实验研究

文章编号:1671-251X(2021)01-0036-07 DOI:10.13272/j.issn.1671-251x.2020080047

# 基于强化学习的煤矸石分拣机械臂智能控制算法研究

张永超, 于智伟, 丁丽林

(山东科技大学 机械电子工程学院, 山东 青岛 266590)

扫码移动阅读

**摘要**:针对传统煤矸石分拣机械臂控制算法如抓取函数法、基于费拉里法的动态目标抓取算法等依赖于精确的环境模型、且控制过程缺乏自适应性,传统深度确定性策略梯度(DDPG)等智能控制算法存在输出动作过大及稀疏奖励容易被淹没等问题,对传统 DDPG 算法中的神经网络结构和奖励函数进行了改进,提出了一种适合处理六自由度煤矸石分拣机械臂的基于强化学习的改进 DDPG 算法。煤矸石进入机械臂工作空间后,改进 DDPG 算法可根据相应传感器返回的煤矸石位置及机械臂状态进行决策,并向相应运动控制器输出一组关节角状态控制量,根据煤矸石位置及关节角状态控制量控制机械臂运动,使机械臂运动到煤矸石附近,实现煤矸石分拣。仿真实验结果表明:改进 DDPG 算法相较于传统 DDPG 算法具有无模型通用性强及在与环境交互中可自适应学习抓取姿态的优势,可率先收敛于探索过程中所遇的最大奖励值,利用改进 DDPG 算法控制的机械臂所学策略泛化性更好、输出的关节角状态控制量更小、煤矸石分拣效率更高。

**关键词**:选煤;煤矸石分拣;分拣机器人;机械臂;关节角状态控制;强化学习;奖励函数;DDPG 算法

中图分类号:TD67 文献标志码:A

## Research on intelligent control algorithm of coal gangue sorting robot arm based on reinforcement learning

ZHANG Yongchao, YU Zhiwei, DING Lilin

(College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

**Abstract**: The problems of the traditional gangue sorting robot arm control algorithms such as the grasping function method and the dynamic target grasping algorithm based on Ferrary method are relying on an accurate environment model and lacking adaptivity in the control process. At the same time, the problems of the traditional intelligent control algorithms such as deep deterministic policy gradient (DDPG) are excessive output actions and sparse rewards that are easily covered. In order to solve these problems, this study improves the neural network structure and reward function in the traditional DDPG algorithm, and proposes an improved DDPG algorithm based on reinforcement learning, which is suitable for handling six-degree-of-freedom gangue sorting robot arms. After the gangue enters the working space of the robot arm, the improved DDPG algorithm can make decisions according to the gangue position and

**收稿日期**:2020-08-14;**修回日期**:2020-11-20;**责任编辑**:张强。

**基金项目**:山东省自然科学基金项目(ZR2018MEE036)。

**作者简介**:张永超(1977-),男,山东金乡人,讲师,博士研究生,主要研究方向为机电智能控制、流体机械,E-mail:jdxzyc@sdust.edu.cn。通信作者:于智伟(1995-),男,山东德州人,硕士研究生,主要研究方向为机器人智能控制,E-mail:yzw_edu@163.com。

**引用格式**:张永超,于智伟,丁丽林. 基于强化学习的煤矸石分拣机械臂智能控制算法研究[J]. 工矿自动化,2021,47(1):36-42.

ZHANG Yongchao, YU Zhiwei, DING Lilin. Research on intelligent control algorithm of coal gangue sorting robot arm based on reinforcement learning[J]. Industry and Mine Automation, 2021, 47(1): 36-42.

robot arm state returned by the corresponding sensor, and can output a set of joint angle state control quantity to the corresponding motion controller. The algorithm can control the movement of the robot arm according to the gangue position and joint angle state control quantity, so that the robot arm moves to the nearby gangue to conduct gangue sorting. The simulation results show that compared with the traditional DDPG algorithm, the improved DDPG algorithm has the advantages of model-free versatility and adaptive learning of grasping pose in interaction with the environment. Moreover, the improved algorithm can be the first to converge to the maximum reward value encountered during exploration. The robot arm controlled by the improved DDPG algorithm has better policy generalization, smaller joint angle state control output and higher gangue sorting efficiency.

**Key words**: coal preparation; coal gangue sorting; sorting robot; robot arm; joint angle state control; reinforcement learning; reward function; DDPG algorithm

## 0　引言

煤矸石分拣是煤炭粗选的首要环节,也是提高煤炭质量以及矿井效益的重要方法[1]。传统煤矸石分拣如人工分拣、湿选和干选等分拣方式正面临工伤风险率高、环境污染严重及智能化程度低的困境[2-3]。而机械臂分拣不仅能有效降低工伤风险率,同时还具有效率高、绿色分拣的优势。《关于加快煤矿智能化发展的指导意见》也明确提出对具备条件的煤矿要加快智能化改造,推进危险岗位的机器人作业,到 2035 年各类煤矿基本实现智能化,建成智能感知、智能决策、自动执行的煤矿智能化体系。煤矸石分拣朝着智能机器人化方向发展符合现代工业发展趋势。

目前,用于机器人分拣机械臂的控制算法主要有抓取函数法[4]、基于费拉里法的动态目标抓取算法[5]、金字塔形寻优算法[6]、比例导引法[7]等。这些控制算法中除抓取函数法仅适用静态目标外,其他控制算法均可在已获取目标位置的前提下,实现快速接近目标的目的。然而将以上控制算法应用于煤矸石分拣机械臂时发现,煤矸石分拣机械臂的工作效率严重受限于控制算法中机械臂运动学或动力学等环境模型的设计精度,若精度低则会使煤矸石分拣机械臂末端执行器不能良好接近目标,造成较高的漏选率[8]。王鹏等[9]设计了一种基于机器视觉技术的多机械臂煤矸石分拣机器人系统,该系统可高效分拣 50～260 mm 粒度的煤矸石,并且能良好适应不同带速,但其机械臂控制算法仍依赖模型的精确性。因此,研究一种可使煤矸石分选机械臂稳定高效运行的无模型智能控制算法对实现煤矸石分选作业无人化、智能化具有重要意义。

强化学习作为一种重要的机器学习方法,因其控制过程不依赖于精确环境模型的特点,在机器人智能控制领域具有重要地位。它强调在与环境的交互中获取反映真实目标达成度的反馈信号,强调模型的试错学习和序列决策行为的动态和长期效应。深度确定性策略梯度(Deep Deterministic Policy Gradient,DDPG)算法是强化学习中常用来处理具有连续性动作空间任务的一种优秀控制算法[10],但传统 DDPG 算法的奖励函数仅是基于欧几里得距离设计的,缺乏动作惩罚项等约束项,训练过程中容易使机械臂学习到输出较大关节角控制量的策略。若奖励函数设计不合理,还会出现稀疏奖励问题,使机械臂学不到期望的策略[11]。

针对以上问题,本文提出了一种基于强化学习的改进 DDPG 算法,并用于煤矸石分拣机械臂的控制。通过设计奖励函数使机械臂能更快学习到平稳接近目标的策略。通过改进 Actor 和 Critic 神经网络结构以及使用批标准化和 Dropout 优化方法,使算法具有更稳健处理机械臂关节角等低维输入观测值的能力[12]。仿真实验结果表明,改进的 DDPG 算法与传统控制算法相比具有无模型通用性强及在与环境交互中可自适应学习抓取姿态的优势,并且学习能力优于传统 DDPG 算法,其在大型连续动作空间具有巨大的潜在应用价值。

## 1　DDPG 算法

DDPG 算法属于 Actor-Critic 算法,它在确定性策略梯度(Deterministic Policy Gradient,DPG)算法[13]的基础上融入了深度神经网络[14]。DDPG 算法的参数更新思想借鉴了深度 Q 网络(Deep Q Network,DQN)算法中的双网络延时更新和经验回放机制来切断数据相关性,训练过程中使用批量标准化[15]来提高学习效率。

DDPG 算法中的 Actor 策略网络可在连续动作空间进行探索以及做出动作决策,Critic 价值网络则负责评判决策优劣,以此指引 Actor 策略网络参数更新。两者均包含 2 个具有相同结构的神经网

络，即目标网络和估计网络。Actor 的估计网络是根据 Critic 的估计网络的指引实时更新，Critic 的估计网络是根据自身目标网络的指引实时更新，因此，估计网络的参数是最新参数，而目标网络参数则是根据估计网络参数使用滑动平均方式来延迟更新的，Critic 的估计网络参数的更新方式为最小化均方误差 $L$：

$$\begin{cases} L = \dfrac{1}{N}\sum_i (y_i - Q(s_i, a_i \mid \theta^Q))^2 \\ y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} \mid \theta^{\mu'}) \mid \theta^{Q'}) \end{cases} \tag{1}$$

式中：$N$ 为从经验池中取出的数据总个数；$y_i$ 为 $i$ 时刻目标 $Q$ 值的更好估计；$Q(s_i, a_i \mid \theta^Q)$ 为估计 $Q$ 值；$s_i$ 为 $i$ 时刻的环境状态；$a_i$ 为 $i$ 时刻的输入动作；$\theta^Q$ 为 Critic网络中估计网络的参数，$Q$ 代表 Critic 网络中的估计网络；$r_i$ 为 $i$ 时刻的即时奖励；$\gamma$ 为折扣率；$Q'(s_{i+1}, \mu'(s_{i+1} \mid \theta^{\mu'}) \mid \theta^{Q'})$ 为目标 $Q$ 值；$\mu'(s_{i+1} \mid \theta^{\mu'})$ 为确定性策略，$\mu'$ 为 Actor 网络中的目标网络，$\theta^{\mu'}$ 为 Actor 网络中目标网络的参数；$\theta^{Q'}$ 为 Critic 网络中目标网络的参数，$Q'$ 为 Critic 网络中的目标网络。

这种利用估计网络和目标网络进行参数更新的双网络机制切断了数据相关性，有效提高了算法收敛性。类似于监督学习标签的 $y_i$ 为 Critic 估计网络参数的更新指明了方向，使参数更新更稳定。

Actor 策略网络基于策略梯度更新网络参数：

$$\nabla_{\theta^\mu} J(\theta^\mu) \approx \frac{1}{N}\sum_i \nabla_a Q(s, a \mid \theta^Q)\big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\big|_{s_i} \tag{2}$$

式中：$\nabla_{\theta^\mu} J(\theta^\mu)$ 为性能指标 $J(\theta^\mu)$ 对神经网络参数 $\theta^\mu$ 的梯度，$\mu$ 为 Actor 网络中的估计网络；$a$ 为输入动作；$s$ 为环境状态；$\mu(s_i)$ 为 Actor 网络中的估计网络在环境状态为 $s_i$ 时输出的确定性策略；$\nabla_a Q(s, a \mid \theta^Q)\big|_{s=s_i, a=\mu(s_i)}$ 表示环境状态为 $s_i$、动作为 $\mu(s_i)$ 时，估计 $Q$ 值对动作 $a$ 的梯度；$\mu(s \mid \theta^\mu)$ 为 Actor 网络中估计网络学习到的确定性策略；$\nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\big|_{s_i}$ 为环境状态为 $s_i$ 时，$\mu(s \mid \theta^\mu)$ 对 $\theta^\mu$ 的梯度。

$\nabla_{\theta^\mu} J(\theta^\mu)$ 可使 Actor 策略网络向着可获得最大奖励的方向不断调整自身网络参数 $\theta^\mu$。DDPG 算法结构如图 1 所示。DDPG 算法在处理具有连续动作空间的任务上具有很大优势，但其收敛性能受奖励函数以及所选特征影响很大，较差的奖励函数不仅会增加机械臂的无效探索次数，还会使收敛难度增大，因此，很难将 DDPG 算法直接应用于机械臂控制任务。

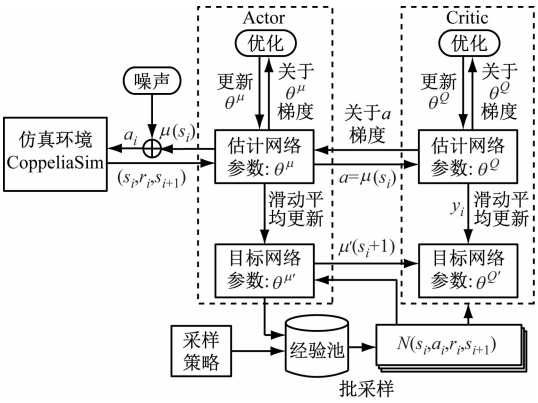


图 1 DDPG 算法结构

Fig. 1 DDPG algorithm structure

## 2 改进 DDPG 算法

### 2.1 神经网络结构

改进 DDPG 算法应用对象为六自由度煤矸石分拣机械臂，其动作空间庞大且应用环境复杂，为使神经网络更容易学习其输入特征数据的分布，同时也为缓解训练过程中神经网络的过拟合问题，Actor 网络使用 4 层计算层并引入批量标准化方法和 Dropout 方法，以使模型训练更稳健，Actor 网络的双网络均采用图 2 所示的神经网络结构。

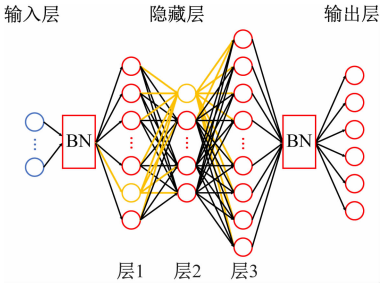


图 2 Actor 神经网络结构

Fig. 2 Actor neural network structure

Actor 框架中的估计网络输入层接收环境实时状态 $s_i$，目标网络输入层接收智能体与环境交互后的下一状态 $s_{i+1}$。批量标准化（Batch Normalization，BN）层通过减小数据在神经网络层传递过程中数据分布的改变，使神经网络参数更新更快且更易学习到知识，又因其是在小批量样本上计算均值和方差，难免引入小噪声，这也同时提高了神经网络的泛化能力。隐藏层中层 1 和层 2 中的黄色神经元表示该层使用了 Dropout 方法，Dropout 方法通过使部分神经元随机失活以减小神经元之间的依赖性，促使神经网络学习更稳健的特征，提高了网络的鲁棒性。

Critic 网络的估计网络和目标网络采用如图 3 所示的 3 层全连接神经网络来拟合 $Q(s_i, a_i \mid \theta^Q)$ 和 $Q'(s_{i+1}, \mu'(s_{i+1} \mid \theta^{\mu'}) \mid \theta^{Q'})$，并使用 BN 层和 Dropout 方法缓解神经网络过拟合，从而使 Critic 网络可更

好评判 Actor 网络输出动作的优劣。

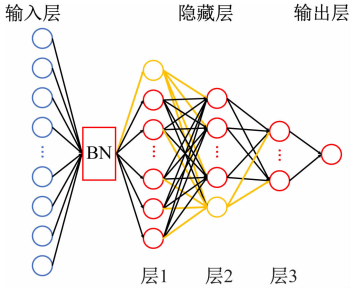


图 3　Critic 神经网络结构

Fig. 3　Critic neural network structure

### 2.2　奖励函数设计

为使煤矸石分拣机械臂可自适应学习到接近目标的方法,同时解决传统机械臂强化学习控制方法输出动作过大以及稀疏奖励容易被淹没的问题,本文在传统 DDPG 算法的基础上设计了可持续输出收益的奖励函数:

$$r_i=-c_1d_t^2-c_2\frac{\exp(d_t^2)-\exp(-d_t^2)}{\exp(d_t^2)+\exp(-d_t^2)}-(\|a_t\|_2)^{\frac{1}{6}}+\frac{10}{d_t^2+1}+b \tag{3}$$

式中:$c_1$,$c_2$ 为常值系数;$d_t$ 为采样时刻 $t$ 机械臂末端执行器特定位置与煤矸石位置的距离,当机械臂末端执行器从较远位置接近目标煤矸石时,$d_t^2$ 迅速减小,当距离较近时,$d_t^2$ 变化缓慢;$\frac{\exp(d_t^2)-\exp(-d_t^2)}{\exp(d_t^2)+\exp(-d_t^2)}$为双曲正切函数,其在机械臂末端执行器特定位置距离目标较远时接近常值,当距离较近时,该项迅速减小,与 $d_t^2$ 动态配合,可使机械臂更快学习到精确接近目标的方法;$(\|a_t\|_2)^{\frac{1}{6}}$ 为输出动作惩罚项,当 $d_t$ 较大时,该项可减少因改进 DDPG 算法输出过大的控制量而使机械臂性能受损等问题;$\frac{10}{d_t^2+1}$项可进一步诱导机械臂末端执行器接近目标位置,距离越小,则获得收益越大;$b$ 为终止状态判定项,如果机械臂末端执行器特定位置与目标煤矸石距离为 0,$b$ 被赋值为一个正值,若下一状态距离持续为 0,则 $b$ 值逐渐增加,当达到指定终止值时,本幕提前结束,若机械臂运行过程中,出现一次 $d_t^2$ 不为 0 的动作,则 $b$ 直接被赋值为 0。

终止状态判定项通过给可输出连续跟踪目标动作的智能体一个较大的奖励值来诱导机械臂学习到跟踪目标的能力,减小目标与机械臂末端执行器的冲击。

## 3　煤矸石分拣机械臂智能控制原理

改进 DDPG 算法依据马尔科夫决策过程[16]控制机械臂在与环境交互中学习知识,任务是找到能最大化智能体期望回报的最优策略[17],即最大化估计 $Q$ 值。在煤矸石分拣机械臂控制任务中,强化学习中的智能体为具有决策能力的煤矸石分拣机械臂。估计 $Q$ 值的更新依据自身梯度以及其与目标 $Q$ 值 $q$ 的均方误差来指引,其中 $q$ 为

$$q=r_i+\gamma Q'(s_{i+1},\mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}) \tag{4}$$

机械臂根据相应传感器返回的煤矸石位置以及机械臂状态进行决策,输出机械臂预测控制量,并根据机械臂执行动作之后的环境状态如末端执行器与煤矸石的距离、机械臂的位置、煤矸石的位置等进行下一步决策。通过不断试错学习,使网络参数向可输出机械臂探索过程中所遇最大 $q$ 值方向更新,直到机械臂可以根据环境状态准确输出接近煤矸石的策略或幕奖励波动趋于稳定时即可提前终止训练。煤矸石分拣机械臂控制流程如图 4 所示。首先初始化改进 DDPG 算法权重,然后加载训练权重,最后根据煤矸石位置及关节角状态输出相应的关节角来控制机械臂运动,以使末端执行器接近煤矸石。煤矸石位置以及机械臂关节角状态由相应检测模块和传感器测得。

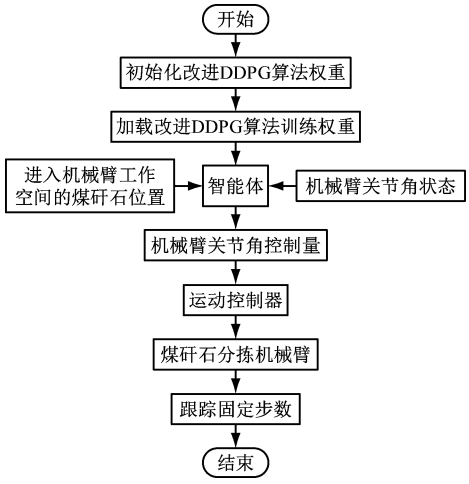


图 4　机械臂控制流程

Fig. 4　Control flow of robotic arm

## 4　实验及结果分析

### 4.1　环境配置

本文采用 CoppeliaSim 搭建虚拟环境模型,并在环境模型中创建 UR5 机械臂、RG2 执行器、输送带、煤和煤矸石,以此模拟实际煤矸石分拣环境,仿真平台如图 5 所示。

使用 Python 在 tensorflow2.2.0 上编写改进 DDPG 算法框架,并调用 CoppeliaSim 创建的虚拟环境来训练算法。算法中 Actor 的双网络和 Critic 的双网络神经元配置见表 1,其神经网络输入状态 $s_i$ 为 6 个关节角度、执行器指定点的绝对坐标、

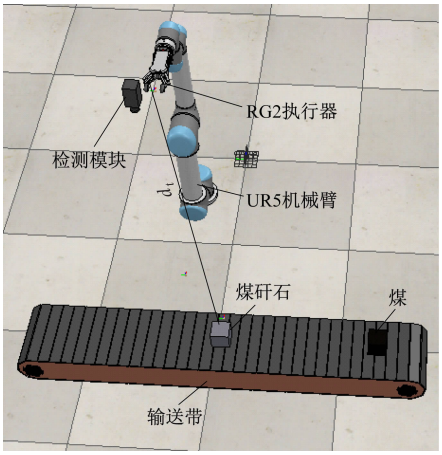


图 5 仿真平台

Fig. 5 Simulation platform

表 1 Actor 和 Critic 神经网络中各层神经元数量

Table1 Number of neurons in each layer in Actor and Critic neural networks

| 神经网络 | 输入层 | 隐藏层 1 | 隐藏层 2 | 隐藏层 3 | 输出层 |
|---|---|---|---|---|---|
| Actor | 14 | 700 | 550 | 800 | 6 |
| Critic | 300 | 200 | 80 | 50 | 1 |

第 5 个关节的绝对坐标、末端执行器指定点与煤矸石中心的距离。奖励函数参数 $c_1$ 和 $c_2$ 经测试设定为 0.1 和 0.2，使用 30 倍的高斯分布噪声，设定算法循环幕数为 5 000，每幕迭代步数为 300，若末端执行器中设定的特定点与煤矸石中心距离为 0 时，超参数 $b$ 被赋值为 10，若下一步距离仍为 0，则 $b$ 继续累加 10，当其累加了 20 步时，即跟踪了 20 步，则提前结束当前幕。

### 4.2 性能对比测试

为测试改进 DDPG 算法的性能，将其与仅使用欧几里得距离作为奖励函数的传统 DDPG 算法在相同环境下进行训练对比，经 5 000 幕迭代后，改进 DDPG 算法每幕奖励结果如图 6 所示，传统 DDPG 算法每幕奖励结果如图 7 所示。

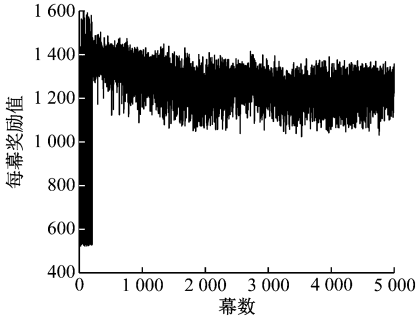


图 6 改进 DDPG 算法每幕奖励

Fig. 6 Reward for each episode of improved DDPG algorithm

强化学习不同于传统监督学习，其训练样本没有确定标签，也就没有真正意义上的损失函数，因此，当幕奖励趋于稳定时，即可认为机械臂在当前环境中学习到了基于所用算法的最优策略。由图 6 和图 7 可看出，改进 DDPG 算法和传统 DDPG 算法指引下的机械臂在环境中经过随机探索后都学习到了一种策略，改进 DDPG 算法训练稳定后的每幕奖励接近机械臂探索过程中的最大值，而传统 DDPG 算法训练稳定后的每幕奖励与机械臂探索过程中遇到的最大奖励相差较大，这表示所设计的奖励函数比传统奖励函数更具有诱导机械臂学习到最优策略的能力。

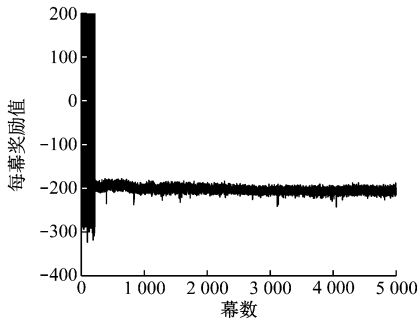


图 7 传统 DDPG 算法每幕奖励

Fig. 7 Rewards for each episode of traditional DDPG algorithm

为了进一步测试 2 种算法训练后的性能，将训练后的改进 DDPG 算法和传统 DDPG 算法进行泛化能力对比验证，测试过程为在虚拟环境中继续添加红色、黄色、绿色物块，如图 8 所示，图中灰色物块为 2 种算法在学习过程中所用的训练目标，使用这 4 种颜色的物块代替实际分拣环境中的煤矸石，将这 4 种煤矸石分别设定为测试目标，并使用 2 种算法分别对其进行 10 次接近测试，结果见表 2。

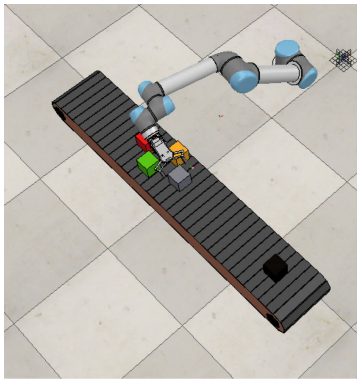

图 8 CoppeliaSim 虚拟环境

Fig. 8 CoppeliaSim virtual environment

表 2 测试结果

Table 2 Test results

| 算法 | 黄色成功接近次数 | 绿色成功接近次数 | 红色成功接近次数 | 灰色成功接近次数 |
|---|---|---|---|---|
| 改进 DDPG | 7 | 7 | 4 | 10 |
| 传统 DDPG | 3 | 1 | 0 | 3 |

由表 2 可看出,改进 DDPG 算法具有良好的泛化能力,即接近训练过程中未曾出现过目标的能力,而传统 DDPG 算法泛化能力较差,这也进一步验证了改进 DDPG 算法控制下的机械臂所学策略质量较高。在接近测试过程中,传统 DDPG 算法控制下的煤矸石分拣机械臂的末端执行器多数以一种垂直于虚拟环境中地面的非常规姿态位于物块之上,这也证明了传统 DDPG 算法控制下的机械臂并未学习到良好的策略。在黄色物块接近测试中,传统 DDPG 算法控制下的煤矸石分拣机械臂末端执行器特定位置的运动轨迹如图 9 所示,改进 DDPG 算法控制下的末端执行器特定位置运动轨迹如图 10 所示。从图 9 和图 10 可看出,改进 DDPG 算法与传统 DDPG 算法相比,其控制策略可以更好地使末端执行器接近目标。

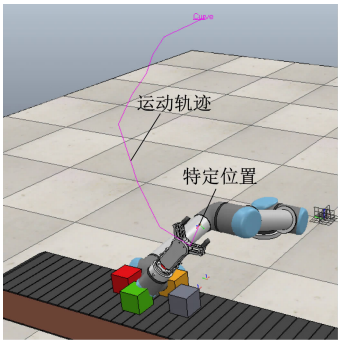


图 9　传统 DDPG 算法控制结果

Fig. 9　Traditional DDPG algorithm control results

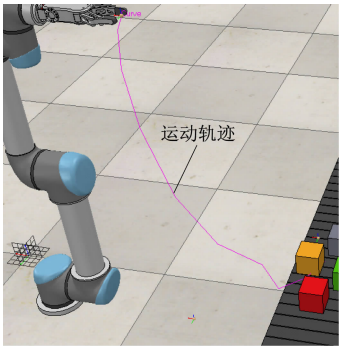


图 10　改进 DDPG 算法控制结果

Fig. 10　Improved DDPG algorithm control results

黄色物块接近测试中,2 种算法输出的一次关节角控制量对比如图 11 所示。从图 11 可看出,改进 DDPG 算法因受奖励函数输出动作惩罚项的影响,其关节角输出控制量小于传统 DDPG 算法的输出,降低了因大幅度更新关节角对机械臂性能造成的影响。

综上可知,改进 DDPG 算法和传统 DDPG 算法控制下的煤矸石分拣机械臂通过与环境交互均可使机械臂学习到一种控制策略,但测试结果表明,传统 DDPG 算法所学策略性能较差,不能良好接近目标,

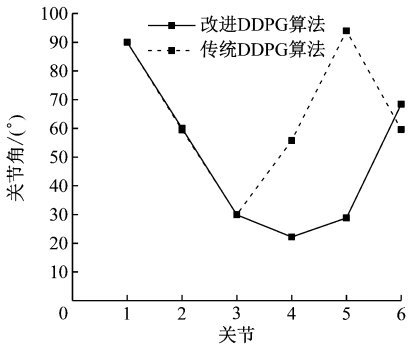


图 11　2 种算法输出的一次关节角控制量对比

Fig. 11　Comparison of output control quantity of a joint angle of two algorithms

改进 DDPG 算法所学策略更接近最优。

## 5　结论

(1) 对传统 DDPG 算法中的神经网络结构和奖励函数进行了改进,提出了适合处理六自由度煤矸石分拣机械臂的改进 DDPG 算法。改进 DDPG 算法经过训练后,Actor 网络即可根据环境状态映射成具有最大估计 $Q$ 值的关节状态控制量,最终经相应运动控制器控制煤矸石分拣机械臂接近煤矸石,实现煤矸石分拣。

(2) 实验结果表明,改进 DDPG 算法相较于传统 DDPG 算法可更快收敛于探索过程中所遇最大奖励值。改进 DDPG 算法可控制机械臂小幅度接近煤矸石,具有较强的泛化能力,为解决实际煤矸石机械臂智能分拣问题提供了一种新方法,同时也为后期与基于深度学习的视觉检测联合使用提供了理论保障。

**参考文献(References):**

[1] MA D,DUAN H,LIU J,et al. The role of gangue on the mitigation of mining-induced hazards and environmental pollution:an experimental investigation [J]. Science of the Total Environment,2019,664:436-448.

[2] YANG X,ZHAO Y,LUO Z,et al. Fine coal dry beneficiation using autogenous medium in a vibrated fluidized bed [J]. International Journal of Mineral Processing,2013,125:86-91.

[3] 夏云凯,李功民. 我国动力煤干选技术现状及展望[J]. 洁净煤技术,2017,23(6):21-29.
XIA Yunkai, LI Gongmin. Current situation and prospects of power coal dry separation technology in China[J]. Clean Coal Technology,2017,23(6):21-29.

[4] JOHNS E,LEUTENEGGER S,DAVISON A J. Deep learning a grasp function for grasping under gripper pose uncertainty[C]//2016 IEEE/RSJ International

Conference on Intelligent Robots and Systems (IROS),2016:4461-4468.

[5] 苏婷婷,张好剑,王云宽,等. 基于费拉里法的 Delta 机器人动态目标抓取算法[J]. 华中科技大学学报(自然科学版),2018,46(6):128-132.
SU Tingting, ZHANG Haojian, WANG Yunkuan, et al. Dynamic picking algorithm based on Ferrari's method for Delta robot[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition),2018,46(6):128-132.

[6] 张朝阳. 基于视觉的机器人废金属分拣系统研究[D]. 北京:中国农业大学,2015.
ZHANG Chaoyang. Research on scrap metal sorting robot with machine vision[D]. Beijing: China Agricultural University,2015.

[7] 崔彦凯,梁晓庚,王斐,等. 弹道导弹助推段拦截最优制导律设计[J]. 飞行力学,2011,29(1):59-62.
CUI Yankai, LIANG Xiaogeng, WANG Fei, et al. Design of optimal guidance law for interception ballistic missile during the boost phase[J]. Flight Dynamics,2011,29(1):59-62.

[8] 王鹏,曹现刚,马宏伟,等. 基于余弦定理-PID 的煤矸石分拣机器人动态目标稳准抓取算法研究[J/OL]. 煤炭学报:1-8[2020-11-18]. https://doi.org/10.13225/j.cnki.jccs.2019.1565.
WANG Peng, CAO Xiangang, MA Hongwei, et al. Research on dynamic target steady and accurate grasping algorithm of gangue sorting robot based on cosine theorem-PID[J/OL]. Journal of China Coal Society:1-8[2020-11-18]. https://doi.org/10.13225/j.cnki.jccs.2019.1565.

[9] 王鹏,曹现刚,夏晶,等. 基于机器视觉的多机械臂煤矸石分拣机器人系统研究[J]. 工矿自动化,2019,45(9):47-53.
WANG Peng, CAO Xiangang, XIA Jing, et al. Research on multi-manipulator coal and gangue sorting robot system based on machine vision[J]. Industry and Mine Automation,2019,45(9):47-53.

[10] QIU C, HU Y, CHEN Y, et al. Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications[J]. IEEE Internet of Things Journal,2019,6(5):8577-8588.

[11] 杨惟轶,白辰甲,蔡超,等. 深度强化学习中稀疏奖励问题研究综述[J]. 计算机科学,2020,47(3):182-191.
YANG Weiyi, BAI Chenjia, CAI Chao, et al. Survey on sparse reward in deep reinforcement learning[J]. Computer Science,2020,47(3):182-191.

[12] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature,2015,521(7553):436-444.

[13] ZHANG Y, ZHAO B, LIU D. Deterministic policy gradient adaptive dynamic programming for model-free optimal control[J]. Neurocomputing,2020,387:40-50.

[14] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature,2015,518(7540):529-533.

[15] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//32nd International Conference on Machine Learning, 2015:448-456.

[16] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press,2018.

[17] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey[J]. IEEE Signal Processing Magazine, 2017,34(6):26-38.