doi:10.6041/j.issn.1000-1298.2020.S2.001

基于深度强化学习的虚拟机器人采摘路径避障规划

熊俊涛　李中行　陈淑绵　郑镇辉
(华南农业大学数学与信息学院,广州 510642)

摘要:针对采摘机器人在野外作业环境中,面临采摘任务数量多,目标与障碍物位置具有随机性和不确定性等问题,提出一种基于深度强化学习的虚拟机器人采摘路径避障规划方法,实现机器人在大量且不确定任务情况下的快速轨迹规划。根据机器人本体物理结构设定虚拟机器人随机运动策略,通过对比分析不同网络输入观测值的优劣,结合实际采摘行为设置环境观测集合,作为网络的输入;引入人工势场法目标吸引和障碍排斥的思想建立奖惩函数,对虚拟机器人行为进行评价,提高避障成功率;针对人工势场法范围斥力影响最短路径规划的问题,提出了一种方向惩罚避障函数设置方法,将障碍物范围惩罚转换为单一方向惩罚,通过建立虚拟机器人运动碰撞模型,分析碰撞结果选择性给予方向惩罚,进一步优化了规划路径长度,提高采摘效率;在 Unity 内搭建仿真环境,使用 ML-Agents 组件建立分布式近端策略优化算法及其与仿真环境的交互通信,对虚拟机器人进行采摘训练。仿真实验结果显示,不同位置障碍物设置情况下虚拟机器人完成采摘任务成功率达 96.7% 以上。在 200 次随机采摘实验中,方向惩罚避障函数方法采摘成功率为 97.5%,比普通奖励函数方法提高了 11 个百分点,采摘轨迹规划平均耗时 0.64 s/次,相较于基于人工势场法奖励函数方法降低了 0.45 s/次,且在连续变动任务实验中具有更高的适应性和鲁棒性。研究结果表明,本系统能够高效引导虚拟机器人在避开障碍物的前提下快速到达随机采摘点,满足采摘任务要求,为真实机器人采摘路径规划提供理论与技术支撑。

关键词:采摘机器人;路径规划;避障;深度强化学习;人工势场法
中图分类号:TP391.9　　**文献标识码**:A　　**文章编号**:1000-1298(2020)S2-0001-10

Obstacle Avoidance Planning of Virtual Robot Picking Path Based on Deep Reinforcement Learning

XIONG Juntao　LI Zhonghang　CHEN Shumian　ZHENG Zhenhui
(*College of Mathematics and Informatics*, *South China Agricultural University*, *Guangzhou* 510642, *China*)

Abstract: In the field environment, picking robots are faced with the problems of a large number of picking tasks, randomness and uncertainty in the positions of targets and obstacles, and so on. Traditional picking path planning methods usually use kinematics equations combined with the shortest path algorithm to solve them, while takes a lot of time to calculate in each planning. In order to improve the efficiency of trajectory planning to adapt to the field picking environment, a virtual robot picking path planning method based on deep reinforcement learning was proposed. Firstly, the virtual robot random action strategies were set according to the real robot physical structure, and the environment observation set was rationally set as the input of the network by analyzing the actual picking behavior. Establishing reward function with the reference to the idea of target attraction and obstacle rejection in artificial potential field method, which was used to evaluate the behavior of virtual robots and improve the success rate of obstacle avoidance. Aiming at the problem that the range repulsion of the artificial potential field method affected the shortest path planning, a directional penalty obstacle avoidance function setting method was proposed, which converted the obstacle range penalty into a single direction penalty. Besides, by establishing a virtual robot motion collision model, the direction penalties were giving selectively by analysis results of the model. Finally, a simulation environment in Unity was built, and the

收稿日期:2020-08-05　修回日期:2020-09-10
基金项目:国家自然科学基金项目(32071912)、广东省自然科学基金项目(2018A030313330)、广州市科技计划项目(202002030423)和国家级大学生创新创业训练计划项目(201910564033)
作者简介:熊俊涛(1981—),男,副教授,主要从事农业机器人和智能设计与制造研究,E-mail:xiongjt2340@163.com

distributed proximal policy optimization algorithm was used to train the virtual robot. The simulation experiment results showed that the success rate of the virtual robot in completing the picking task was over 96.7% under the condition of obstacles in different positions. In 200 random picking experiments, the directional penalty obstacle avoidance function method had a picking success rate of 97.5%, which was 11 percentage points higher than the ordinary reward function method, and the picking trajectory planning took an average of 0.64 s/time, which was 0.45 s/time shorter than the artificial potential field method. The research results showed that the system can efficiently guide virtual robots to quickly reach random picking points under the premise of avoiding obstacles, and met the requirements of picking tasks, which provided theoretical and technical support for real robot picking path planning.

Key words: picking robot; route planning; obstacle avoidance; deep reinforcement learning; artificial potential field method

0 引言

中国是世界上重要的水果生产国之一,其主要水果品种产量自 2012 年以来一直居世界第一[1]。水果采摘所需劳动力约占整个生产过程所需劳动力的 40%,采摘的效率与品质很大程度上影响到水果的经济效益[2]。随着人口老龄化速度加快,人工成本越来越高,仅依靠人工采摘不能满足经济高效的市场需求。农业自动采摘机器人能够有效代替人工完成采摘任务,很大程度上解决了人工不足的问题[3]。目前针对采摘机器人的研究主要分为 4 部分:机械结构设计、视觉识别定位、路径规划以及轨迹控制。在非结构性的工作环境下,如何让采摘机器人在保证安全作业的前提下准确、快速地到达采摘点,完成采摘任务,是机械臂控制领域的难点,也是决定机器人采摘是否成功和高效的关键之处,因此,采摘机器人的路径规划十分重要[4]。

在机械臂路径规划方面,国内外学者开展了相关研究,大部分方法集中在通过传统的路径规划算法达到目标,如:快速扩展随机数法[5](Rapidly exploring random tree, RRT)、A * 算法[6]、人工势场法[7]等。由于机械臂的运动规划是在三维空间下,从起始位姿到目标位姿的一组连续关节轨迹序列,导致使用传统的路径规划算法会出现“指数爆炸”等难题[8]。为此,GÓMEZ-BRAVO 等[9]在使用 RRT 算法提供一组可行轨迹的基础上,利用遗传算法根据其适应度函数优化原始解,并进行了有效性验证;CAO 等[10]提出一种结合 CMPSO 和 IPM 的混合算法进行柔性冗余机械臂的运动规划,成功对三连杆机器人规划最佳位姿。薛阳等[11]通过改进人工势场算法,加入虚拟吸引点避免算法陷入局部最优,实现对双机械臂避障的轨迹规划;王毅等[12]根据采摘机械臂的特殊结构,提出第 5 关节分离的姿态规划方法,对机械臂不同采摘方式进行路径规划。上述方法大多使用运动学方程和动力学方程对机械臂进行建模分析,并且通过转换空间的方式进行路径搜索。这导致其在路径规划的过程中利用运动学方程判断位置,计算量大,并且随着环境障碍和目标点的改变,需要重新计算空间映射,难以实现实时动态规划[13-14]。强化学习的出现恰好可以解决这一问题,通过不断试错的方式进行对环境的感知和学习,达到轨迹规划的目的。MEYES 等[15]使用强化学习对机器人进行沿线轨迹规划,用于解决生产焊接等问题;周祺杰等[16]利用深度 Q 网络算法训练机械手对矩形区域固体放射性废物进行抓取;李跃等[17]利用深度强化学习算法,通过设计新型奖励函数解决机械臂学习效率偏低的问题。上述方法主要针对工业生产、维修等领域,机器人作业环境稳定,其作业对象位置固定,作业行为相对规范。而在野外采摘环境中,采摘的目标点和障碍物都具有随机性和不确定性,需要进一步提高机器人轨迹规划的效率。

针对上述问题,本文以六自由度采摘机器人为研究对象,设计基于强化学习的虚拟机器人采摘路径避障规划系统。首先,根据机器人物理结构设定虚拟机器人随机动作策略,通过分析实际采摘行为合理设置环境观测量,作为网络的输入。引入人工势场法目标吸引和障碍排斥的思想建立奖惩函数,针对人工势场法范围斥力影响路径规划的问题,通过对虚拟机器人进行运动碰撞分析,提出一种方向惩罚避障函数,对虚拟机器人行为进行评价,引导虚拟机器人在躲避障碍物前提下尽快到达目标采摘点。最后使用 ML-Agents(Machine learning Agents)组件建立仿真环境与强化学习间的交互通信,利用分布式近端策略优化算法(Distributed proximal policy optimization, DPPO)对虚拟机器人进行采摘训练,以期实现对虚拟采摘机器人避障路径的智能规划。

1 系统整体设计

虚拟机器人采摘路径避障规划系统在 Unity 下

完成构建，总体结构如图 1 所示。搭建机器人仿真模型与采摘环境，依据机器人物理结构设定各虚拟机器人转动臂的动作及活动范围；根据机器人采摘机制，添加虚拟机器人末端执行器位置，采摘点及障碍物三者三维坐标等必要状态信息（观测量），设计导引奖赏机制，对虚拟机器人的行为进行评价；所有的环境信息通过 ML－Agents 组件与 Python API 完成交互，DPPO 算法与神经网络根据所收集的信息，进行网络参数的迭代更新，并将效果实时反馈于仿真环境中；在完成训练后，系统能够引导虚拟机器人到达采摘范围内的采摘任务点，实现虚拟机器人避障路径规划。为了进一步提高系统可操作性，搭建交互界面，用户可设置确定采摘点进行路径规划，获取结果路径与机器人位姿，用以进行实物规划参考与分析。

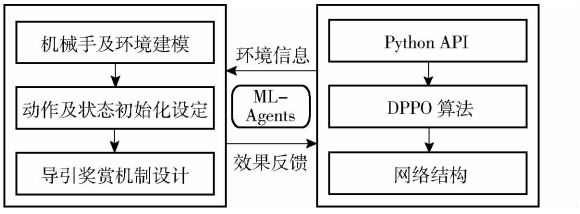


图 1　虚拟机器人采摘路径避障规划系统总体结构图

Fig. 1　Overall structure diagram of virtual robot picking path obstacle avoidance planning system

2　分布式近端策略优化算法

分布式近端策略优化算法 DPPO[18] 的概念由谷歌的 DeepMind 提出，核心是近端策略优化算法 PPO。PPO 算法[19] 是 OpenAI 最先提出的一种基于 Actor－Critic 框架的深度强化学习算法，其目的是解决 Actor 中策略梯度（Policy gradient，PG）算法网络参数更新慢、难以确定学习率（学习步长）的问题。学习速率如果设置过大，其动作策略会产生大幅度波动，导致其难以收敛；学习速率如果偏小，参数更新慢，其训练时间会极大提升，无法满足实验要求。PPO 算法则根据新旧策略比例，限制新策略的更新幅度，使得 PG 算法可以在更大的学习率上进行训练并且收敛。

Actor 中 PG 算法的目标函数为

$$J(\theta)=E_t[A(s_t,a_t)\pi_\theta(a_t|s_t)] \tag{1}$$

式中　$\pi(\cdot)$——策略函数　　θ——Actor 网络参数

s_t、a_t——第 t 步状态、动作

$A(s_t,a_t)$——优势函数

E_t——对时间步长的经验期望

优势函数在状态 s_t 下，选择 a_t 所得分数与平均分比较，如果高，则优势函数为正向的，否则为反向的；利用梯度上升法更新 θ，使目标函数 $J(\theta)$ 最大化，实现策略得分最优化的目的。

由于 PG 算法采取的是在线更新策略，每一次参数更新都进行重新采样，导致其学习率不易确定。PPO 算法则将在线更新策略转换为离线更新策略，即采用新旧 Actor 策略的方式，新 Actor 的训练数据可以通过旧的 Actor 得到。使用新旧策略的行动概率比值 $r_t(\theta)$ 表示新策略权重，该比值表示为

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)} \tag{2}$$

式中　θ'——旧策略网络参数

PPO 算法目标函数可表示为

$$J^{\mathrm{CPI}}(\theta')=E_t\left[A(s_t,a_t)\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)}\right]=E_t[A(s_t,a_t)r_t(\theta)] \tag{3}$$

在梯度上升法求解中，为了保证策略的回报期望是单调递增的，PPO 算法限制新旧策略分布不能差别太大[20]，即在同样的状态下，2 个网络得到的动作概率不能相差太远，否则需要进行大量采样才能得到近似结果。在针对策略可能发生大幅度变化的问题上，PPO 算法有 2 种解决方式：一种是采用 KL－penalty 惩罚，通过 KL 散度判断策略波动幅度，当 KL 散度大时，增大这一部分的惩罚，当 KL 散度小时，减少这一部分的惩罚，达到约束策略突变的目的。另一种方法是采用 Clip 进行裁剪处理，即将新旧策略变化的范围限定在一个区间内。实验表明，采用裁剪处理避免了 KL－penalty 参数自适应能力差的缺点，能够在实际连续控制任务中取得更好的效果。因此，本文采用裁剪处理方法进行训练，其目标函数为

$$J^{\mathrm{CLIP}}(\theta')=E_t[\min(A(s_t,a_t)r_t(\theta),\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)A(s_t,a_t))] \tag{4}$$

式中　ε——超参数

$\mathrm{clip}(\cdot)$——裁剪处理函数

式(4)表示将 $r_t(\theta)$ 值限定在$[1-\varepsilon,1+\varepsilon]$范围内。目标函数最终取原始值及裁剪值的最小项，防止新策略网络参数 θ 更新过快。

DPPO 在 PPO 的基础上，增加了多线程处理，如图 2 所示，主线程负责更新 Actor 与 Critic 参数，多个辅线程与环境独立交互，收集数据，通过计算梯度并进行汇总共同更新网络参数。具体表现为在环境中添加多个智能体，同时并行训练，一个智能体能够与其他智能体进行相互通信，进行策略间的相互反馈，极大提高了训练的效率，解决了收敛速度慢等问题。

3　虚拟采摘机器人避障路径规划系统搭建

3.1　ML－Agents 主要功能

ML－Agents 为 Unity 的一个插件，利用外部通

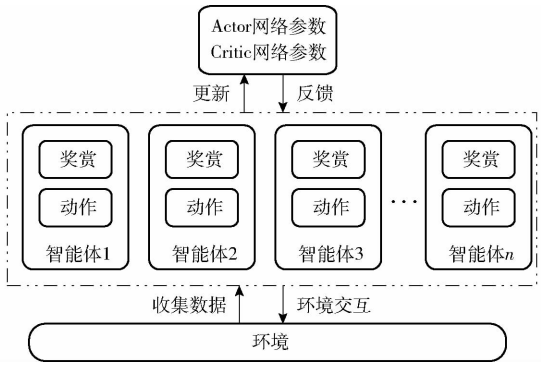


图 2 DPPO 多线程处理框架

Fig. 2 Multi-thread processing framework of DPPO

信器(External Communicator)组件和 Socket 通信机制实现了学习环境与 Python API 之间的通信功能,使用者在 C#编程环境下间接调用 Python API,达到在 Unity 中进行强化学习的目的。其主要结构如图 3 所示,其中,学习环境主要包括 3 部分:

(1)机器代理智能体(Agent):执行实际动作的智能体,可附加于 Unity 的任何物体对象上。主要负责执行所接收的动作,实时感知环境状态,设置并反馈每一步动作奖惩结果。所有 Agent 的决策由决策大脑 Brain 执行;本文视虚拟机器人为 Agent。

(2)决策大脑(Brain):包含了 Agent 的决策逻辑,在训练环境下,根据收集的环境信息以及奖惩信息决定 Agent 在当前环境下的下一步动作;保存并更新动作信息(训练结果),在运行环境下指挥 Agent 的每一步决策。

(3)决策组织(Academy):配置环境参数、运行步数等,收集 Agent 与 Brain 的奖惩信息、决策信息等,通过外部通信器将信息与 Python API 进行交互,反馈信息并指挥观测与决策过程。

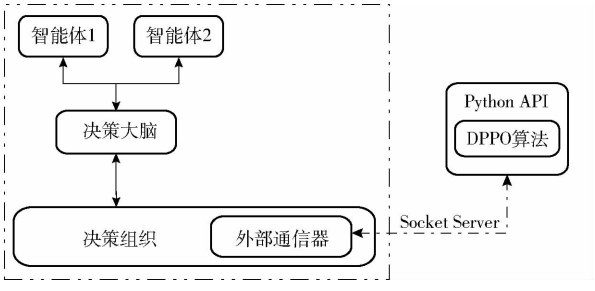


图 3 ML-Agent 组件基本结构

Fig. 3 Basic structure of ML-Agent components

在强化学习训练前,需要定义 3 个实体:动作、观测量(状态)和奖励信号,其中动作表示智能体可采取的动作;观测量用于智能体对环境的感知,收集可供使用的信息;奖励信号表示智能体行为好坏的标量值。ML-Agents 提供了这 3 项实体设置的接口。

3.2 动作及状态初始化设定

虚拟机器人动作根据其机械结构进行设定,本文采用的机器人共有 6 个转动关节轴,如图 4 所示。序号 1~6 分别对应机器人转动关节轴 1~6,主从级别从 1~6 递减,即轴 1 转动可带动所有其他关节轴运动,轴 2 转动可带动转动关节轴 3、4、5、6 运动。其中关节轴 1 以 z 轴为中心,在 xy 平面运动;关节轴 2 以 y 轴为中心,在 xz 平面运动;关节轴 3 以 y 轴为中心,在 xz 平面运动,其他关节轴依次类推。根据关节轴转动特点,初始化 6 个关节轴的动作。

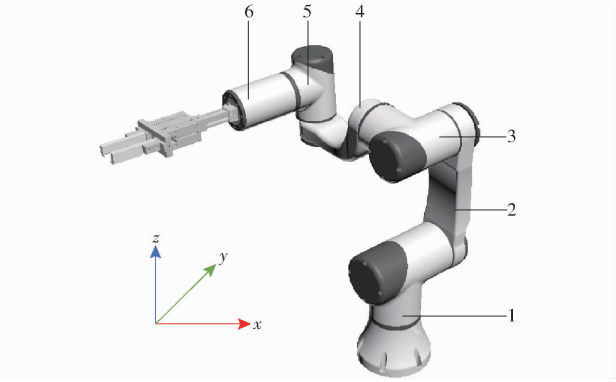


图 4 机器人整体结构示意图

Fig. 4 Schematic of overall structure of robot

环境中的各类实体需要进行合理的设置,才能更有效地进行训练。采摘目标点的位置设定需要综合考虑机器人位置及障碍物的位置,根据机器人有效工作范围及合理采摘区域,将目标采摘点限定在规划区域内随机出现;同时,障碍物的位置也应在规划区域随机出现,为了确保训练过程顺利,采摘点、虚拟机器人、障碍物三者位置应避免发生冲突。在复杂环境下,实现在学习环境下对虚拟机器人路径规划采摘的有效训练,需要观测虚拟机器人位置信息、目标采摘点信息以及障碍物信息等。结合实际采摘环境中能够获取的已知信息:虚拟机器人关节轴转动位置、采摘点三维坐标、障碍物三维坐标,将位置观测集合 O_P 设定为(A_{R_i},P_G,P_{R_iO}),其中,$i=1$、2、3、4、5、6,同时为了让虚拟机器人靠近采摘点并且远离障碍物,还需要输入距离信息,因此设置距离观测集合 O_D(D_G,D_{R_iO}),所有符号定义及所有的观测集合可表示为

$$O=\{A_{R_i},P_G,P_{R_iO},D_G,D_{R_iO}\} \tag{5}$$

式中 O——所有观测集合

R_i——第 i 个机器人转动关节轴

A_{R_i}——虚拟机器人各转动关节轴转向位置

P_G——虚拟机器人末端执行器切割点与采摘目标点的相对位置

P_{R_iO}——虚拟机器人各转动关节轴关键点与

障碍物的相对位置

D_G——虚拟机器人末端执行器切割点到采摘点的距离

D_{R_tO}——虚拟机器人各转动关节轴关键点与障碍物的距离

3.3　导引奖赏机制设计

初始化动作和状态后，智能体可以根据状态随机得出不同的动作策略，但是无法根据状态评价动作的品质。设计导引奖赏函数可以对智能体行为进行评估，提高高分行为的发生概率，降低低分行为的发生概率，进而引导智能体在各种环境状态做出正确的行动。奖赏机制决定了训练结果的效果，合理的奖惩函数设计能够提高训练速度，减少计算机资源消耗，使训练结果可以更快收敛。多数情况下，连续的奖惩信息能够不断让智能体对采取的动作策略得到反馈，比稀疏奖惩信号更加有效。本文在人工势场法思想的基础上构建连续型奖励函数，通过设计引导函数及避障函数引导智能体正确导向目标；同时，设置时间函数引导智能体更快地完成任务。

3.3.1　引导函数设定

采用传统人工势场法设定引力时，其值由物体当前的位置决定；在设置奖励函数时，引力用奖励信号表示，其奖励信号应该由智能体的动作决定，当智能体的动作发生改变，使得末端执行器切割点靠近目标点时，进行奖励，并与接近距离成正比，反之进行惩罚。如在某一状态 s_t 下，虚拟机器人末端执行器切割点 $p_{\text{end}}=(x_{s_t},y_{s_t},z_{s_t})$ 与目标采摘点 $p_{\text{goal}}=(x_0,y_0,z_0)$ 存在一定的距离，用目标距离 D_{s_t} 表示。如果在训练过程中，该距离能够不断减少，说明机器人的动作策略是正确的，该行为应当得到奖励。引导函数设置公式为

$$D_{s_t}=\sqrt{(x_{s_t}-x_0)^2+(y_{s_t}-y_0)^2+(z_{s_t}-z_0)^2} \tag{6}$$

$$D_{\min}=\begin{cases}D_{s_0} & (i=0)\\ \min(D_{s_i},D_{\min}) & (i\neq 0)\end{cases} \tag{7}$$

$$R_{\text{goal}}=\begin{cases}(D_{\min}-D_{s_t})k_1 & (D_{s_t}\neq 0)\\ k_2 & (D_{s_t}=0)\end{cases} \tag{8}$$

式中　$D_{\min}$——一个采摘回合中目标距离最小值

D_{s_0}——采摘回合最初状态 $D_{\min}$ 初始值

R_{goal}——目标距离奖惩值

k_1、k_2——常数

随着智能体进行随机动作，D_{s_i} 发生改变，$D_{\min}$ 根据式(7)条件取两者最小值。当目标距离缩小时，根据缩短距离给予其低奖赏；反之给予其低惩罚；当目标距离为 0 时，表示机器人正确到达目标采摘点，给其高奖赏，并结束本回合。

3.3.2　避障函数设定

采用传统人工势场法设定斥力时，只要物体进入障碍物的势场作用范围就会受到斥力的影响，这些影响有时并无必要，会影响最短路径的规划[21]，如图 5 所示，基于人工势场法思想设置避障函数时，斥力用惩罚信号表示，在该情况下虚拟机器人到达目标点的最短路径为直线，但由于受到障碍物作用范围的影响，其路径发生了偏移。

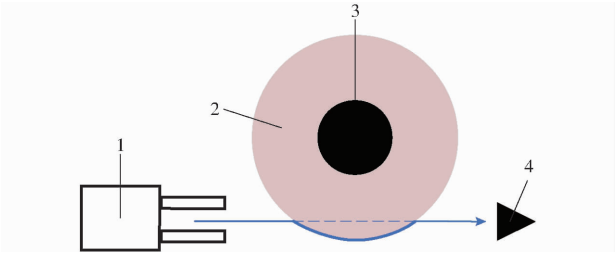


图 5　传统人工势场法局限示意图

Fig. 5　Schematic of limitations of traditional artificial potential field method

1. 机器人　2. 作用范围　3. 障碍物　4. 目标点

为了解决这一问题，本文将虚拟机器人的空间运动分为 3 部分进行考虑：沿 x 轴的左右运动、沿 y 轴的前后运动、沿 z 轴的上下运动。将障碍物的范围惩罚分为上述 3 个运动方向的惩罚，判断虚拟机器人各转动轴以当前位姿沿 x、y、z 轴方向的运动是否可能与障碍物发生碰撞，如果有可能，则进行该方向的惩罚，反之则不进行该方向的惩罚，避免虚拟机器人在进行正确的无碰撞运动时受到不必要的惩罚。

根据采摘机器人物理结构(图 4)，将虚拟机器人每个转动关节轴视为圆柱体进行分析。虚拟机器人的随机运动策略拆分成沿 x、y、z 轴方向的运动，由于各轴 3 个方向的运动碰撞分析具有相似性，本文仅对转动关节轴 6 沿 z 轴的上下运动进行碰撞分析，其他运动情况可以类似得到。假设虚拟机器人转动关节轴 6 当前坐标为 (x_q,y_q,z_q)，障碍物当前的坐标为 (x_o,y_o,z_o)，要判断转动轴 6 上下运动是否会与障碍物发生碰撞，进而确定是否进行惩罚，首先将该转动轴的 z 轴平移到与障碍物 z 轴一致的位置，即令该转动轴坐标为 (x_q,y_q,z_o)，此时两者处于 $x-y$ 平面，如图 6 所示。

图 6 中 o 表示障碍物中点，R 表示障碍物半径；q 表示虚拟机器人转动轴关键点，l_1 表示 q 到其水平圆柱面的距离，即圆柱半径，l_2 表示圆柱长度的 $\frac{1}{2}$；向量 $\boldsymbol{p}_{qo}$ 表示以 q 为起点指向障碍物中点 o 的方向向量，向量 $\boldsymbol{p}_q$ 表示虚拟采摘机器人转动轴当前正

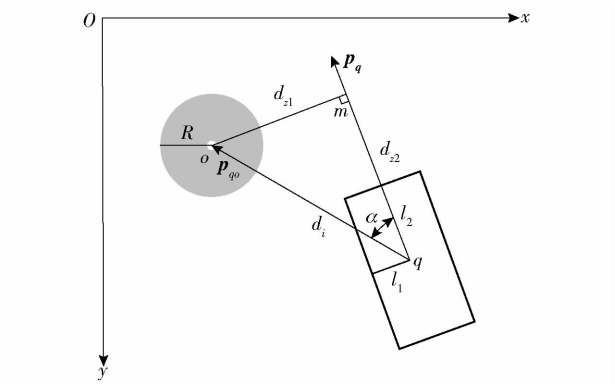


图 6 虚拟机器人转动轴上下运动方向碰撞分析

Fig. 6 Collision analysis of virtual robot's rotating axis up and down direction

方向向量;α 表示向量 $\boldsymbol{p}_{qo}$ 和向量 $\boldsymbol{p}_q$ 的夹角;d_i 表示 q 点与 o 点间距离,可由坐标计算得到,d_{z1} 表示 o 点与 q 点的横向距离,d_{z2} 表示 o 点与 q 点的纵向距离。由图 6 可知,d_{z1} 和 d_{z2} 可由 d_i 计算得到,公式为

$$\begin{cases} d_{z1} = d_i \sin\alpha \\ d_{z2} = d_i \cos\alpha \quad (\alpha > \pi/2) \\ \alpha = \pi - \alpha \end{cases} \tag{9}$$

当 o 点与 q 点的横向距离 $d_{z1} < R + l_1$,且 o 点与 q 点的纵向距离 $d_{z2} < R + l_2$ 时,转动轴 6 进行上下运动时才可能与障碍物发生碰撞,否则不会发生碰撞,因此,设置 z 方向的避障系数 k_z,当符合上述碰撞条件时,k_z 为 1,否则为 0。该转动轴上下运动方向的避障函数表示为

$$R_{\text{obs}-z} = k_{zi} \frac{1}{d_{zi}} \tag{10}$$

式中 d_{zi}——上下距离

类似地,设置左右运动方向、前后运动方向的避障系数及避障函数,避障系数和整体避障函数表示为

$$\begin{cases} k_{xi} = \begin{cases} 1 & (d_{xi} \leqslant R + l_{i1} \,\&\&\, d_{xi} \leqslant R + l_{i2}) \\ 0 & (\text{其他}) \end{cases} \\ k_{yi} = \begin{cases} 1 & (d_{yi} \leqslant R + l_{i1}) \\ 0 & (d_{yi} > R + l_{i1}) \end{cases} \\ k_{zi} = \begin{cases} 1 & (d_{zi} \leqslant R + l_{i1} \,\&\&\, d_{zi} \leqslant R + l_{i2}) \\ 0 & (\text{其他}) \end{cases} \end{cases} \tag{11}$$

$$R_{\text{obs}} = \begin{cases} \sum_{i=1}^{6} k_{xi} \frac{1}{d_{xi}} + k_{yi} \frac{1}{d_{yi}} + k_{zi} \frac{1}{d_{zi}} & (H \cap O_{\text{obs}} = \varnothing) \\ k_{\text{obs}} & (H \cap O_{\text{obs}} \neq \varnothing) \end{cases} \tag{12}$$

式中 d_{xi}——虚拟机器人转动轴 i 关键点的 x 坐标与障碍物 x 坐标的距离,即左右距离

d_{yi}——前后距离

l_{i1}、l_{i2}——机械臂 i 圆柱半径、圆柱长度的 1/2

k_{obs}——碰撞惩罚常数

H、O_{obs}——虚拟机器人空间集合和障碍物空间集合

式(12)仅当虚拟机器人进入障碍物斥力作用范围时生效,其惩罚随着虚拟机器人各个转动轴关键点与障碍物的特点方向距离减少而增大;当机器人与障碍物发生碰撞时,给予高惩罚并结束本回合。

3.3.3 时间函数设定

时间惩罚函数 R_P 根据初始状态下路程进行设置,公式为

$$R_P = \frac{k_t}{D_{s_0}} \tag{13}$$

式中 k_t——时间惩罚常数

总奖励函数的设计为上述 3 个函数的累加和,可表示为

$$R = R_{\text{goal}} + R_{\text{obs}} + R_P \tag{14}$$

3.4 DPPO 算法网络训练流程

DPPO 算法实现需要建立 3 个神经网络:新策略网络 Actor - New、旧策略网络 Actor - Old 以及评价网络 Critic - NN。新旧策略网络负责预测动作策略,评价网络负责评价动作效果,训练流程如图 7 所示。首先环境状态信息(观测量)输入到新策略网络,新策略网络根据状态得到正态分布参数并抽样出一个动作,动作与环境交互产生奖赏值和得到下一步状态 s',将状态、动作与奖赏存储进记忆库中后,状态 s' 再输入到新策略网络中,不断循环,直到记忆库中存储量到达要求。在评价网络中,通过不断获取状态和奖赏值,进行网络参数的反向传播更新,使评价网络对不同情况的评价值越来越接近奖励函数设定值;同时,新旧策略网络根据状态输出策略集合,计算权重后根据式(4)进行新策略网络参数更新,一定步数后,用新策略网络参数来更新旧策略网络参数。不断循环上述过程,直到达到设定的步数限制,完成训练。

4 实验数据分析

实验平台配置为搭载 NVIDIA GTX1050Ti 显卡,i7 - 7700 处理器的 Window 10 系统。为了验证本系统的有效性,从训练结果数据以及实际运行效果两个方面进行分析。

4.1 训练结果数据分析

实验中,训练最大步数设为 20 万步,同时训练智能体设为 10 个,训练环境采用轻量化布局,如图 8 所示。梯度下降初始学习率设为 3×10^{-4},更

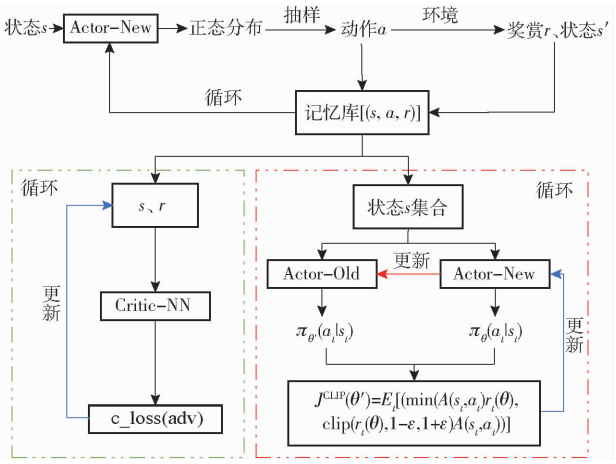


图 7　DPPO 算法网络训练流程图

Fig. 7　Network training flowchart of DPPO algorithm

新策略前收集的经验数量设为 1.024×10^4，每次梯度下降迭代中的经验数设为 1 024，熵正则化强度 β 设为 5×10^{-3}，新旧策略可接受差异范围值 ε（式(4)中的超参数）设为 0.2。在训练过程中，在机器人到达采摘点、碰撞到障碍物、超出预设工作范围或长时间未能到达采摘点 4 种情况下，都立即结束本回合，并计算回合平均奖励。在上述条件下进行农业机器人采摘路径规划训练，训练数据结果由 Tensorboard 获取，并将下载数据导入 Matlab 绘制，如图 9 所示。从图 9a 可以看出，随着步数的增加，代理机器人所获得的奖励逐渐升高，在接近 8 万步时达到收敛状态，此时机器人大部分能在短时间内避开障碍物到达采摘点，说明其选择的动作能够获得更大的奖励值，达到更好的规划效果。随着时间推移，训练算法搜索最优策略的最大步骤逐渐减少，如图 9b 所示，说明其能够在短时间内准确完成路径规划任务。而在图 9c 中，平均策略损失（Policy loss）在正确训练期间也呈下降趋势，进一步表明该方法的有效性。

4.2　仿真实验分析

训练完成后，利用训练结果模型进行路径规划结果测试及对比实验，实验分为 3 部分：有无障碍物影响采摘实验、不同奖励函数下的采摘性能对比实验以及路径规划实际效果实验。

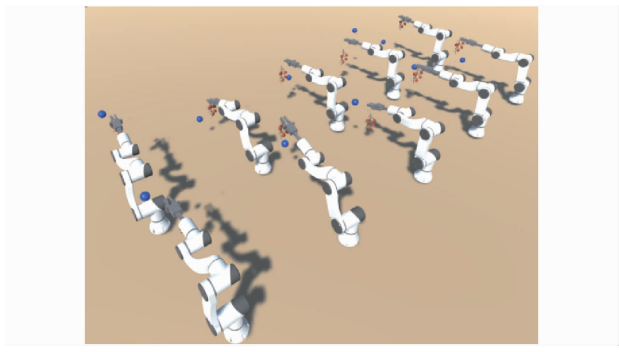
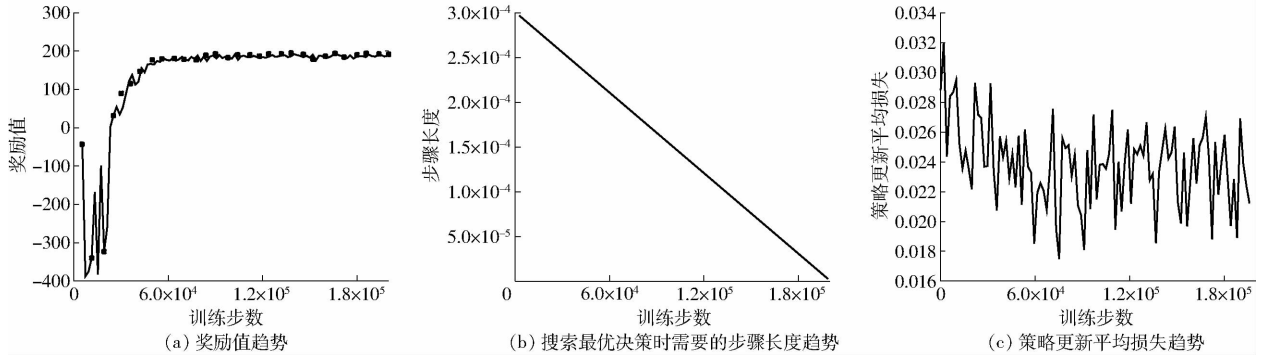


图 8　强化学习训练环境

Fig. 8　Training environment of deep reinforcement learning

4.2.1　有无障碍物影响采摘实验

首先基于本文设计方法进行有无障碍物影响采摘实验。根据障碍物与采摘点的随机生成策略，训练运行结果可分为 2 类：一类是障碍物对机器人路径规划结果有影响的情况，即机器人在从初始位置到达采摘点的过程中，曾进入障碍物警示区域，避障函数根据接近程度做出惩罚反馈；另一类是障碍物对机器人路径规划结果无影响的情况，即机器人在从初始位置到达采摘点的过程中，从未进入障碍物警示区域，此时避障函数不会做出任何的奖惩反馈。2 种情况的成功率如表 1 所示。可以看出，在各自实验次数为 90 次、障碍物无影响时，机器人在训练区域采摘的成功率达到 98.9%；障碍物有影响时，机器人在训练区域采摘的成功率达到 96.7%。结果表明，系统在大多数情况下能够完成避障采摘任务，满足实际采摘需要。其中 3 次发生碰撞中 2 次为障碍物和采摘点距离较小的情况，此时虚拟机器人绕行到达采摘点，避障函数和时间函数产生的惩罚之和可能高于其直接碰撞障碍物产生的惩罚，虚拟机器人执行相对评价得分较高的行为导致其采摘失败。在后续的改进中，可针对特殊情况进行特殊奖惩函数的设置，或根据具体环境评估该点是否值得进行采摘，进一步提高系统的鲁棒性。



(a) 奖励值趋势　(b) 搜索最优决策时需要的步骤长度趋势　(c) 策略更新平均损失趋势

图 9　强化学习训练数据

Fig. 9　Training data of reinforcement learning

表 1 有无障碍物影响下机器人完成采摘任务成功率

Tab. 1 Success rate of robot to complete picking task under influence of obstacles

实验类别	实验数	成功数	成功率/%
障碍物产生影响	90	87	96.7
障碍物不产生影响	90	89	98.9

4.2.2 不同奖励函数下的采摘性能对比实验

为了验证本文方法相较于其他方法的有效性，进行不同奖励函数下的采摘性能对比实验，不同奖励函数分为 4 种：①无障碍惩罚奖励函数，即虚拟机器人在训练过程中与障碍物发生碰撞只单纯进行次数记录，不会进行惩罚，用于判断随机采摘点与随机障碍物对虚拟机器人采摘产生影响的概率。②普通奖励函数，即虚拟机器人在训练过程中与障碍物发生碰撞进行惩罚，而在接近障碍过程中无惩罚。③基于人工势场法的奖励函数，即虚拟机器人在训练过程中，当其靠近障碍物时，给予惩罚，惩罚值与虚拟机器人和障碍物的距离成反比。④改进人工势场法方向惩罚奖励函数，即本文方法。所有奖励函数的目标导引机制相同。随机采摘实验下设置不同奖励函数的采摘性能如表 2 所示。采摘成功数即虚拟机器人成功避开障碍物到达目标采摘点的次数，使用时间指成功完成 200 次采摘任务所需时间。由无障碍惩罚奖励函数采摘成功数可知，随机障碍对采摘任务的影响率大约 28%，相较于普通奖励函数，基于人工势场法的奖励函数和本文方法的避障效果显著，其采摘成功率分别达到 93.5% 和 97.5%，分别比普通奖励函数方法提高了 7 个百分点和 11 个百分点；但使用时间上，基于人工势场法奖励函数由于会产生不必要的惩罚，导致虚拟机器人为了躲避障碍物而选择一条路程较远的路径，采摘时间较长，本文方法能够有效减少这种情况产生的影响，只使用了 124 s，平均耗时 0.64 s/次，相较于基于人工势场法奖励函数方法降低了 0.45 s/次，说明其规划路程相对最短，有利于提高采摘效率，减少资源消耗。

表 2 随机采摘实验下设置不同奖励函数的采摘性能

Tab. 2 Picking performance with different reward functions under random picking experiment

奖励函数类别	采摘成功数（实验数）	采摘成功率/%	使用时间/s
无障碍惩罚奖励函数	144(200)	72.0	141
普通奖励函数	173(200)	86.5	145
基于人工势场法奖励函数	187(200)	93.5	203
本文方法	195(200)	97.5	124

在固定采摘时间 200 s 内，不同奖励函数下虚拟机器人成功采摘个数变化曲线如图 10 所示。结果显示，本文方法在采摘个数以及采摘稳定性上相对最优。由于障碍物具有随机性，每次执行采摘任务的难度是不一致的，普通奖励函数在 0 ~ 70 s 时采摘效率较高，但在后续采摘过程中连续出现障碍物距离采摘点比较接近的情况，导致虚拟机器人采摘失败，进行重新采摘。基于人工势场法的奖励函数虽然保持较高的采摘成功率，但由于虚拟机器人受到范围惩罚的影响，其规划的路径长度大于其他 2 种方法，导致虚拟机器人单个采摘时间较长，在 200 s 内采摘个数最少。本文方法在 200 s 内成功采摘个数最多，在不同采摘任务难度下其采摘效率仅会发生细微波动，说明本文方法在连续变动任务实验中具有更高的适应性和鲁棒性。

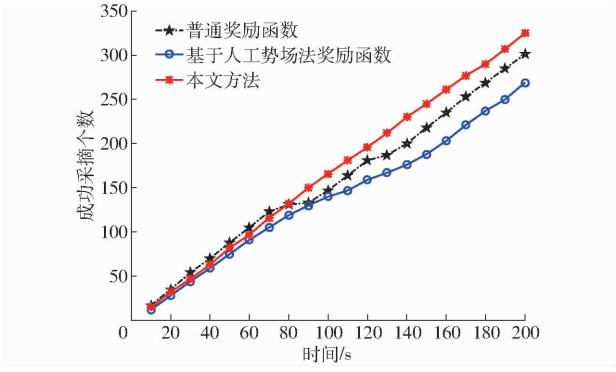


图 10 不同奖励函数下成功采摘个数变化曲线

Fig. 10 Variation curves of number of successful picking under different reward functions

4.2.3 路径规划实际效果

采用本文方法进行路径规划效果如图 11 所示，其中红色圆柱为机器人末端执行器起始点，蓝色球状物为障碍物，红色荔枝串为采摘对象，绿色线条为虚拟机器人末端执行器移动路径。可以看出，虚拟机器人能够根据随机采摘点和障碍物位置，对不同情况进行路径规划，完成无碰撞前提下快速到达采摘点的任务。

为了便于控制与查看结果，搭建仿真运行环境以及仿真交互式控制界面，整体效果如图 12a 所示；可通过输入采摘点起始坐标，障碍物起始坐标进行目的性的路径规划；到达采摘点后，虚拟机器人根据收集管位置将水果放置于收集箱处，并回到初始位姿，继续下一任务，如图 12b 所示；最后可以在保存数据表中，查看系统根据不同采摘点和障碍物所规划的路线，作为实物规划路径的参考。

5 结论

(1)针对水果采摘任务数量多及果实分布随机性强的问题，设计了一套基于深度强化学习的虚拟

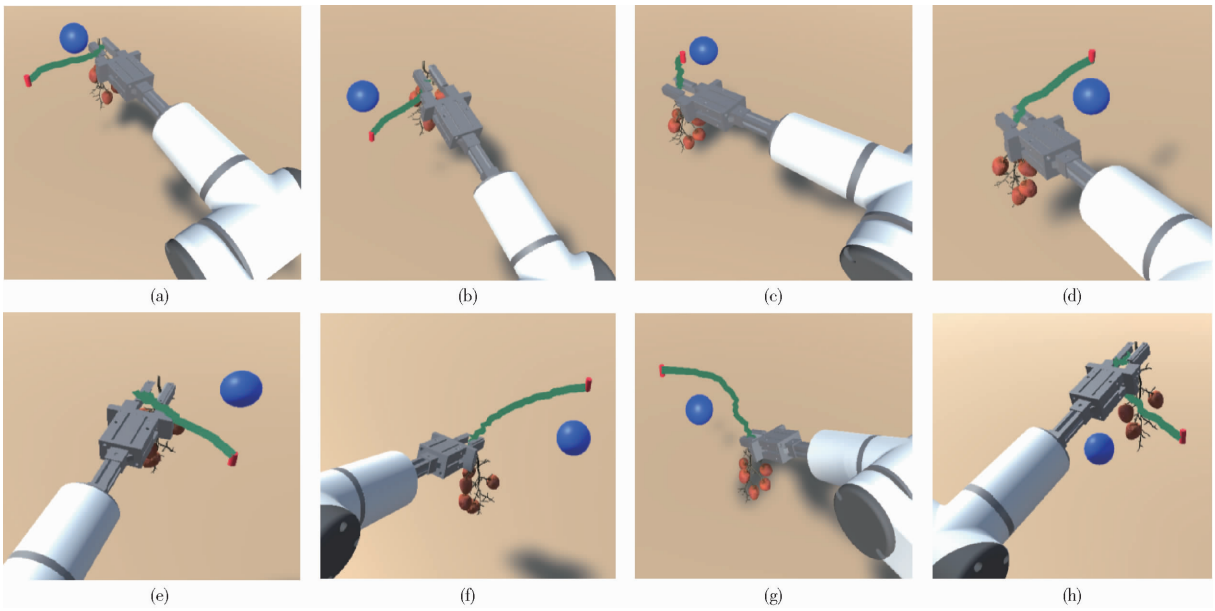
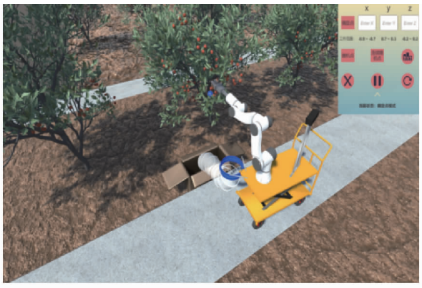
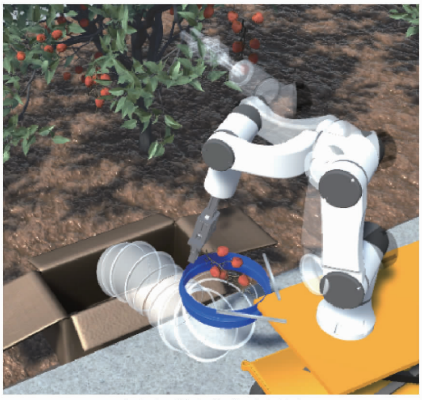


图 11　路径规划结果

Fig. 11　Path planning test results



(a) 仿真运行环境整体效果



(b) 机器人收集水果图

图 12　仿真采摘环境效果图

Fig. 12　Simulation picking environment renderings

机器人采摘路径规划系统，可直接根据采摘点和障碍物位置进行实时高效路径规划，仿真实验结果显示系统采摘成功率达到 96.7% 以上。

(2)提出了一种奖赏函数设置方法。引入人工势场法目标吸引和障碍排斥思想建立奖励函数，提高采摘避障的成功率。针对人工势场法范围斥力影响最短路径规划的问题，提出了一种方向惩罚避障函数及机器人碰撞分析模型，避免机器人正确决策受到不必要的惩罚，提高采摘效率。

(3)为验证本文方法的有效性，设计了不同奖励函数下的采摘性能对比实验。结果显示，本文方法在避障采摘成功率上比普通奖励函数方法提高了 11 个百分点，在采摘轨迹规划平均耗时上比基于人工势场法奖励函数降低了 0.45 s/次，表明本文方法具有更优路径规划效率。

(4)本系统能够快速完成多种情况下的采摘任务路径规划，但在苛刻采摘条件下还有小概率采摘失败的情况，存在一定的不足。后续研究中，将对特殊条件下的采摘情况进行处理，并对不规则形状障碍情况下的路径规划进行探讨，进一步提高系统的鲁棒性。

参　考　文　献

[1]　熊俊涛，郑镇辉，梁嘉恩，等. 基于改进 YOLO v3 网络的夜间环境柑橘识别方法[J/OL]. 农业机械学报，2020，51(4)：199－206.
XIONG Juntao, ZHENG Zhenhui, LIANG Jia'en, et al. Citrus detection method in night environment based on improved YOLO v3 network[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(4): 199－206. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20200423&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2020.04.023. (in Chinese)

[2]　王甲甲，程志强，张伏，等. 果园采摘机械手研究现状综述[J]. 农机化研究，2020，42(5)：258－262.
WANG Jiajia, CHENG Zhiqiang, ZHANG Fu, et al. Design of the divided no-till wheat planter orchard picking manipulator research[J]. Journal of Agricultural Mechanization Research, 2020, 42(5): 258－262. (in Chinese)

[3] 毕松,高峰,陈俊文,等.基于深度卷积神经网络的柑橘目标识别方法[J/OL].农业机械学报,2019,50(5):181 - 186.
BI Song, GAO Feng, CHEN Junwen, et al. Detection method of citrus based on deep convolution neural network[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019,50(5):181 - 186. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20190521&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2019.05.021. (in Chinese)

[4] 滕举元,许洪斌,王毅,等.采摘机器人机械臂运动轨迹规划设计仿真[J].计算机仿真,2017,34(4):362 - 367.
TENG Juyuan, XU Hongbin, WANG Yi, et al. Design and simulation of motion trajectory planning for manipulator of picking robot[J]. Computer Simulation, 2017,34(4):362 - 367. (in Chinese)

[5] KAVRAKI L E, LATOMBE J C, OVERMARS M H. Probabilistic roadmaps for path planning in high-dimensional configuration spaces[C]// IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE,1996:566 - 580.

[6] HART P E, NILSSON N J, RAPHAEL B. A formal basis for the heuristic determination of minimum cost paths in graphs[J]. IEEE Trans. Syst. Sci. and Cybernetics, SSC,1968,4(2):100 - 107.

[7] KHATIB O. Real-time obstacle avoidance for manipulators and mobile robots[J]. International Journal of Robotics Research, 1986,5(1):90 - 98.

[8] 马冀桐,王毅,何宇,等.基于构型空间先验知识引导点的柑橘采摘机械臂运动规划[J].农业工程学报,2019,35(8):100 - 108.
MA Jitong, WANG Yi, HE Yu, et al. Motion planning of citrus harvesting manipulator based on informed guidance point of configuration space[J]. Transactions of the CSAE, 2019,35(8):100 - 108. (in Chinese)

[9] GÓMEZ-BRAVO F, CARBONE G, FORTES J C. Collision free trajectory planning for hybrid manipulators[J]. Mechatronics, 2012,22(6):836 - 851.

[10] CAO Hongxin, SUN Shouqian, ZHANG Kejun, et al. Visualized trajectory planning of flexible redundant robotic arm using a novel hybrid algorithm[J]. Optik,2016,127(20):9974 - 9983.

[11] 薛阳,俞志程,吴海东,等.基于改进人工势场法的双机械臂避障路径规划[J].机械传动,2020,44(3):39 - 45.
XUE Yang, YU Zhicheng, WU Haidong, et al. Obstacle avoidance path planning for double manipulator based on improved artificial potential field method[J]. Journal of Mechanical Transmission, 2020,44(3):39 - 45. (in Chinese)

[12] 王毅,滕举元,张哲,等.六自由度采摘机械臂采摘姿态规划研究[J].机械设计与制造,2019(8):235 - 238,242.
WANG Yi, TENG Juyuan, ZHANG Zhe, et al. Research on the attitude planning of six degrees of freedom picking manipulator[J]. Machinery Design & Manufacture,2019(8):235 - 238,242. (in Chinese)

[13] 王曌,胡立生.基于深度 Q 学习的工业机械臂路径规划方法[J].化工自动化及仪表,2018,45(2):141 - 145,171.
WANG Zhao, HU Lisheng. Industrial manipulator path planning based on deep Q-learning[J]. Control and Instruments in Chemical Industry,2018,45(2):141 - 145,171. (in Chinese)

[14] MILICA L, NĂSTASE A, ANDREI G. Optimal path planning for a new type of 6RSS parallel robot based on virtual displacements expressed through Hermite polynomials[J]. Mechanism and Machine Theory,2018,126: 14 - 33.

[15] MEYES R, TERCAN H, ROGGENDORF S, et al. Motion planning for industrial robots using reinforcement learning[J]. Procedia CIRP,2017,63:107 - 112.

[16] 周祺杰,刘满禄,李新茂,等.一种基于深度强化学习的固体放射性废物抓取方法研究[J/OL].计算机应用研究[2020 - 05 - 29]. https://doi.org/10.19734/j.issn.1001-3695.2019.07.0288.
ZHOU Qijie, LIU Manlu, LI Xinmao, et al. Research on solid radioactive waste grasping method based on deep reinforcement learning[J/OL]. Application Research of Computers. https://doi.org/10.19734/j.issn.1001-3695.2019.07.0288. (in Chinese)

[17] 李跃,邵振洲,赵振东,等.面向轨迹规划的深度强化学习奖励函数设计[J].计算机工程与应用,2020,56(2):226 - 232.
LI Yue, SHAO Zhenzhou, ZHAO Zhendong, et al. Design of reward function in deep reinforcement learning for trajectory planning[J]. Computer Engineering and Application,2020,56(2):226 - 232. (in Chinese)

[18] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv Preprint arXiv:1707.06347, 2017.

[19] HEESS N, SRIRAM S, LEMMON J, et al. Emergence of locomotion behaviours in rich environments[J]. arXiv Preprint arXiv:1707.02286, 2017.

[20] 冯苏柳,姜秀华.基于强化学习的 DASH 自适应码率决策算法研究[J].中国传媒大学学报(自然科学版),2020,27(2):59 - 64,83.
FENG Suliu, JIANG Xiuhua. DASH adaptive bitrate decision algorithm based on reinforcement learning[J]. Journal of Communication University of China(Science and Technology), 2020,27(2):59 - 64,83. (in Chinese)

[21] 熊超,解武杰,董文瀚.基于碰撞锥改进人工势场的无人机避障路径规划[J].计算机工程,2018,44(9):314 - 320.
XIONG Chao, XIE Wujie, DONG Wenhan. Obstacle avoidance path planning for UAV based on artificial potential field improved by collision cone[J]. Computer Engineering,2018,44(9):314 - 320. (in Chinese)