熊春源, 熊俊涛, 杨振刚, 等. 基于深度强化学习的柑橘采摘机械臂路径规划方法 [J]. 华南农业大学学报, 2023, 44(3): 473-483.
XIONG Chunyuan, XIONG Juntao, YANG Zhengang, et al. Path planning method for citrus picking manipulator based on deep reinforcement learning[J]. Journal of South China Agricultural University, 2023, 44(3): 473-483.

基于深度强化学习的柑橘采摘机械臂路径规划方法

熊春源✉, 熊俊涛✉, 杨振刚, 胡文馨
(华南农业大学 数学与信息学院, 广东 广州 510642)

摘要:【目的】为解决非结构化环境下采用深度强化学习进行采摘机械臂路径规划时存在的效率低、采摘路径规划成功率不佳的问题，提出了一种非结构化环境下基于深度强化学习 (Deep reinforcement learning, DRL) 和人工势场的柑橘采摘机械臂的路径规划方法。【方法】首先，通过强化学习方法进行采摘路径规划问题求解，设计了结合人工势场的强化学习方法；其次，引入长短期记忆 (Longshort term memory，LSTM) 结构对 2 种 DRL 算法的 Actor 网络和 Critic 网络进行改进；最后，在 3 种不同的非结构化柑橘果树环境训练 DRL 算法对采摘机械臂进行路径规划。【结果】仿真对比试验表明：结合人工势场的强化学习方法有效提高了采摘机械臂路径规划的成功率；引入 LSTM 结构的方法可使深度确定性策略梯度 (Deep deterministic policy gradient，DDPG) 算法的收敛速度提升 57.25%，路径规划成功率提升 23.00%；使软行为评判 (Soft actor critic，SAC) 算法的收敛速度提升 53.73%，路径规划成功率提升 9.00%；与传统算法 RRT-connect(Rapidly exploring random trees connect) 对比，引入 LSTM 结构的 SAC 算法使规划路径长度缩短了 16.20%，路径规划成功率提升了 9.67%。【结论】所提出的路径规划方法在路径规划长度、路径规划成功率方面存在一定优势，可为解决采摘机器人在非结构化环境下的路径规划问题提供参考。

关键词: 采摘机械臂；柑橘；路径规划；深度强化学习；非结构化环境；LSTM
中图分类号: S666；S233.4 **文献标志码:** A **文章编号:** 1001-411X(2023)03-0473-11

Path planning method for citrus picking manipulator based on deep reinforcement learning

XIONG Chunyuan✉, XIONG Juntao✉, YANG Zhengang, HU Wenxin
(College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China)

Abstract: 【Objective】In order to solve the problems of poor training efficiency and low success rate of picking path planning of manipulator using deep reinforcement learning (DRL), this study proposed a path planning method combined with DRL and artificial potential field for citrus picking manipulator in unstructured environments.【Method】Firstly, the picking path planning problem was solved by the DRL with artificial potential field method. Secondly, the longshort term memory (LSTM) structure was introduced to improve the Actor network and Critic network of two DRL algorithms. Finally, the DRL algorithms were trained in three different unstructured citrus growing environments to perform path planning for picking manipulator.【Result】The comparison of simulation experiments showed that the success rate of path planning was effectively

收稿日期:2022–06–17 **网络首发时间:**2023–03–29 16:00:04
首发网址:https://kns.cnki.net/kcms/detail/44.1110.S.20230329.1424.002.html
作者简介:熊春源，硕士研究生，主要从事采摘机器人研究，E-mail: 20203165015@stu.scau.edu.cn；通信作者：熊俊涛，教授，博士，主要从事智慧农业方向研究，E-mail: xiongjt2340@163.com
基金项目:国家自然科学基金 (32071912)；广州市基础研究计划 (202102080337)

improved by combining DRL with the artificial potential field method, the method with LSTM structure improved the convergence speed of the deep deterministic policy gradient (DDPG) algorithm by 57.25% and the success rate of path planning by 23.00%. Meanwhile, the method improved the convergence speed of the soft actor critic (SAC) algorithm by 53.73% and the path planning success rate by 9.00%. Compared with the traditional algorithm RRT-connect (Rapidly exploring random trees connect), the SAC algorithm with LSTM structure shortened the planned path length by 16.20% and improved the path planning success rate by 9.67%. 【Conclusion】The proposed path planning method has certain advantages for path planning length and path planning success rate, which can provide references for solving path planning problems of picking robots in unstructured environments.

Key words: Picking manipulator; Citrus; Path planning; Deep reinforcement learning; Unstructured environment; LSTM

随着智慧农业的快速发展，柑橘、荔枝、番石榴等季节性水果的采摘过程正往自动化方向转型[1]，在非结构化环境下，果树自然生长形态呈无序性[2]，机械臂采摘任务中存在树枝和非目标果实等障碍物，可能会发生碰撞。碰撞会降低水果质量和产量，并对机械臂和果树都可能造成损伤。因此，为实现水果无损高效的采摘，需要规划出可行的路径来保证机械臂在采摘的过程中尽量避开障碍物。

传统的机械臂路径规划算法有快速扩展随机数法[3-5]、遗传算法[6]、A*算法[7]、栅格法[8]、蚁群算法[9-10] 等。这些算法对于三维环境中的非结构化障碍物和多自由度机械臂来说，过于依赖环境的实时精确建模，且随着机械臂的自由度增加，算法的计算复杂度呈指数级增加，同时依赖参数设定，复杂环境下算法规划的计算时间较长，难以保证稳定性和可靠性。人工势场 (Artificial potential field, APF) 自 1986 年由 Khatib[11] 首次提出以来，已被广泛应用于机械臂的避障。史亚飞等[12] 提出了一种基于改进的人工势场的运动规划方法，使机械臂能够在多障碍物环境下实现动态避障，但是障碍物的斥力势场产生的斥力仅作用于末端执行器，很难保证机械臂的其他部分与障碍物不发生碰撞。为了正确处理障碍物的斥力势场与整个机械臂之间的相互作用，一些研究集中在作用点的选择上。Wang 等[13] 通过定义虚拟力与障碍物上最近点和关节的距离之间的关系来实现避障。上述运动规划方法均在笛卡尔空间中进行，每个规划步骤都需要逆运动学求解。然而，机械臂的逆运动学是一个复杂的多解问题，可能导致关节角路径不连续和奇异点的问题。谢龙等[14] 通过结合机械臂末端与障碍物最近点的排斥力来描述作用于机械臂的排斥力。Zhang 等[15] 提出了一种改进的六自由度串行机器人运动规划的人工势场方法来解决这个问题，该规划直接在关节空间进行，以避免逆运动学求解，虽然这种方法不再需要求解逆运动学，但它在每一步中都需要大量的遍历，并且容易落入局部最小值，鲁棒性不强。

近年来，深度强化学习 (Deep reinforcement learning, DRL) 的快速发展给传统的人工势场法带来了更多的可能。深度强化学习通常基于几个神经网络作为预测操作的策略，通过与环境交互的方式获取数据，对获取的数据进行训练学习控制决策，在计算方面拥有更好的鲁棒性，从而管理一系列复杂的机械臂任务[16-18]。采摘机械臂执行采摘作业任务可以抽象化为一个马尔可夫决策过程。因此对于非结构化环境下采摘机械臂的路径规划问题可以通过深度强化学习进行求解。深度强化学习中的深度确定性策略梯度算法 (Deep deterministic policy gradient，DDPG) 算法和软行为评判 (Soft actor critic，SAC) 算法可以应用于机械臂连续动作控制的领域。但是，采摘机械臂工作的非结构化环境复杂导致了网络的训练效率低。Lu 等[19] 提出一种信息瓶颈理论，尽可能地限制从环境观测传递到状态标识的信息，鼓励神经网络去学习一些高维的特征，提升了算法的训练效果。DeepMind 的研究者在文献 [20] 中提出一种结合了 LSTM 的改进架构 (Contrastive bert for reinforcement learning，CoBERL) 进行深度强化学习，以提高数据的处理效率。Lin 等[21] 提出了一种结合了循环神经网络的 DDPG 算法，应用在番石榴采摘机器人上，获得了较好的鲁棒性。

本文以六自由度采摘机械臂在柑橘果树的非结构化环境中的路径规划问题为研究对象，基于深度强化学习算法进行研究。首先针对深度强化学习

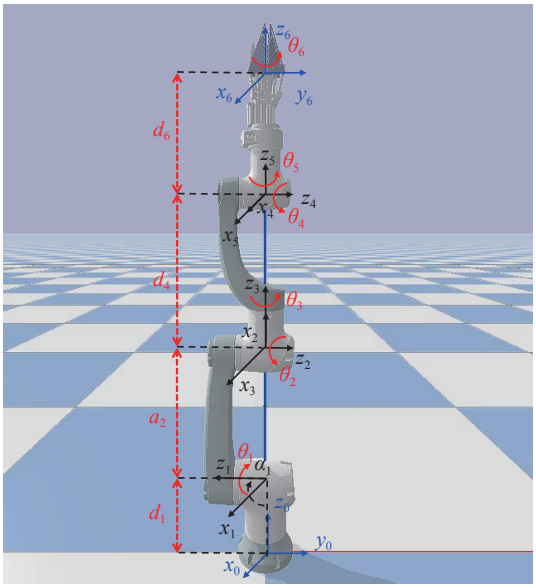
在非结构化环境中采摘路径规划成功率低的问题，提出一种结合人工势场的强化学习方法，让深度强化学习算法获得更高维的信息，通过仿真试验说明该方法可以提升路径规划的成功率。基于结合人工势场的强化学习方法，进一步引入 LSTM 结构，对包含 Actor-critic 网络的 2 种深度强化学习算法 (DDPG 和 SAC) 进行改进，通过仿真试验表明，改进后的算法的收敛速度和采摘路径规划成功率得到了显著提升。最后，与传统方法进行对比，结果表明所提出的基于 SAC 的改进算法在采摘路径规划耗时、路径规划长度和路径规划成功率方面具有优越性。

1　研究方法

1.1　采摘机械臂工作环境

1.1.1　*采摘机械臂*　本研究所采用的机械臂主体为 Hans robot Elfin-5 机械臂，由机械臂、爪型采摘末端执行器组成。机械臂包含 6 个连杆和 6 个转动关节。爪型采摘末端执行器连接到连杆 6 上，作为连杆 6 的一部分。根据标准 Denavit-Hartenberg (D-H) 方法建立机械臂连杆坐标系，如图 1 所示；采摘机械臂的 D-H 参数如表 1 所示。

1.1.2　*碰撞检测和路径规划问题*　在采摘机械收获目标果实的过程中，需要定义采摘点，一般有



θ: 关节角；d: 关节偏移量；a: 连杆长度；α: 连杆扭转角；xyz: 机器人坐标系，其中，蓝色坐标系为机械臂的原点和末端坐标系，黑色坐标系为关节坐标系

θ: Joint angle; d: Joint distance; a: Link length; α: Link twist angle; xyz: Robot coordinate system, in which the blue coordinate system is the origin and end coordinate system of the manipulator, and the black coordinate system is the joint coordinate system

图 1　采摘机械臂

Fig. 1　Picking manipulator

表 1　采摘机械臂 D-H 参数[1)]

Table 1　D-H parameters of picking manipulator

关节编号 Joint No.	θ	d/m	a/m	α/(°)
1	θ_1	0.22	0	90
2	θ_2	0	0.38	180
3	θ_3	0	0	90
4	θ_4	0.42	0	−90
5	θ_5	0	0	90
6	θ_6	0.4	0	0

1)θ: 关节角，d: 关节偏移量，a: 连杆长度，α: 连杆扭转角

1)θ: Joint angle, d: Joint distance, a: Link length, α: Link twist angle

2 种方式：一种对于葡萄、荔枝等串型水果，将采摘点定义在果串的末端以让执行器收获成串水果；另一种对于苹果、柑橘等单果球型水果，定义其质心作为采摘点以让执行器收获单个果实。对于单果球型的柑橘果实的采摘，取果实的质心作为采摘点以指导末端执行器到达采摘位置[22-23]。将采摘任务路径规划过程简化为一个运动规划问题：将末端执行器从初始位置移动到采摘点位置。在检测末端执行器与采摘目标点位置时，若欧氏距离小于等于 0.02 m 则判定为到达。

本文采用几何包络法实现障碍物的近似描述从而进行碰撞检测[5,24-25]。因树叶具有良好的柔性，对机械臂采摘作业的影响很小，可以不计算树叶的碰撞[26]。如图 2 所示，使用圆柱体包络表示枝干障碍，使用球体包络表示非目标果实障碍。对于目标果实，不采取碰撞检测，而使用采摘点作为近似替代。

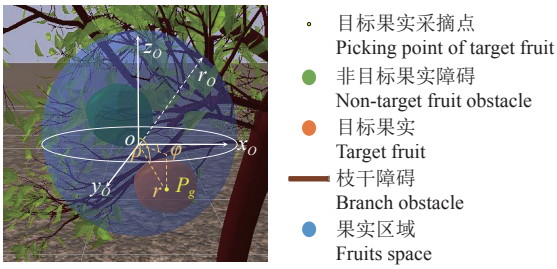
a: 枝干障碍
a: Branch obstacle

b: 果实障碍
b: Fruit obstacle

图 2　碰撞检测模型

Fig. 2　Collision test model

果实的位置应根据柑橘的疏花疏果操作[27]，综合考虑机械臂位置及障碍物的位置以及机械臂有效工作范围进行合理设置。如图 3 所示，将果实区域限定在树枝末梢附近。果实区域 $O(x_O,y_O,z_O,r_O)$ 是以 O 为原点，半径为 r_O 的球体，其中 r_O 取 0.08 m。在果实区域中以 O 为原点建立球极坐标系，非目标果实的质心和目标果实的采摘点以球极坐标 $p_g(r,\varphi,\rho)$ 的形式表示，其中，根据正态分布



P_g: 采摘点, O: 果实区域原点, r_O: 果实区域半径, r: 径向距离, φ: 方位角, ρ: 极角

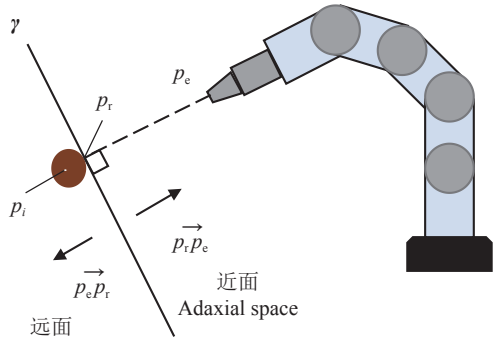
P_g: Picking point, O: Origin point of fruits space, r_O: Radius of fruits space, r: Radial distance, φ: Azimuth angle, ρ: Polar angle

图 3 果实区域

Fig. 3 Fruits space

$r \sim N(0.04, 0.5^2)$的方式生成径向距离 r，使用 Python 中 random 函数随机取值的方式生成方位角 φ 和极角 ρ。

1.1.3 采摘平面定义 为描述果实、枝干与机械臂的空间位置关系[28]，引入采摘平面的概念。如图 4 所示，对枝干的圆形横截面中心点p_i与末端执行器p_e进行连线，线段p_ip_e与枝干的横截面圆形交点为p_r，则于p_r可作一个与p_ip_e垂直的切平面 γ，将该切平面 γ 称作采摘平面。定义向量$\overrightarrow{p_rp_e}$为 γ 到末端执行器方向的空间向量，以 γ 为分界平面，该方向上的空间定义为近面；相反地，以 γ 为分界，向量$\overrightarrow{p_ep_r}$所表示方向的空间为远面。根据位置关系可以推断出：当机械臂需要到达位于远面上的目标时，路径规划中需要对枝干进行避障。



γ: 采摘平面, p_i: 枝干横截面中心点, p_e: 末端执行器, p_r: 线段p_ip_e与横截面的交点, $\overrightarrow{p_rp_e}$: 采摘平面法向量, $\overrightarrow{p_ep_r}$: 采摘平面法向量 (方向与$\overrightarrow{p_rp_e}$相反)

γ: Picking plane, p_i: Center point of branch cross section, p_e: End effector, p_r: Intersection of line segment p_ip_e and cross section, $\overrightarrow{p_rp_e}$: Normal vector of picking plane, $\overrightarrow{p_ep_r}$: Normal vector of picking plane(opposite direction to $\overrightarrow{p_rp_e}$)

图 4 二维采摘平面

Fig. 4 2D picking plane

1.1.4 采摘试验环境搭建 PyBullet 是基于 Bullet 物理引擎的一款仿真器，通过 Python 连接深度强化学习算法。本文选择 PyBullet 搭建仿真采摘环境，进行机械臂运动仿真测试，如图 5 所示。试验中设定了 3 种采摘环境，3 种环境中树枝和果实分布位置不同。在机械臂可达的空间内设置果实区域，引用“1.1.3”中对采摘平面以及近面、远面的定义可以对 3 种环境描述如下：环境 A 中的果实区域均位于采摘平面的近面上，在起始情况下从p_e到采摘点p_g的空间内只存在细小分枝和非目标果实障碍，因此机械臂能够以更小的难度到达采摘点，并且果实分布更加密集。环境 B 中设置了更多的主枝障碍，其中 80% 的果实区域位于远面，机械臂的运动范围中的障碍包括主枝、分枝和非目标果实，相较于环境 A 更复杂，难度更大。环境 C 中，20% 的果实区域在远面上，在机械臂的初始位姿正前方较近的位置存在一根需要避障的主枝，且工作环境相对狭窄。

环境 A 中机械臂和果树的基座标分别为 (0,0,0.75) 和 (0.25,−1.70,0)；环境 B 中机械臂和果树的基座标分别为 (0,−0.30,0.75) 和 (0,1.30,0)；环境 C 中机械臂和果树的基座标分别为 (0,0,0.75) 和 (−1.00,−1.20,0)。

1.2 强化学习设计

在机械臂执行采摘任务的过程中，机械臂的状态与动作之间存在映射关系，即下一个状态取决于当前状态和当前动作，可以用马尔可夫决策过程 (Markov decision process, MDP) 进行建模。此过程主要分为 4 个部分，即状态集合 S、动作集合 A、状态转移概率 P 和奖励回报 R。在运动路径规划过程中机械臂作为智能体，在某一时刻 t，机械臂的姿态以及与周围障碍物的位置关系都会对应

a：环境 A
a: Environment A

b：环境 B
b: Environment B

c：环境 C
c: Environment C

图 5 非结构化环境中的采摘测试

Fig. 5 Picking test in unstructured environment

一个状态s_t，机械臂根据策略$\pi(s_t)$以概率$P(s_{t+1}|s_t)$选择动作a_t，执行动作a_t之后，机械臂的姿态以及与周围障碍物的位置关系将会进入一个新的状态s_{t+1}，机械臂得到奖励R_t。智能体在这样的交互中不断更新自己的策略模型，直至学习到满足任务要求的最优策略模型。智能体与环境之间的交互关系流程如图 6 所示。

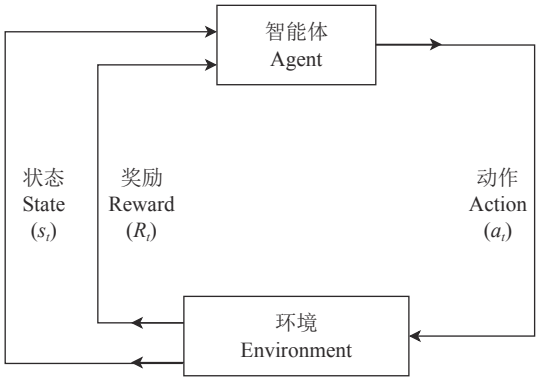


图 6　强化学习流程

Fig. 6　Process of reinforcement learning

1.2.1　基于人工势场法的状态空间　人工势场是一种局部路径规划的方法，通过人工势场辅助机械臂进行避障，其基本思想是：目标点对机械臂产生引力作用，障碍物对机械臂产生斥力作用，引力和斥力的合力方向决定机械臂的运动方向。为了让智能体获取更多的环境信息，本研究结合了人工势场法对状态信息进行设计。下文中，结合人工势场的状态记为s_{APF}，不含人工势场法的基础状态记为s_{basic}。

对于采摘机械臂在非结构化环境下工作可获取的状态信息，可以分类为障碍物位置信息、机械臂姿态信息。文献 [29-32] 对于机械臂在非结构化环境下工作的研究中，基础状态空间的设计主要基于机械臂姿态、目标位置、障碍物位置、机械臂与障碍物之间的距离、末端与障碍物之间的距离。如式 (1) 设计基础状态s_{basic}：

$$\begin{cases} s_{\mathrm{r}} = [\theta_n, p_{\mathrm{e}}, d_g] \\ H_i = [p_i, d_{ni}] \\ s_{\mathrm{basic}} = \{p_g, s_{\mathrm{r}}, H_i\} \end{cases}, \tag{1}$$

式中，s_{r}为机械臂状态，H_i为障碍物i状态。θ_n为机械臂n轴的角度，p_{e}表示机械臂末端位置。d_g表示机械臂末端与目标的距离，p_i表示障碍物位置，d_{ni}表示机械臂n连杆与障碍物i的距离，p_g表示目标位置。

将采摘点、障碍物和机械臂的部分状态用人工势场描述。考虑到机械臂工作环境下的距离特点，构建目标点 g 的引力势场 U_g 和障碍物i对机械臂n连杆的斥力势场 U_{ni}：

$$U_g(d_g) = \frac{1}{2}\xi d_g^2, \tag{2}$$

$$U_{ni}(d_{ni}) = \begin{cases} \dfrac{1}{2}\eta\left(\dfrac{1}{10d_{ni}} - \dfrac{1}{10d_{\mathrm{m}}}\right)^2, & d_{ni} \leqslant d_{\mathrm{m}}, \\ 0, & d_{ni} > d_{\mathrm{m}}, \end{cases} \tag{3}$$

式中，ξ为引力参数，η为斥力参数，d_{m}表示障碍物的影响半径。根据式 (2) 和式 (3)，将人工势场加入机械臂状态 s_{r} 和障碍物i状态H_i中，得到s'_{r}和H'_i，进而构建人工势场状态s_{APF}：

$$\begin{cases} s'_{\mathrm{r}} = [\theta_n, p_{\mathrm{e}}, d_g, U_g] \\ H'_i = [p_i, d_{ni}, U_{ni}] \\ s_{\mathrm{APF}} = \{p_g, s'_{\mathrm{r}}, H'_i\} \end{cases}。 \tag{4}$$

1.2.2　奖励函数与动作空间　在采摘作业任务中，机械臂需要准确避开障碍物，并驱动末端执行器尽量以最短的时间到达采摘目标点，因此需要设置一种合理的奖励策略。一般地，在奖励设定时，设置负奖励的目的是避免某情况的发生，正奖励的目的是希望得到某个结果。首先基于前述的基本状态和参考文献 [30-32] 中描述的避障奖励方法进行奖励设定，具体如公式 (5) 的 basic 方法。

$$\begin{aligned} R_{\mathrm{basic}} &= R_{\mathrm{a}} + R_{\mathrm{c}} + R_{\mathrm{step}}, \\ R_{\mathrm{a}} &= \begin{cases} -(0.5d_g^2), & d_g < \delta, \\ -[\delta(|d_g| - 0.5\delta)], & d_g \geqslant \delta, \end{cases} \\ R_{\mathrm{c}} &= \begin{cases} 0, & d_{ni} > 0, \\ -10, & d_{ni} = 0, \end{cases} \\ R_{\mathrm{step}} &= \frac{\tau_1}{t_{\mathrm{step}}}, \end{aligned} \tag{5}$$

式中，R_{a} 为平滑距离奖励，δ为平滑参数，R_{c} 为避障奖励，R_{step} 为步数奖励，τ_1为步数奖励常数，t_{step}为该回合内的仿真时间步长。

对人工势场方法的奖励 R_{APF} 进行设定，使用公式 (2) 的引力势场函数构建目标引导奖励函数 R_{att}，使用公式 (3) 的斥力势场构建避障奖励函数 R_{rep}。

$$\begin{aligned} R_{\mathrm{att}}(d_g) &= \begin{cases} \tau_2[U_0 - U_g(d_g)], & d_g > 0, \\ \tau_3, & d_g = 0, \end{cases} \\ R_{\mathrm{rep}}(d_{ni}) &= \begin{cases} \tau_4 U_{ni}(d_{ni}), & d_{ni} > 0, \\ \tau_5, & d_{ni} = 0, \end{cases} \end{aligned} \tag{6}$$

式中，U_0为末端位于初始位置的引力势场，末端离采摘点越近，奖励越大。τ_2、τ_3均为正值，当距离d_g=0 时，说明末端已到达采摘点，完成采摘任务，给予一个高值奖励。τ_4、τ_5取负值，触发在斥力范围内工作的情况会给出一定的惩罚。如果在此过程中发

生碰撞的情况，则给予一个高值的负奖励。

综上所述，总奖励为目标引导奖励函数、避障斥力函数、步长奖励函数的累计，如公式 (7) 所示。

$$R_{\mathrm{APF}} = R_{\mathrm{att}} + \sum_{i=0}^{n} R_{\mathrm{rep}}(d_{ni}) + R_{\mathrm{step}}。 \tag{7}$$

基于关节空间定义机械臂的动作空间 A：

$$A = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6], \tag{8}$$

式中，θ_1、θ_2、θ_3、θ_4、θ_5、θ_6分别为机械臂 1、2、3、4、5、6 轴的关节角度。

1.3 Actor-critic 算法

行为−评判算法 (Actor-critic algorithm) 是一种结合策略网络和价值函数的深度强化学习算法，Actor 为行为网络，使用策略梯度算法，负责生成动作，并与环境交互，一般用神经网络实现，输入是当前状态，输出是动作。该网络的训练目标是最大化累计回报的期望。Critic 为评判网络，使用价值函数 Q，负责评价 Actor 的表现，并指导 Actor 后续行为动作，同样可以通过神经网络实现，该网络可以对当前策略函数进行估计，可以评价策略函数 Actor 的好坏。本研究中使用 2 种主流的基于 Actor-critic 算法的深度强化学习算法——DDPG 和 SAC 进行机械臂的路径规划。

1.3.1 *深度确定性策略梯度算法* 深度确定性策略梯度[33](Deep deterministic policy gradient, DDPG) 算法结合深度 Q 网络 (Deep Q network, DQN) 算法的原理引入了神经网络，不仅可用来解决基于高维连续动作空间的强化学习问题，而且使得值函数收敛问题也得到了很好的解决。训练过程中，一共有 4 个神经网络，即 Actor 网络、Target actor 网络、Critic 网络和 Target critic 网络，分别对应：行为网络、目标行为网络、评判网络和目标评判网络。每个学习步骤完成后，将 Actor 网络和 Critic 网络中的参数分别复制到 Target actor 网络和 Target critic 网络中。因此，Actor 网络和 Target actor 网络的架构必须完全相同，Critic 网络和 Target critic 网络的架构必须完全相同。DDPG 算法中ϕ_{a}和ϕ_{c}分别表示 Actor 和 Critic 的神经网络参数。在 DDPG 算法的训练过程中，根据公式 (9) 进行参数更新。

$$\begin{aligned} \phi_{\mathrm{a}}' &= \lambda\phi_{\mathrm{a}} + (1-\lambda)\phi_{\mathrm{a}}', \\ \phi_{\mathrm{c}}' &= \lambda\phi_{\mathrm{c}} + (1-\lambda)\phi_{\mathrm{c}}', \end{aligned} \tag{9}$$

式中，λ为目标平滑因子，它影响目标网络的更新速度和智能体的学习稳定性。时间步状态估计误差 $y(t)$ 由公式 (10) 计算。损失函数 L 由公式 (11) 通过梯度下降法最小化，其中 T 为累计训练步数。

$$y(t) = R(s_t, a_t) + \gamma \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t), \tag{10}$$

$$L = \frac{1}{T}\sum_{t=1}^{T}\left[R(s_t, a_t) + \gamma \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right]^2。 \tag{11}$$

构建的深度强化学习 DDPG 网络如图 7 所示。

1.3.2 *软行为评判算法* 软行为评判 (SAC) 算法[34-35]是一种以离线学习方式更新优化随机策略的算法，与 DDPG 算法相比，该算法训练的是一种随机策略，适用于涉及连续动作的强化学习任务。熵正则化是 SAC 的一个创新点，该策略经过训练，以权衡最大化预期回报和熵之间的关系，熵是衡量策略随机性的一个指标。熵增加时，会发生更多的探索。因此，可能加快学习，还可以避免过早收敛到局部最优。SAC 的目标函数包括奖励项和熵$\mathcal{H}$，同时熵由温度参数 α_{temp} 进行加权，目标函数$J(\pi)$的定义如公式 (12) 所示，熵函数定义如公式 (13) 所示。

$$J(\pi) = \sum_{t=0}^{T} E_{(s_t,a_t)\sim\rho_\pi}\{R(s_t, a_t) + \alpha_{\mathrm{temp}}\mathcal{H}[\pi(\cdot|s_t)]\}, \tag{12}$$

$$\mathcal{H}[\pi(\cdot|s_t)] = -\sum_{a_t}\pi(a_t|s_t)\ln\pi(a_t|s_t), \tag{13}$$

式 (12) 中，α_{temp} 值越大，策略的随机性越强。E代表期望，ρ_π表示策略的状态分布和状态作用边缘。SAC 致力于学习 3 种功能：$\pi_{\phi_{\mathrm{a}}}(s_t, a_t)$与神经网络参数$\phi_{\mathrm{a}}$、软动作值函数$Q_{\phi_{\mathrm{c}}}(s_t, a_t)$与神经网络参数$\phi_{\mathrm{c}}$、软状态值函数$V_\psi$。由于$V_\psi$可以由 Q 和π导出，原则上不需要为状态值单独设置函数逼近器。

在大多数情况下，决定训练成功与否的温度参数 α_{temp} 很难确定。然而，由于奖励的不断变化，使用固定的 α_{temp} 是不合理的，这会使整个训练不稳定，失去 SAC 对超参数依赖性低的优势。因此，我们希望神经网络能够自动调整 α_{temp} 的大小，以确保 α_{temp} 在不同状态下可以调整到不同的值。当某一状态达到最优策略时，α_{temp} 逐渐收敛。我们可以用熵作为约束来求解策略和 α_{temp} 的优化问题。

$$\max E_{\rho_\pi}\left[\sum_{t=0}^{T}\pi R(s_t, a_t)\right], \text{ s.t. } \forall t, \mathcal{H}(\pi_t) \geqslant \mathcal{H}_0, \tag{14}$$

式中，$\mathcal{H}_0$为最小熵阈值。基于上述约束优化问题，可得到如公式 (14) 所示的 α_{temp} 的目标函数。

$$J(\alpha_{\mathrm{temp}}) = E_{a_t\sim\pi_t}[-\alpha_{\mathrm{temp}}\ln\pi_t(a_t|s_t) - \alpha_{\mathrm{temp}}\mathcal{H}_0]。 \tag{15}$$

构建的深度强化学习 SAC 网络如图 8 所示。

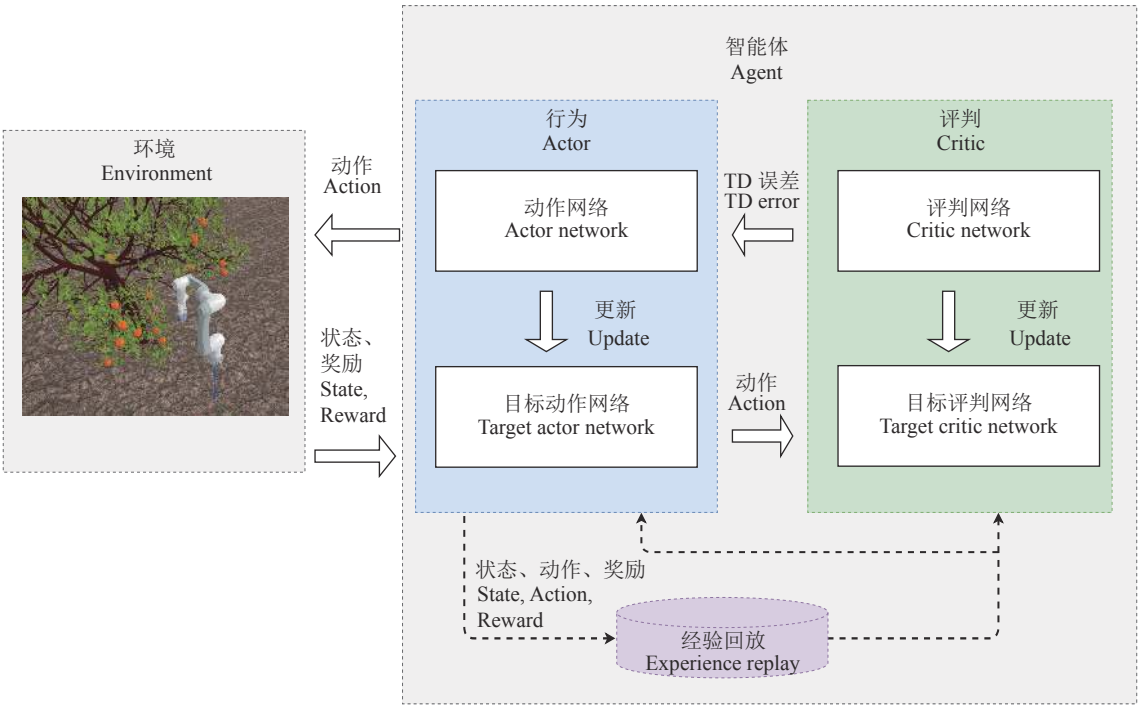


图 7　DDPG 网络结构

Fig. 7　DDPG network structure

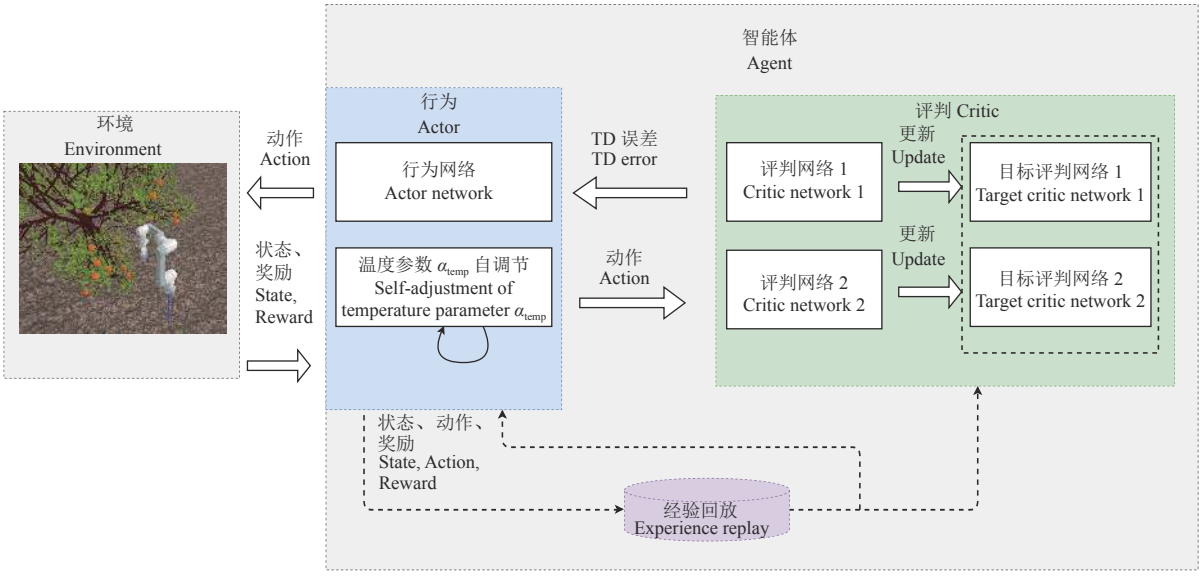


图 8　SAC 网络结构

Fig. 8　SAC network structure

1.4　结合 LSTM 的 Actor-critic 网络改进

由于机械臂执行任务中与环境的关系具有时间序列特征，其路径规划过程不仅与当前时刻的状态有关，还与历史运动信息有关。智能体在充分认知环境的前提下，还需要学习足够多的运动前后关系才能获得更好的鲁棒性。本研究采用 LSTM 结构处理时序关系。LSTM 网络有时间步长，将状态向量中属于不同时刻的状态向量s_t馈入 LSTM 的不同时间步长中。LSTM 最后一个时间步长的输入向量是s_t，而前一个时间步长的输入向量是s_{t-1}，以此类推进行输入。图 9a 为 LSTM-actor 网络结构，LSTM 的输出被输入到一个全连接的深度神经网络 (Deep neural network，DNN) 中，该 DNN 生成一个动作a_t，然后根据 Critic 网络来估计动作的 Q 值。图 9b 为 LSTM-critic 网络结构，它的输入是状态s_t和动作a_t，LSTM 用于提取状态s_t的特征，动作向量a_t也同样通过 LSTM 进行状态提取，二者的输出通过特征归一化后输入到全连接的 DNN。此全连接的 DNN 负责学习从每个状态−动作到 Q 值的映射。

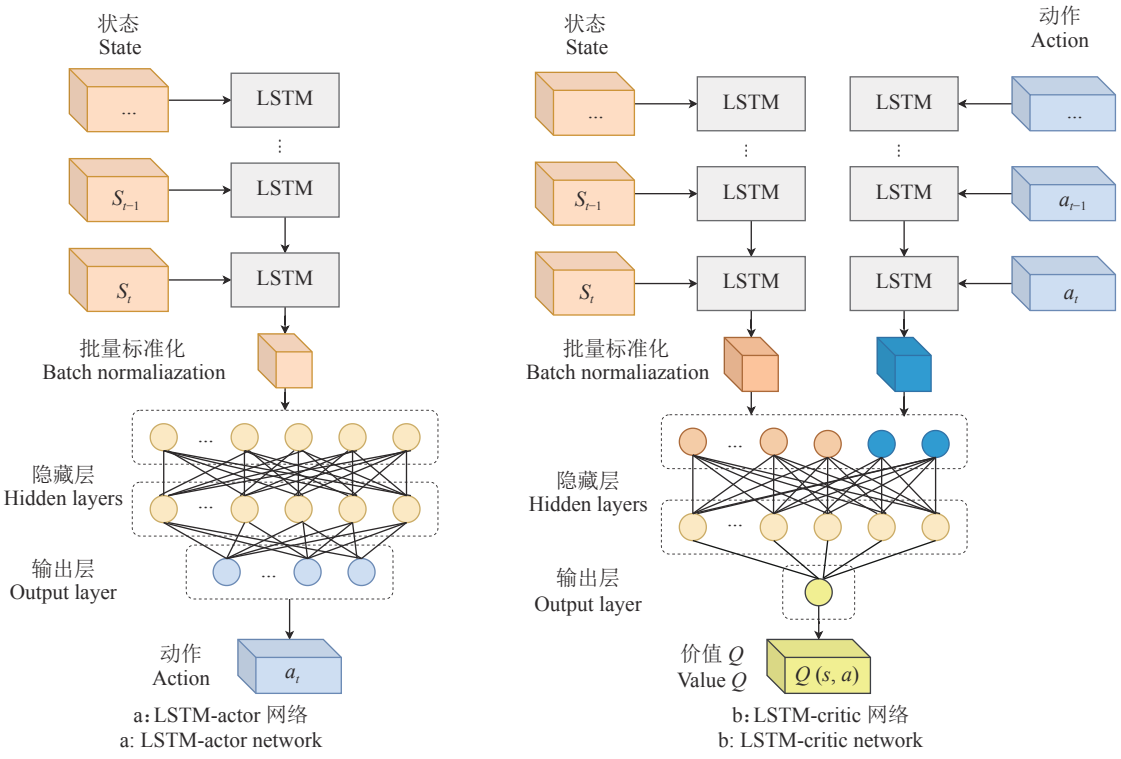


图 9 LSTM-actor 网络与 LSTM-critic 网络结构

Fig. 9 LSTM-actor network and LSTM-critic network structure

2 试验验证与结果分析

为验证方法的有效性，基于 PyBullet 搭建仿真试验平台，在 3 种不同难度的采摘环境下进行柑橘采摘机械臂的路径规划试验。试验主机基于 Windows10 操作系统和 Python3.7 语言，硬件配置为 Intel Core I7 11700K，Nvidia RTX 3090 和 RAM 32 GB。在仿真试验中，人工势场的参数ξ、η、d_m和δ分别设置为−1.0、1.0、0.5 和 0.5；奖励函数的参数τ_1、τ_2、τ_3、τ_4和τ_5分别设置为 10、10、100、−1 和−10。

2.1 人工势场有效性试验

基于 SAC 算法进行试验，将结合人工势场设计的方法记作 APF 方法，对应的状态和奖励分别为s_{APF}和R_{APF}；不含人工势场的方法记作 basic 方法，对应的状态和奖励分别为s_{basic}和R_{basic}，2 种方法均基于前述的动作空间 A 进行试验。在 3 种环境中分别执行基于 APF 训练方法和 basic 训练方法的深度强化学习算法，每个环境中执行 5×10^5 集迭代训练，深度强化学习神经网络的学习率设置为 0.001，学习开始轮数设置为 1 000，batch_size 参数选取 512，强化学习折扣率设置为 0.95，神经网络目标平滑因子λ取 0.005。为直观分析训练效果，在学习过程中定期评估 100 次测试中采摘路径规划的成功率，并以它为评价指标进行分析。

图 10 显示了在环境 A、B 和 C 试验中不同方法训练的 SAC 算法的采摘路径规划的成功率，可以发现，基于 APF 方法训练的采摘路径规划成功率

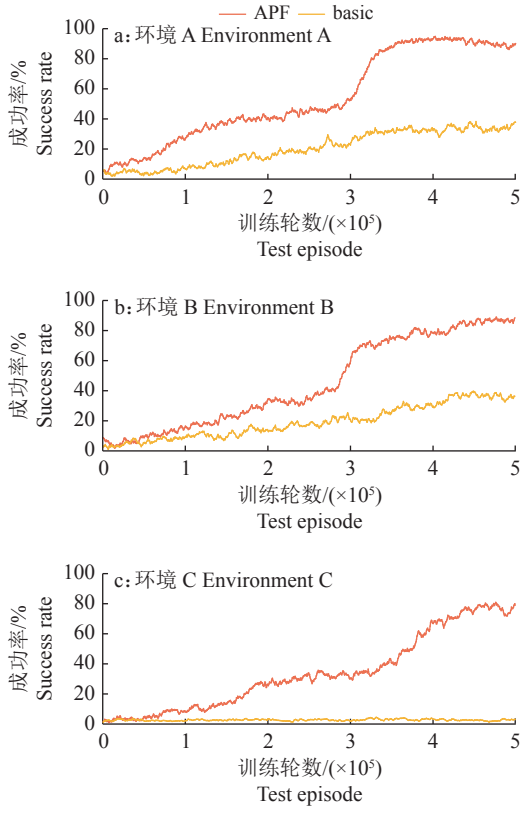


图 10 不同方法与环境下的试验结果

Fig. 10 Experiment results of different methods and environments

高于 basic 方法。

在环境 A 和 B 中，训练曲线的变化趋势大体相似。在训练初期，APF 方法的上升幅度高于 basic。之后在 3×10^5 集左右，APF 方法训练的效果出现大幅提升，远超过 basic 方法。基于 basic 方法训练的算法在环境 A 和环境 B 中的成功率分别为 38% 和 36%，而基于 APF 方法训练的算法的路径规划成功率更高，在环境 A 和环境 B 中分别达到 90% 和 88%。在环境 A 中，APF 方法的成功率出现了回落，随着训练次数的增加，降低的情况更加严重，不利于算法获得更好的性能。这是因为，在避障问题中进行更多的探索并不能解决动作价值高估的问题，使算法陷入了局部最优问题。在环境 C 中，basic 方法无法完成任务，APF 方法的收敛速度略慢于前 2 个环境，在 4×10^5 集左右才开始收敛，并且震荡幅度相较于前 2 个环境更大，最终成功率为 79%。

从对比试验可知，结合人工势场法后，智能体可以获取更多的环境信息，使深度强化学习算法尽可能学习策略以完成采摘点路径规划的任务。

2.2　基于 LSTM 结构的改进算法的有效性试验

试验对基于 LSTM 结构的改进 DDPG 和 SAC 算法的有效性进行分析。将机械臂规划任务强化学习问题看作是一个与时间序列相关的任务，引入 LSTM 结构处理时间序列的状态–动作信息，对 DDPG 和 SAC 算法的 Actor 网络和 Critic 网络进行改进。

试验同样基于环境 A、B 和 C，每个环境中执行 5×10^5 集迭代训练，在学习过程中定期评估 100 次测试的采摘路径规划成功率，并将它作为评价指标进行分析。沿用“2.1”中的深度强化学习算法训练参数，将包含 LSTM 结构的 2 种改进算法分别记作 LSTM-DDPG 和 LSTM-SAC，试验所得不同算法在不同环境下的采摘路径规划成功率情况如图 11 所示。

由图 11 可知，加入 LSTM 结构后，算法在收敛速度和路径规划成功率方面均得到提升。一方面，在 3 种采摘环境的试验中，DDPG 及其改进算法收敛后的波动幅度均大于 SAC 及其改进算法。因此在训练效果稳定性方面，SAC 及其改进算法优于 DDPG 及其改进算法。另一方面，在 3 种环境中，DDPG 算法在训练的后期都出现了回落的情况，而 SAC 算法只在环境 A 中出现这种情况。这说明改进算法可以有效解决动作价值高估的问题，从而避免陷入局部最优的情况。

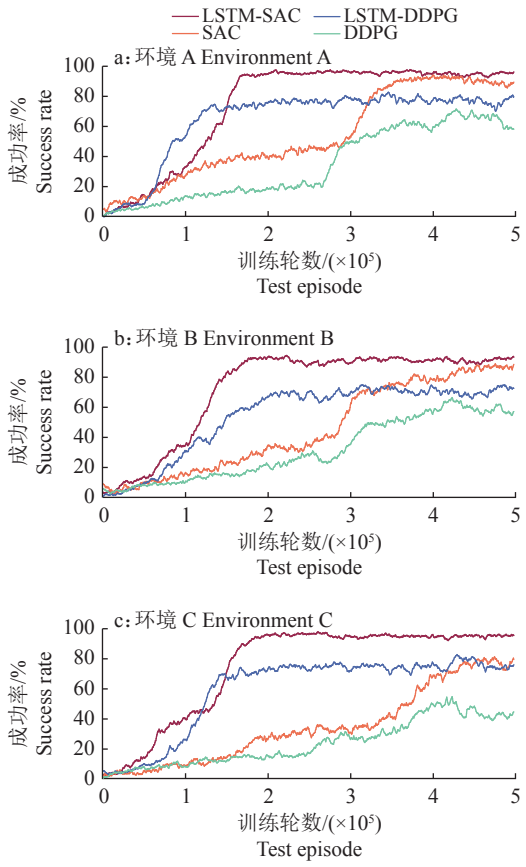


图 11　不同环境下算法的训练结果

Fig. 11　Training results of algorithms in different environments

在环境 A 中，LSTM-SAC、LSTM-DDPG、SAC 和 DDPG 算法的收敛点分别是 1.780×10^5、1.580×10^5、3.760×10^5 和 4.035×10^5，环境 B 中，4 个算法的收敛点分别是 1.765×10^5、2.115×10^5、3.855×10^5 和 4.085×10^5，环境 C 中，4 个算法的收敛点分别是 1.910×10^5、1.410×10^5、4.180×10^5 和 3.780×10^5。在环境 A、B 和 C 中，与原算法对比，LSTM-SAC 算法的采摘路径规划成功率分别提升了 6%、5% 和 16%，综合提升了 9%，算法收敛速度分别提升了 52.66%、54.21% 和 54.31%，综合提升了 53.73%。LSTM-DDPG 算法的采摘路径规划成功率分别提升了 22%、15% 和 32%，综合提升了 23%，算法收敛速度分别提升了 60.84%、48.22% 和 62.70%，综合提升了 57.25%，且算法规避了局部最优情况的发生。DDPG 算法在该任务上存在更大的提升空间，因此本文提出的改进方法在 DDPG 算法上对路径规划成功率的提升幅度大于 SAC，相对地，SAC 在该任务上拥有更好的鲁棒性，提升空间更小。另外，本研究结果表明，SAC 及其改进算法 LSTM-SAC 在该任务中的表现效果分别优于 DDPG 算法及其改进算法 LSTM-DDPG。

2.3 路径规划效果对比

为了进一步量化分析算法对采摘机械臂路径规划的效果，本节在 3 种采摘试验环境下分别测试 100 次 LSTM-SAC、LSTM-DDPG、SAC 和 DDPG 算法对采摘路径进行规划的平均耗时、路径平均长度以及路径规划的成功率。其中，通过计算末端执行器运行的距离来测量路径长度。同时，选择 RRT-connect 算法和 RRT 算法作为传统算法进行对比。为保证试验的统一性，测试不同算法时，每种采摘环境中的果实生成坐标均相同。不同算法在环境 A、B 和 C 中的试验结果如表 2 所示。由表 2 可知，对于路径规划的成功率，环境 A 和 B 中 LSTM-SAC 和 SAC 的路径规划成功率高于传统算法，环境 C 中仅有 LSTM-SAC 的路径规划成功率高于传统算法。对于路径平均长度，环境 A 和 C 中 LSTM-SAC 和 LSTM-DDPG 的路径平均长度短于传统算法，环境 B 中仅有 LSTM-SAC 的路径长度短于传统算法。对于平均规划耗时，深度强化学习算法耗时远小于传统算法，这是因为深度强化学习算法使用了神经网络来预测无碰撞路径，相较于 RRT-connect 和 RRT 需要在构型空间中获取大量样本来探索最优路径的计算量更小。

表 2 不同算法在 3 种环境中的试验结果[1)]

Table 2 Experiment results of different algorithms in three environments

环境 Environment	算法 Algorithm	t/s	l/m	成功率/% Success rate
A	LSTM-SAC	0.03	0.721	96
	LSTM-DDPG	0.03	0.764	80
	SAC	0.05	1.237	90
	DDPG	0.05	1.432	58
	RRT-connect	7.28	0.813	90
	RRT	11.36	0.896	85
B	LSTM-SAC	0.04	1.103	93
	LSTM-DDPG	0.04	1.864	72
	SAC	0.06	2.034	88
	DDPG	0.07	2.339	57
	RRT-connect	9.64	1.337	81
	RRT	17.32	1.431	77
C	LSTM-SAC	0.04	0.793	95
	LSTM-DDPG	0.04	0.937	76
	SAC	0.06	1.361	79
	DDPG	0.07	1.581	44
	RRT-connect	8.72	0.973	84
	RRT	16.56	1.038	81

1) t: 平均规划耗时; l: 路径平均长度

1) t: Average planning time; l: Average path length

综合在 3 种环境中各算法的表现可知，LSTM-SAC 算法在路径平均长度方面相较于 SAC 算法缩短了 43.51%，在路径规划成功率方面提升了 9.00%。相较于 DDPG 算法，LSTM-DDPG 算法使路径平均长度缩短了 33.39%，路径规划成功率提升了 23.00%。另一方面，与传统算法 RRT-connect 相比，LSTM-SAC 算法使路径平均长度缩短了 16.20%，路径规划成功率提升了 9.67%。

3 结论

为了实现采摘机器人野外环境的高效规划，本文提出了一种基于深度强化学习的柑橘采摘机械臂路径规划方法。于搭建的柑橘采摘仿真试验平台下 3 种不同的采摘环境进行仿真试验，结论如下：

1) 本文将人工势场结合到强化学习的状态中，并构建包含人工势场的奖励函数，该方法让深度强化学习算法获得更高维度的信息以学习更具体的行为策略，从而提高深度强化学习算法的训练效率和路径规划的成功率。

2) 本文将 LSTM 结构应用于深度强化学习的 Actor 网络和 Critic 网络中，设计了 LSTM-actor 和 LSTM-critic 网络结构。综合 3 种环境中的表现，相较于 SAC 算法，LSTM-SAC 算法的采摘路径规划成功率的提升幅度达到 9.00%，算法收敛速度快了 53.73%，路径长度缩短了 43.51%；相较于 DDPG 算法，LSTM-DDPG 算法的采摘路径规划成功率的提升幅度达到 23.00%，算法收敛速度快了 57.25%，路径长度缩短了 33.39%。

3) 本文将深度强化学习算法与传统算法在路径规划用时、长度和成功率方面的表现进一步对比，仿真试验结果表明，在路径规划用时方面深度强化学习算法均优于传统算法，充分表明利用神经网络可以在复杂环境中更快地搜索可行路径。其中，本文提出的 LSTM-SAC 算法比 RRT-connect 传统算法在 3 种环境中的路径平均长度综合缩短了 16.20%，路径规划成功率综合提升了 9.67%。

参考文献：

[1] 常有宏, 吕晓兰, 蔺经, 等. 我国果园机械化现状与发展思路[J]. 中国农机化学报, 2013, 34(6): 21-26.

[2] 徐丹琦. 基于 kinect 相机的自然生长状态下果树枝干的三维构建[D]. 桂林: 广西师范大学, 2021.

[3] 崔永杰, 王寅初, 何智, 等. 基于改进 RRT 算法的猕猴桃采摘机器人全局路径规划[J]. 农业机械学报, 2022, 53(6): 151-158.

[4] 张勤, 乐晓亮, 李彬, 等. 基于 CTB-RRT~*的果蔬采摘机械臂运动路径规划[J]. 农业机械学报, 2021, 52(10):

129-136.
[5] 王怀震, 高明, 王建华, 等. 基于改进 RRT~*-Connect 算法的机械臂多场景运动规划[J]. 农业机械学报, 2022, 53(4): 432-440.
[6] 马宇豪, 梁雁冰. 一种基于六次多项式轨迹规划的机械臂避障算法[J]. 西北工业大学学报, 2020, 38(2): 392-400.
[7] 贾庆轩, 陈钢, 孙汉旭, 等. 基于 A*算法的空间机械臂避障路径规划[J]. 机械工程学报, 2010, 46(13): 109-115.
[8] 张敦凤, 赵皓, 徐亮, 等. 基于栅格法的机械臂工作空间解析方法研究[J]. 制造业自动化, 2019, 41(4): 69-70.
[9] 张强, 陈兵奎, 刘小雍, 等. 基于改进势场蚁群算法的移动机器人最优路径规划[J]. 农业机械学报, 2019, 50(5): 23-32.
[10] 刘可, 李可, 宿磊, 等. 基于蚁群算法与参数迁移的机器人三维路径规划方法[J]. 农业机械学报, 2020, 51(1): 29-36.
[11] KHATIB O. Real-time obstacle avoidance for manipulators and mobile robots[M]// COX I J, WILFONG G T. Autonomous robot vehicles. New York: Springer, 1986: 396-404.
[12] 史亚飞, 张力, 刘子煊, 等. 基于速度场的人工势场法机械臂动态避障研究[J]. 机械传动, 2020, 44(4): 38-44.
[13] WANG W, ZHU M, WANG X, et al. An improved artificial potential field method of trajectory planning and obstacle avoidance for redundant manipulators[J]. International Journal of Advanced Robotic Systems, 2018, 15(5): 1729881418799562.
[14] 谢龙, 刘山. 基于改进势场法的机械臂动态避障规划[J]. 控制理论与应用, 2018, 35(9): 1239-1249.
[15] ZHANG N, ZHANG Y, MA C, et al. Path planning of six-DOF serial robots based on improved artificial potential field method[C]// 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau: IEEE, 2017.
[16] GU S, HOLLY E, LILLICRAP T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]//2017 IEEE international conference on robotics and automation (ICRA), Singapore: IEEE, 2017: 3389-3396.
[17] WEN S, CHEN J, WANG S, et al. Path planning of humanoid arm based on deep deterministic policy gradient[C]//2018 IEEE International Conference on Robotics and Biomimetics (ROBIO). Kuala Lumpur: IEEE, 2018: 1755-1760.
[18] KIM M, HAN D, PARK J, et al. Motion planning of robot manipulators for a smoother path using a twin delayed deep deterministic policy gradient with hindsight experience replay[J]. Applied Sciences, 2020, 10(2): 575.
[19] LU X, LEE K, ABBEEL P, et al. Dynamics generalization via information bottleneck in deep reinforcement learning[EB/OL]. arXiv, 2020: 2008.00614 [2020-08-03]. https://arxiv.org/abs/2008.00614.
[20] BANINO A, BADIA A, WALKER J, et al. CoBERL: Contrastive BERT for reinforcement learning[EB/OL]. arXiv, 2021: 2107.05431 [2022-02-22]. https://arxiv.org/abs/2107.05431.
[21] LIN G, ZHU L, LI J, et al. Collision-free path planning for a guava-harvesting robot based on recurrent deep reinforcement learning[J]. Computers and Electronics in Agriculture, 2021, 188: 106350.
[22] 毕松, 张潞. 自然环境下的柑橘采摘点识别方法研究[J]. 计算机仿真, 2021, 38(12): 227-231.
[23] 杨长辉, 刘艳平, 王毅, 等. 自然环境下柑橘采摘机器人识别定位系统研究[J]. 农业机械学报, 2019, 50(12): 14-22.
[24] 尹建军, 武传宇, YANG S, 等. 番茄采摘机器人机械臂避障路径规划[J]. 农业机械学报, 2012, 43(12): 171-175.
[25] CAO X, ZOU X, JIA C, et al. RRT-based path planning for an intelligent litchi-picking manipulator[J]. Computers and Electronics in Agriculture, 2019, 156: 105-118.
[26] 郑嫦娥, 高坡, GAN H, 等. 基于分步迁移策略的苹果采摘机械臂轨迹规划方法[J]. 农业机械学报, 2020, 51(12): 15-23.
[27] 邓钊. 柑橘省力化疏果和促进果实膨大技术研究[D]. 武汉: 华中农业大学, 2018.
[28] 张哲. 柑橘采摘机器人采摘姿态及序列研究[D]. 重庆: 重庆理工大学, 2018.
[29] 熊俊涛, 李中行, 陈淑绵, 等. 基于深度强化学习的虚拟机器人采摘路径避障规划[J]. 农业机械学报, 2020, 51(S2): 1-10.
[30] ZHANG T, ZHANG K, LIN J, et al. Sim2real learning of obstacle avoidance for robotic manipulators in uncertain environments[J]. IEEE Robotics and Automation Letters, 2021, 7(1): 65-72.
[31] XIE J, SHAO Z, LI Y, et al. Deep reinforcement learning with optimized reward functions for robotic trajectory planning[J]. IEEE Access, 2019, 7: 105669-105679.
[32] DUGULEANA M, BARBUCEANU F, TEIRELBAR A, et al. Obstacle avoidance of redundant manipulators using neural networks based reinforcement learning[J]. Robotics and Computer-Integrated Manufacturing, 2012, 28(2): 132-146.
[33] LILLICRAP T, HUNT J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. arXiv, 2015: 1509.02971 [2019-07-05]. https://arxiv.org/abs/1509.02971.
[34] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[EB/OL]. arXiv, 2018: 1801.01290 [2018-08-08]. https://arxiv.org/abs/1801.01290.
[35] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications[EB/OL]. arXiv, 2018: 1812.05905 [2019-01-29]. https://arxiv.org/abs/1812.05905.

【责任编辑　庄　延】