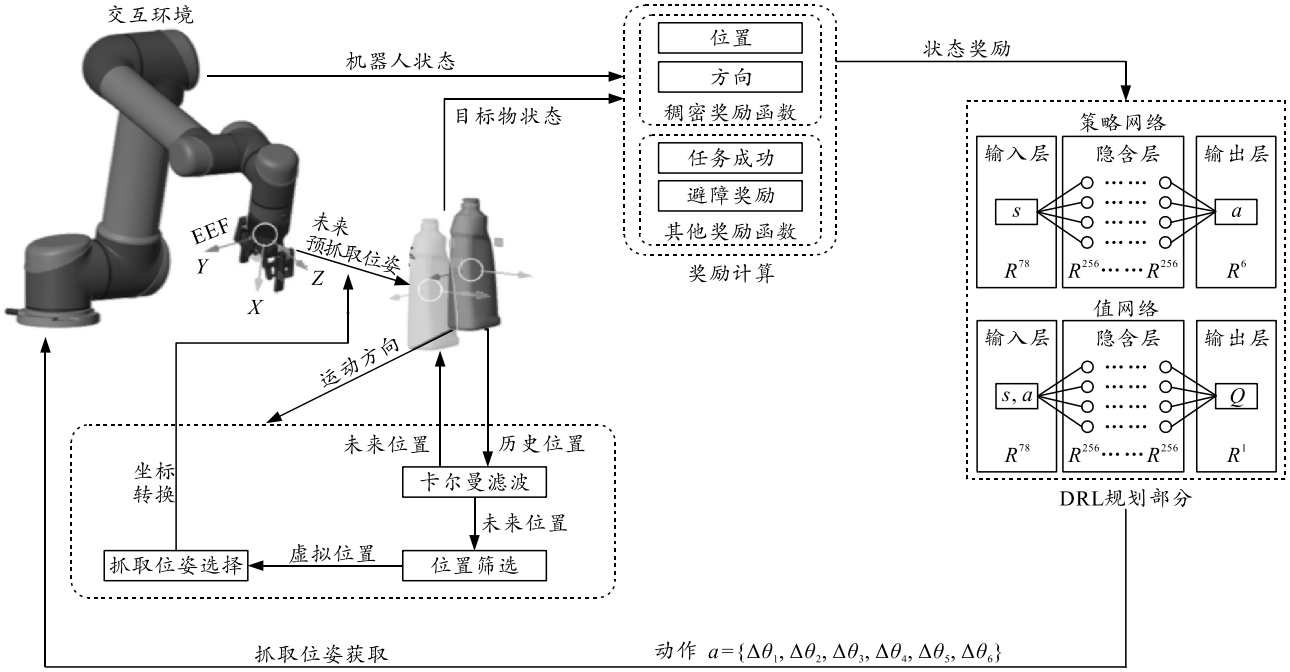


图 1 机械臂动态目标抓取训练流程

1.1 抓取位姿选取

在动态目标抓取任务中，获取一个合适的抓取位姿是整个工作的前提，步骤如图 1 中抓取位姿获取框选所示。

为降低抓取位姿生成需要的时间，笔者提前为待抓取目标生成一个抓取位姿数据库[6]，包含多个可靠的抓取位姿，后续的期望抓取位姿即从数据库中选取产生。

考虑到机械臂运动需要时间，笔者使用卡尔曼滤波为动态目标进行轨迹预测，输出目标 1 和 2 s 后的坐标。再根据机械臂末端执行器的工具中心点(tool center point，TCP)与目标质心的欧氏距离从预测结果中筛选出一个，作为替代目标坐标的虚拟位置。选取规则：欧氏距离大于 20 cm 则选择目标 2 s 后的坐标作为虚拟位置；小于 20 cm 且大于 10 cm 选择目标 1 s 后的坐标作为虚拟位置；小于 10 cm 使用目标坐标真值作为虚拟位置。

为快速在数据库中获取合适的抓取位姿，笔者使用可达性[7]作为筛选标准：假定目标处于虚拟位置，在预先计算的可达性空间中，快速对数据库中的抓取位姿进行可达性评分，选择得分最高的抓取位姿(可达性值越高，机械臂末端处于该抓取位姿下的可操控性也就越高)。

为了保证后续机械臂规划过程的安全性，笔者将选取到的抓取位姿沿其抓取方向的反方向退后 7.5 cm(7.5 cm 为 UR5 机械臂末端执行器的长度)得到预抓取位姿，避免规划过程碰倒目标。

最后经过坐标转换，假定目标处于前文虚拟位置，将预抓取位姿从物体坐标系转换到以机械臂基座为原点的机械臂坐标系下，结果记作未来预抓取位姿(future pre, grasp)pose_{fpg}为机械臂规划过程的目标。这样在整个抓取过程中，机械臂末端靠近的目标，会由预测位置逐渐变为当前位置，避免出现机械臂难以跟上目标的状况。

1.2 抓取过程机械臂规划问题 MDP 建模

为避免机械臂运动学建模，使用深度强化学习方法直接对机械臂的关节空间进行规划，将动态目标抓取过程中的机械臂规划问题建模为：以机械臂和目标物信息为状态空间，以机械臂关节角度增量为动作空间的马尔科夫决策过程，并根据任务的具体要求构造奖励函数。详细设计如下：

1.2.1 动作空间设计

智能体每时间步输出的动作是六自由度 UR5

机械臂各关节的角度增量 $\Delta\theta_i$，具体动作 a 公式如下：

$$a=\{\Delta\theta_1, \Delta\theta_2, \Delta\theta_3, \Delta\theta_4, \Delta\theta_5, \Delta\theta_6\}。\tag{1}$$

同时，$\Delta\theta_i$ 的大小由智能体的控制频率决定，笔者选取 6.67 Hz 作为智能体获取状态，输出动作的频率。考虑到 UR5 机械臂各个关节电机运动的最大角速度为 3.2 rad/s，将每个关节角度增量 $\Delta\theta_i$ 的范围设置为[0°, 18°]。

1.2.2　状态空间设计

要实现机械臂对动态目标的六自由度抓取，相比工作[7-8]限制抓取方向为自上而下。笔者设计状态空间时，除动态目标外，还需考虑抓取和机械臂末端方向。将前文得到的 pose_{fpg} 作为状态空间的一员。

如此机械臂末端执行器需要同时探索不同位置和方向，从而扩大了机械臂的工作空间，增加了机械臂碰撞的风险。为了避免这种情况，笔者将目标位姿 pose_o 和机械臂各连杆在机械臂坐标系下的位姿 l_1～l_7 作为状态空间的一部分，以便智能体更好地理解自身姿态和环境(上述标量包含 7 个值，分别为表位置的 3 维坐标和表旋转的四元数)。

状态空间包括目标位姿、机械臂连杆位姿等信息，通过仿真软件获取真实值，避免了误差对策略训练的影响。

除此以外，还有其他常规状态变量，如 TCP 与 pose_{fpg} 的欧式距离 d_{dist}，夹角弧度值 d_{dir}，以及机械臂各个关节的旋转角度 p_1～p_6 和角速度 w_1～w_6。具体状态 s 公式如下：

$$\begin{aligned}s=\{&d_{\text{dist}}, d_{\text{dir}}, p_1, \cdots, p_6, w_1, \cdots, w_6,\\ &\text{pose}_o, \text{pose}_{\text{fpg}}, l_1, \cdots, l_6, l_7\}。\end{aligned}\tag{2}$$

需要说明的是，这些状态变量都通过仿真软件直接获取真值，而非通过深度相机或机械臂坐标正运动学计算得到，避免观测或运动学建模误差对策略训练产生的影响。

1.2.3　奖励函数设计

笔者在稠密奖励与稀疏奖励的基础上，结合要实现的具体任务即动态目标无碰六自由度抓取，增加避障惩罚和自碰撞惩罚，具体设置如下：

$$R=R_{\text{dense}}+R_{\text{sparse}}+R_{\text{avoid}}。\tag{3}$$

式中：R_{dense} 为稠密奖励；R_{sparse} 为稀疏奖励；R_{avoid} 为避障奖励函数；R 为总奖励。

1) 稠密奖励 R_{dense} 设计：

$$R_{\text{dense}}=W_{\text{dist}}\bullet r_{\text{dist}}- W_{\text{dir}}\bullet r_{\text{dir}}。\tag{4}$$

式中：r_{dist} 为归一化后的位置奖励，每回合由 d_{dist} 除以回合开始时 TCP 与目标质心的欧式距离 d_o 得到；W_{dist} 为位置奖励所占权重；r_{dir} 为归一化后的方向奖励，由 d_{dir} 除以 π 后得到；W_{dir} 为方向奖励所占权重。将位置和方向奖励进行归一化，可以平衡它们在训练中所占比重，避免某一因素主导其他因素。在面对运动轨迹、初始位置或目标物体变化的情况下，都能保持一致的尺度，增强算法的稳定性并加速收敛过程。

r_{dist} 详细计算公式如下：

$$r_{\text{dist}}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}\Big/d_o。\tag{5}$$

式中：(x_1, y_1, z_1)为 TCP 在笛卡尔空间中的位置坐标；(x_2, y_2, z_2)为 pose_{fpg} 在笛卡尔空间中的位置坐标。

r_{dir} 的详细计算公式如下：

$$r_{\text{dir}}=\left|2*\cos^{-1}\left(\left|w_1w_2+x_1x_2+y_1y_2+z_1z_2\right|\right)\right|\Big/\pi。\tag{6}$$

式中：(w_1, x_1, y_1, z_1)为机械臂末端执行器方向的单位四元数；(w_2, x_2, y_2, z_2)为 pose_{fpg} 方向的单位四元数。

2) 稀疏奖励 R_{sparse} 设计：

R_{sparse} 的设置可以进一步明确具体的任务和成功标准。当机械臂末端执行器与 pose_{fpg} 距离误差＜1.5 cm，方向误差在 15°以内，视作该时间步规划成功，给予一个正奖励。连续规划成功 5 次视作任务成功满足抓取条件，给予一个巨大的正奖励。需保证此正奖励大于智能体一回合内所有其他奖励之和，否则难以训练出全局最优解。满足规划成功奖励的任意一项，都会单独给予正奖励，加快学习速度。

具体的离散奖励 R_{sparse} 公式如下：

$$R_{\text{sparse}}=\begin{cases}-0.5, \text{if self collision}\\ 1, \text{if} r_{\text{dist}}\leqslant 1.5\ \text{cm}\\ 1, \text{if} r_{\text{dir}}\leqslant 15°\\ 2, \text{if} r_{\text{dist}}\leqslant 1.5\ \text{cm}\wedge r_{\text{dir}}\leqslant 15°\\ 900, \text{if task successful}\end{cases}。\tag{7}$$

3) 避障奖励 R_{avoid} 设计：

为确保机械臂和目标物体的安全，每当任一连杆与目标之间的最小距离 d_{m_o} 低于阈值时，就会触发避障处罚。

具体来说，笔者在目标对象周围定义了 2 种类型的空间。第 1 种类型是警告区，如果 d_{m_o} 的距离小于 2 cm，则智能体开始收到负奖励，该负奖励的绝对值随着 d_{m_o} 的减少而线性增加，当 d_{m_o} 降至

1 cm 以下时，机械臂进入危险区，避障处罚的绝对值达到最大值。

避障奖励函数 R_{avoid} 具体公式如下：

$$R_{\text{avoid}}=\begin{cases}0, \text{if } d_{m\_o}>0.02\ \text{m}\\ -W_{\text{avoid}}\dfrac{0.02-d_{m\_o}}{0.02}, \text{if } 0.01\ \text{m}<d_{m\_o}\leqslant 0.02\ \text{m}\\ -W_{\text{avoid}}, \text{if } d_{m\_o}\leqslant 0.01\ \text{m}\end{cases}。\quad (8)$$

式中 W_{avoid} 为避障奖励函数所占的权重。

1.3 网络设计与策略训练

传统的近端策略优化算法(proximal policy optimization，PPO)[8]，策略网络与值网络结构简单难以适应高维状态空间路径规划问题。笔者采用 PPO 算法进行策略迭代，为其重新设计的网络结构，具体参数如表 1 所示。

表 1 PPO 算法超参数设置

网络名称	超参数
策略网络	(78, 256, 256, 256, 6)
值网络	(78, 256, 256, 256, 1)
采样批次大小	128
策略网络学习率	$3\times e^{-4}$
值网络学习率	$3\times e^{-4}$

策略网络与值网络的输入为 78 维的状态向量，具体为前文状态空间中的值。笔者使用神经元个数为 256 的 3 层感知机网络提取特征，激活函数为线性整流函数，避免训练中出现梯度消失问题。

策略网络得到当前状态对应的动作概率分布，根据贪心原则选取概率最大的动作，作为输出，具体为 6 维的机械臂电机角度增量值。

值网络得到该状态对应的估计值，使用这个估计值与该状态实际奖励间的差异评价当前策略，更新策略网络与值网络。

重复上述过程，最终使得值网络估计的值函数与真实奖励差异减小，使得策略网络输出的一系列动作能得到更大累计回报，训练出最优策略。

2 机械臂动态目标抓取仿真试验

2.1 实验设置

以 pybullet 作为仿真平台，环境中包含一个配有 robotiq 夹爪的 UR5 机械臂，目标物体和一块可以进行指定运动的传送带，箭头代表 pose_{fpg}。具体如图 2 所示。

选取机械臂正前方，圆心角度数为 40°，小圆半径为 0.70 m，大圆半径为 0.85 m 的扇环区域，在其内随机生成直线轨迹或曲线轨迹，并让传送带以一定速率沿轨迹运动。目标每回合会随机旋转[−45°, 45°]保证实验的随机性(直线运动分为 3～5 cm/s 的加速运动、5～3 cm/s 的减速运动或 5 cm/s 的匀速直线运动；曲线运动轨迹为半径不同的同心圆弧，速率 4 cm/s)，具体示意如图 3 所示。

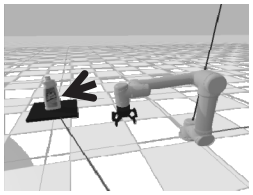


图 2 Pybullet 仿真场景搭建

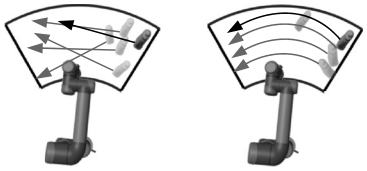


图 3 训练过程中的随机轨迹设置

实验中抓取为固定动作，即沿末端执行器的前进方向伸 7.5 cm 同时闭合夹爪，并举起。

2.2 仿真训练结果与分析

图 4 和 5 展示了抓取随机匀速曲线与随机变速直线运动目标的训练结果。其中横坐标代表训练的时间步，纵坐标表示每 2 048 个时间步的平均回合累计奖励，阴影部分代表多次训练计算得到的奖励标准差。

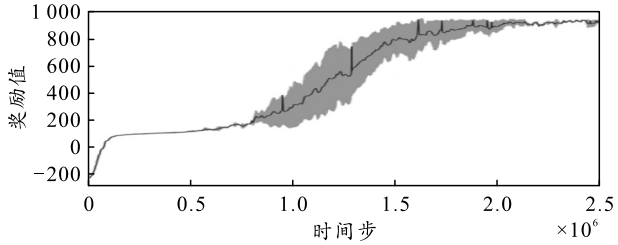


图 4 随机匀速曲线运动目标抓取策略训练曲线

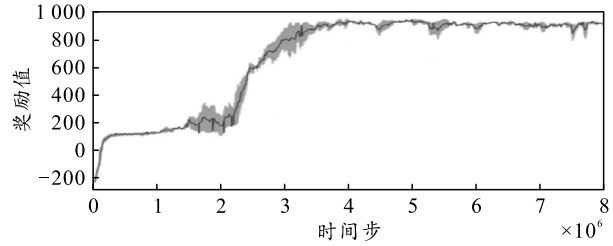


图 5 随机变速直线运动目标抓取策略训练曲线

从图中可以看出：整个训练初期奖励值稳定在 100 左右代表机械臂和很快学习到如何控制末端接近并跟随运动目标；之后经过较长时间步的精细调整逐渐尝试抓取目标。整个训练过程相对稳定，奖

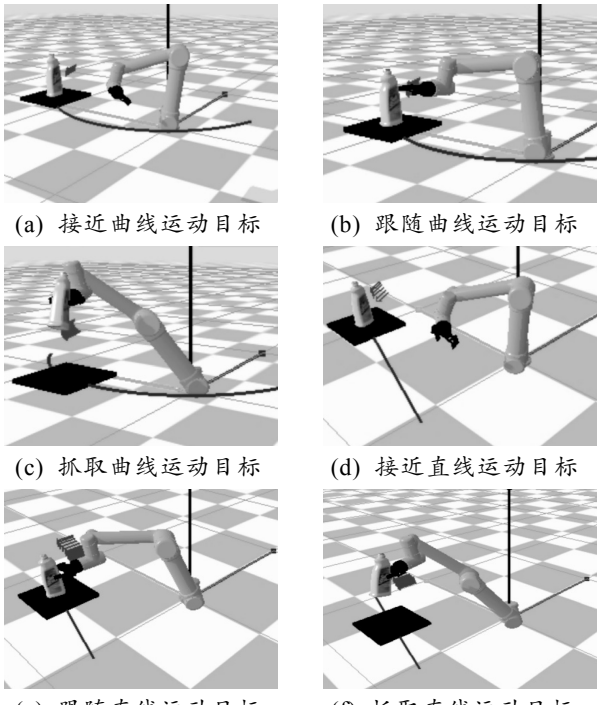
励值稳定增长，最终收敛到最高奖励 900 附近，证明了笔者提出的奖励函数能有效指导机械臂动态目标抓取任务的训练图 5 相比图 4，目标进行随机变速直线运动，每回合运动方向和运动速率都会发生变化，难度更大，因此收敛需要的时间步更多。

2.3　仿真结果定性分析

分析训练得到的动态目标抓取策略在新颖场景下的泛化性。使用图 5 中经过训练的策略进行泛化性测试，抓取的对象为未经训练的 7 cm/s 直线运动目标和 4 cm/s 曲线运动的目标，共计 200 次测试，抓取成功率与抓取时间如表 2 所示，过程如图 6 所示。

表 2　新颖运动轨迹目标抓取结果

运动模式	速度/(cm/s)	成功率	平均抓取时间/s
直线运动	7	0.61 ± 0.03	8.097 ± 0.15
曲线运动	4	0.78 ± 0.08	7.080 ± 0.15

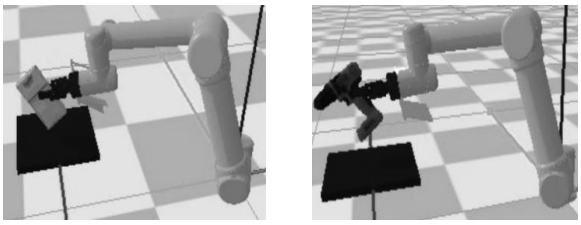


(a) 接近曲线运动目标　(b) 跟随曲线运动目标　(c) 抓取曲线运动目标　(d) 接近直线运动目标　(e) 跟随直线运动目标　(f) 抓取直线运动目标

图 6　机械臂在仿真环境中抓取新颖运动轨迹目标的流程

从上图可以看出，即使目标物处于新颖的运动轨迹，前文训练的抓取策略依然表现出良好的泛化能力。无需重新训练就能有效指导机械臂运动，抓取不同的运动轨迹与运动速率的目标。

除此以外，使用上述策略还可抓取其他物体，如酸奶盒子或电钻等，结果如图 7 所示。这进一步证明笔者的动态目标抓取方法具有较优的泛化能力。



(a) 酸奶盒　(b) 电动螺丝刀

图 7　机械臂在仿真环境中抓取新颖动态目标

2.4　仿真结果定量分析

动态目标抓取方法的可行性与泛化性已经得到初步验证，接下来与其他方法进行定量对比，指标设置如下：

成功率：成功抓取次数与总测试回合数的比值。

平均抓取时间：从物体开始运动到举起物体所用的时间，进行多次试验取平均值得到。

定量对比均为抓取运动速率为 4 cm/s 的随机曲线运动目标。采用图 5 中训练得到的策略，对比：使用 PID 控制，无迹卡尔曼滤波(unscented Kalman filter，UKF)预测的动态目标抓取方法(UKF+CP)，试验共计 200 次，结果如表 3 所示。

表 3　不同算法在相同试验条件下的抓取结果

方法	成功率	平均抓取时间/s
UKF+CP	0.32 ± 0.05	13.38 ± 0.30
OURS	0.78 ± 0.08	7.08 ± 0.15

综上所述，仿真结果表明笔者提出的基于深度强化学习的动态目标抓取方法能够应对不同速率、不同轨迹的运动目标，具有可行性与较优的泛化能力，相比基于经典规划控制的机械臂动态目标抓取方法有着更高的抓取成功率和更快的成功抓取速度。

3　结论

该方法利用机械臂自身的状态信息，环境中的目标信息和未来预抓取位姿，在奖励函数的指导下，经过训练，能够学习到使机械臂成功抓取动态目标的轨迹规划策略。在新颖的试验场景中，对该策略进行测试。定性分析结果表明：该策略能够有效地控制机械臂关节角度，实现对动态目标的六自由度抓取，并具备较优泛化能力。定量分析结果表明：该方法相比基于经典规划控制的机械臂动态目标抓取方法具有更高的抓取成功率与更快的抓取速度，进一步验证该方法能有效提升机械臂动态目标抓取任务的效率，具有可行性。

参考文献：

[1] AKINOLA I, XU J, SONG S, et al. Dynamic grasping with reachability and motion awareness[C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2021: 9422-9429.

[2] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. MIT press, 2018.

[3] CHEN P, LU W. Deep reinforcement learning based moving object grasping[J]. Information Sciences, 2021, 565: 62-76.

[4] WU T, ZHONG F, GENG Y, et al. Grasparl: Dynamic grasping via adversarial reinforcement learning[J]. arXiv preprint arXiv: 2203. 02119, 2022.

[5] PUTERMAN M L. Markov decision processes: discrete stochastic dynamic programming[M]. John Wiley & Sons, 2014.

[6] EPPNER C, MOUSAVIAN A, FOX D. A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set[C]// The International Symposium of Robotics Research. Cham: Springer International Publishing, 2019: 890-905.

[7] AKINOLA I, VARLEY J, CHEN B, et al. Workspace aware online grasp planning[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 2917-2924.

[8] SCHULMAN J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv: 1707.06347, 2017.

****

（上接第 86 页）

4 实验验证

为测试该备用电源在低温、常温和高温环境条件下的放电时间指标性能，将备用电源放入高低温环境试验箱内，按照测试连接示意图搭建实验测试平台。

将温度箱调节至-40 ℃后，待备用电源充电至 14.3 V 且均衡后，开始测试在底温下的放电时间，测试 3 次(电子负载调节为恒功率模式 10 W，电压 4.94 V，电流 2.02 A)。

将温度箱调节至 25 ℃后，待备用电源充电至 14.3 V 且均衡后，开始测试在常温下的放电时间，测试 3 次(电子负载调节为恒功率模式 10 W，电压 4.91 V，电流 2.03 A)；

将温度箱调节至 55 ℃后，待备用电源充电至 14.3 V 且均衡后，开始测试在高温下的放电时间，测试 3 次(电子负载调节为恒功率模式 10 W，电压 4.91 V，电流 2.03 A)。

备用电源放电 3 次试验结果见表 1。从测试数据可以看出备用电源在低温、常温和高温工况下，放电时间稳定，满足装备使用要求。

表 1 备用电源放电时间

环境温度/℃	第 1 次放电时间/s	第 2 次放电时间/s	第 3 次放电时间/s
低温-40	584	570	572
常温 25	581	573	576
高温 55	586	580	575

5 结束语

该系留多旋翼无人机备用电源充分利用了超级电容器固有的循环寿命长、工作范围宽等优点，严格落实国产化元器件应用尽用的自主可控要求，具有宽温度适应性、免维护、结构简单、国产化率较高等特点。试验结果表明：该备用电源在低温、高温、常温下的放电时间指标均优于设计要求，能够满足系留多旋翼无人机装备使用要求。

参考文献：

[1] 王锋, 逯振坤, 周国庆, 等. 系留多旋翼无人机技术进展及设计方法研究[J]. 机械工程师, 2019, 334(4): 68-72, 75.

[2] 龙文彪. 系留多旋翼无人机及其应用[J]. 科技创新导报, 2020, 17(2): 57-59.

[3] 王伟, 薛松, 廖士楠, 等. 系留多旋翼无人机模拟训练系统设计与实现[J]. 兵工自动化, 2023, 42(9): 1-5.

[4] 李志远, 傅航, 蔡华华. 系留多旋翼无人机在无线电监测中的应用分析[J]. 信息系统工程, 2021, 332(8): 10-12.

[5] 王世勇, 刘满, 倪蜂棋, 等. 系留无人机视觉定位技术[J]. 兵工自动化, 2022, 41(11): 77-83.

[6] 肖谧, 宿玉鹏, 杜伯学.超级电容器研究进展[J]. 电子元件与材, 2019, 38(9): 1-12.

[7] 梁晰童, 潘伟, 陈昆峰, 等. 新型超级电容器的研发进展[J]. 应用化学, 2016, 33(8): 867-875.