

基于深度强化学习的六足机器人运动规划

傅汇乔¹, 唐开强¹, 邓归洲², 王鑫鹏², 陈春林¹

(1. 南京大学工程管理学院, 江苏 南京 210046; 2. 西南科技大学制造科学与工程学院, 四川 绵阳 621010)

摘要: 六足机器人拥有多个冗余自由度, 适用于复杂的非结构环境。离散环境作为非结构环境的一个苛刻特例, 需要六足机器人具备更加高效可靠的运动策略。以平面随机梅花桩为例, 设定随机起始点与目标区域, 利用深度强化学习算法进行训练, 并得到六足机器人在平面梅花桩环境中的运动策略。为了加快训练进程, 采用具有优先经验重放机制的深度确定性策略梯度算法。最后在真实环境中进行验证, 实验结果表明, 所规划的运动策略能让六足机器人在平面梅花桩环境中高效平稳地从起始点运动到目标区域。为六足机器人在真实离散环境中的精确运动规划奠定了基础。

关键词: 六足机器人; 运动规划; 深度强化学习

中图分类号: TP242.6

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202039

Motion planning for hexapod robot using deep reinforcement learning

FU Huiqiao¹, TANG Kaiqiang¹, DENG Guizhou², WANG Xinpeng², CHEN Chunlin¹

1. School of Management and Engineering, Nanjing University, Nanjing 210046, China

2. School of Manufacturing Science and Engineering, Southwest University of Science and Technology, Mianyang 621010, China

Abstract: Hexapod robot have multiple redundant degrees of freedom and are suitable for complex unstructured environments. Discrete environments, as a harsh special case of unstructured environments, require hexapod robots to have more efficient and reliable motion strategies. A plane random plum-blossom pile environment was taken as an example. A random starting point and a target area were set, and the deep reinforcement learning algorithm was applied to plan a motion strategy for a hexapod robot in the plane plum-blossom pile environment. To speed up the training process, a deep deterministic policy gradient algorithm with a prioritized experience replay mechanism was used. Finally the policy was verified in a real environment. The results show that the planned motion strategy can make the hexapod robot move efficiently and smoothly from a starting point to a target area in a plane plum-blossom pile environment. This work lays the foundation for the precise motion planning of hexapod robots in the real discrete environment.

Key words: hexapod robot, motion planning, deep reinforcement learning

1 引言

相对于传统的轮式与履带式机器人, 足式机器人拥有更强的地形适应能力与越障能力。在多足机

器人中, 六足机器人具有多个冗余自由度和丰富的行走步态, 具备更好的容错性和静态稳定性, 在抢险救灾、物资运输、军事侦察、星球探索等领域具有广泛的应用前景。如何提升六足机器人在非结构

收稿日期: 2020-11-16; 修回日期: 2020-12-04

通信作者: 陈春林, clchen@nju.edu.cn

基金项目: 国家自然科学基金资助项目 (No.71732003, No.62073160); 国家重点研发计划资助项目 (No.2018AAA0101100); 载人航天第四批预研项目 (No.030602)

Foundation Items: The National Natural Science Foundation of China (No.71732003, No.62073160), The National Key Research and Development Program of China (No.2018AAA0101100), The 4th Pre-Research Project of Manned Space (No.030602)

环境乃至离散环境中的运动能力是当前的研究热点之一^[1-4]。

六足机器人的运动主要分为规划和控制两个部分。规划部分在环境模型的基础上,给定机器人的起始状态与目标状态,规划出一条可行的质心与足端运动轨迹;控制部分通过驱动机器人各关节运动来跟随目标轨迹。参考文献[5]聚焦于六足机器人在半结构环境中的运动规划问题,通过 RGB-D 相机建立环境 2.5D 模型,经分割处理后去除环境中不适合作为落足点的禁止区域,最后利用标准的 A* 算法得到机器人在所创建环境中的运动轨迹。参考文献[6]利用灰度腐蚀获得不连续地形边界图像,然后基于前进、旋转和自适应步态设计六足机器人在不连续地形上行走时的步态选择策略,通过图像处理技术提高六足机器人在不连续地形中的行走效率。

传统的 A*、广度优先搜索等算法能够在状态空间维度较低的情况下为机器人快速规划出可行运动轨迹,但在维度较高的状态动作空间中往往会面临维数灾难问题,因此使用此类算法需要将机器人的状态空间以及动作空间进行充分的离散精简,在降低优化指标的前提下得到次优策略。近年来,随着深度强化学习的发展,越来越多的研究人员开始利用深度强化学习算法在高维连续状态动作空间中为游戏 AI^[7]、机器人等智能体生成运动策略^[8-14]。参考文献[15]结合基于模型的运动规划方法和深度强化学习算法,利用可达性评价准则替代物理模拟来构建马尔可夫决策过程,在高维连续状态动作空间中为足式机器人规划运动轨迹,并控制各关节进行轨迹跟随,使得机器人能够在各类非结构环境中稳定运行。参考文献[16]提出一种基于强化学习和模糊奖励的步态生成策略,利用 Q-learning 算法和模糊奖励机制来判断六足机器人的每一步是否为工作空间中的最佳位置,通过迭代更新在不连续地形上规划运动策略。参考文献[17]提出基于迁移强化学习的六足机器人步态生成策略,通过基于规则的强化学习和迁移学习,加快了六足机器人的步态学习速率。

随机梅花桩作为离散环境的典型例子,能够在一定程度上表征任意非结构环境,适合作为验证六足机器人运动策略可靠性的基本环境。本文以平面梅花桩为例,随机设定六足机器人的初始位置与目标区域,利用深度强化学习算法在平面梅花桩环境中为六足机器人规划运动策略。文章首先对六足机

器人在平面梅花桩环境中的正逆运动学、质心和足端的运动空间以及质心静态稳定裕度进行分析求解,建立基本的运动学模型;随后采用具有优先经验重放(prioritized experience replay, PER)机制的深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法^[18-20]为六足机器人在连续动作空间中规划运动策略;最后建立平面随机梅花桩实验环境,并在真实场景中验证所规划策略的可行性。

2 六足机器人运动学建模

六足机器人具有多个冗余自由度以及丰富的行走步态,如三足步态、四足步态、五足步态、自由步态等^[1],由于各时刻都有3条及以上的腿处于支撑相,在多数情况下六足机器人始终处于静态稳定的状态。在众多步态中,三足步态能够在保证静态稳定性的前提下具备较高的运动效率,本文将三足步态作为六足机器人在平面梅花桩环境下运动的基本步态,并忽略其动力学影响。本节首先建立六足机器人运动学模型,然后分别分析六足机器人的足端、质心在固定平面内的运动空间,最后对六足机器人在三足步态下的静态稳定性进行分析。

2.1 单腿正逆运动学建模

六足机器人的单腿结构连杆示意图如图1所示。每条腿包含3个旋转关节,分别是基节轴、大腿轴和小腿轴,分别对应了坐标系0、坐标系1和坐标系2,连杆长度分别为 l_1 、 l_2 和 l_3 ,坐标系3为机器人足端坐标系。本文利用标准 D-H 建模法对单腿进行正逆运动学求解,D-H 参数见表1,其中角标 i 代表连杆编号, α_i 代表关节扭转角, a_i 代表连杆长度, d_i 代表连杆偏移, θ_i 代表关节转角。六足机器人的尺寸参数和各关节转角范围见表2。

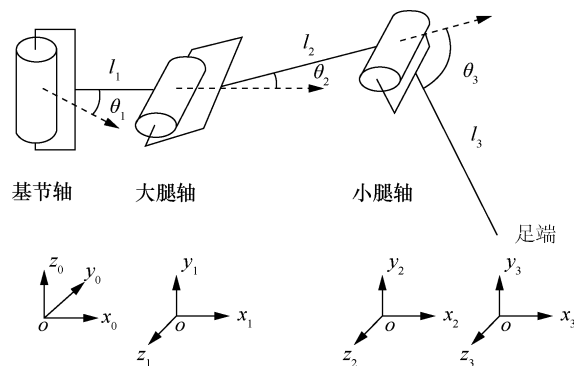


图1 六足机器人的单腿结构连杆示意

表1 D-H 参数

i	α_i	a_i	d_i	θ_i
1	90°	l_1	0	θ_1
2	0	l_2	0	θ_2
3	0	l_3	0	θ_3

表2 六足机器人尺寸参数和各关节转角范围

名称	机身半径 (D)	基节	大腿	小腿
长度/mm	150	50	130	140
转角范围	—	$[-45^\circ, 45^\circ]$	$[0^\circ, 45^\circ]$	$[-135^\circ, -90^\circ]$

利用 D-H 参数对相邻连杆进行旋转平移变换, 得到式 (1) 所示的变换矩阵。

$${}^{i-1}_i T = \begin{bmatrix} \cos \theta_i & -\sin \theta_i \cos \alpha_i & \sin \theta_i \sin \alpha_i & a_i \cos \theta_i \\ \sin \theta_i & \cos \theta_i \cos \alpha_i & -\cos \theta_i \sin \alpha_i & a_i \sin \theta_i \\ 0 & \sin \alpha_i & \cos \alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

利用齐次变换矩阵, 可以得到从坐标系 0 到坐标系 3 的变换矩阵:

$${}^0_3 T = {}^0_1 T {}^1_2 T {}^2_3 T = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

其中,

$$\begin{cases} p_x = l_1 \cos \theta_1 + l_2 \cos \theta_1 \cos \theta_2 + \\ l_3 \cos \theta_1 \cos \theta_2 \cos \theta_3 - l_3 \cos \theta_1 \sin \theta_2 \sin \theta_3 \\ p_y = l_1 \sin \theta_1 + l_2 \sin \theta_1 \cos \theta_2 + \\ l_3 \sin \theta_1 \cos \theta_2 \cos \theta_3 - l_3 \sin \theta_1 \sin \theta_2 \sin \theta_3 \\ p_z = l_2 \sin \theta_2 + l_3 \cos \theta_2 \sin \theta_3 + l_3 \sin \theta_2 \cos \theta_3 \end{cases}$$

对式 (2) 中的 p_x 、 p_y 、 p_z 联立求解, 最终可得到单腿 3 个关节的逆运动学表达式:

$$\begin{cases} \theta_1 = \arctan \frac{p_y}{p_x} \\ \theta_2 = \arcsin \frac{p_z^2 + k^2 + l_2^2 - l_3^2}{2l_2 \sqrt{p_x^2 + p_y^2}} + \arctan \frac{k}{p_z} \\ \theta_3 = \arccos \frac{p_z^2 + k^2 - l_2^2 - l_3^2}{2l_2 l_3} \end{cases} \quad (3)$$

其中, $k = p_x \cos \theta_1 + p_y \sin \theta_1 - l_1$ 。

2.2 足端和质心在平面内的运动空间

假设六足机器人的整体质心等价于机身质心, 其在平面梅花桩环境中运动时, 质心始终处于同一高度, 因此支撑相足端与质心的垂直方向距离保持不变, 在求解足端运动空间时只需考虑足端在固定平面内的运动空间。六足机器人单腿平面投影如图 2 所示, 其中 l_1 、 l_2 和 l_3 为机器人腿部连杆长度, 见表 1, l' 为 A 点到 C 点的直线距离。以六足机器人的一条腿为例, 假设足端与基节的垂直距离为 h , 水平距离为 L , 约束 $|h| < l_3$, 可得:

$$L = l_1 + \sqrt{l'^2 - h^2} = l_1 + \sqrt{(l_2^2 + l_3^2 + 2l_2 l_3 \cos \theta_3) - h^2} \quad (4)$$

在 $\triangle ADC$ 中, 当 h 一定时, 随着 l' 的增大, $\overline{AD}$ 增大; 随着 l' 的减小, $\overline{AD}$ 减小。由表 2 可知, 当小腿轴转角为 -90° 时, $\overline{AD}$ 取最大值, 当小腿轴转角为 -135° 时, $\overline{AD}$ 取最小值。取 $h = 80 \text{ mm}$, 可得 $L_{\min} \approx 116 \text{ mm}$, $L_{\max} \approx 223 \text{ mm}$ 。

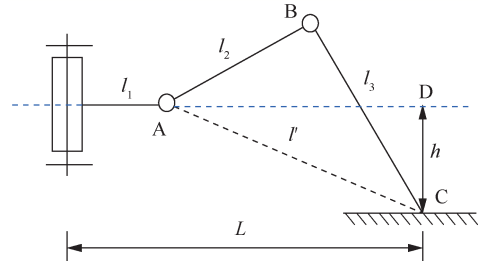


图2 六足机器人单腿平面投影

六足机器人在平面梅花桩环境中运动时, 若要求得环境中的某个梅花桩是否在某条腿的运动范围以内, 需要求得各个足端相对于质心的运动空间, 如图 3 蓝色阴影部分所示。以 1 号腿为例, 其足端相对于机身坐标系 0 的可行域数学约束为:

$$\begin{cases} (y_{13} - y'_1) \geq \tan\left(\frac{\pi}{12}\right)(x_{13} - x'_1) \\ (y_{13} - y'_1) \leq \tan\left(\frac{7\pi}{12}\right)(x_{13} - x'_1) \\ L_{\min}^2 \leq (x_{13} - x'_1)^2 + (y_{13} - y'_1)^2 \leq L_{\max}^2 \end{cases} \quad (5)$$

其中, $x'_1 = D \cos(\pi/3)$, $y'_1 = D \sin(\pi/3)$, 代表 1 号腿基节坐标系 1 相对于机身坐标系 0 的平移变换量, D 为机身半径, (x_{13}, y_{13}) 为 1 号腿足端相对于

质心的坐标。同理,可求得其余5条腿足端相对于机身坐标系0的可行域数学约束。

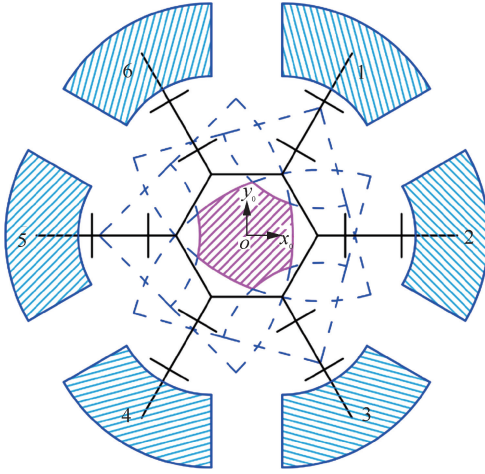


图3 六足机器人足端与质心在平面内的运动空间

由于六足机器人在平面梅花桩环境中运动时机身始终在同一平面内做平移运动,类比于足端运动空间求解方法,在求解质心运动空间时,固定足端坐标,质心围绕足端做平移运动,以质心为坐标原点,已知足端坐标,可求得质心相对于单个足端的运动空间。以1号腿为例,将质心坐标系0作为基坐标系,求解质心的可行域数学约束,如式(6)所示。同理可求质心分别相对于其余5条腿的可行域数学约束。

$$\begin{cases} (y_0 + y'_1 - y_{13}) \leq \tan\left(\frac{\pi}{12}\right)(x_0 + x'_1 - x_{13}) \\ (y_0 + y'_1 - y_{13}) \geq \tan\left(\frac{7\pi}{12}\right)(x_0 + x'_1 - x_{13}) \\ L_{\min}^2 \leq (x_0 + x'_1 - x_{13})^2 + (y_0 + y'_1 - y_{13})^2 \leq L_{\max}^2 \end{cases} \quad (6)$$

在三足步态下,已知当前支撑相3条腿相对于质心的坐标,通过式(6)对所求3组质心运动空间求交集,便可得出当前质心在固定平面内的运动空间,如图3红色阴影部分所示。

2.3 稳定性分析

六足机器人在平面梅花桩环境中运动时始终保持静态稳定,因此可以利用静态稳定裕度法(static stability margin, SSM)^[21]对其稳定性进行评估。以此方法求得的静态稳定裕度表示足式机器人质心在其支撑腿各足端形成的支撑多边形内的投影到支撑多边形各边的最小垂直距离。其计算式如下:

$$SSM = \min(S_1, S_2, S_3, \dots, S_i) \quad (7)$$

其中, S_i 表示机器人质心投影到支撑多边形第 i 条边的垂直距离, i 表示支撑腿数目。

在忽略六足机器人腿部运动对其整体质心的影响后,六足机器人的质心可以始终等价于其机身质心。六足机器人在三足步态下的静态稳定裕度如下:

$$SSM = \min(S_1, S_2, S_3) \quad (8)$$

其中, S_1 、 S_2 、 S_3 分别为六足机器人质心投影到3个支撑相所形成的三角形各边的垂直距离,SSM 的值越大,六足机器人的静态稳定裕度越大,当前时刻的静态稳定性越好。当SSM 的值为负时,说明六足机器人质心投影处于支撑相所构成的三角形范围以外,此时六足机器人处于失稳状态。为了避免六足机器人处于临界稳定状态,即在运动过程中质心投影处于支撑相所构成的三角形边界上,在第4.2节对支撑相所构成的三角形进行了一定比例的缩放。六足机器人在平面梅花桩环境中运动时,只关注其静态稳定裕度是否为正。

3 具有 PER 机制的 DDPG 算法

3.1 强化学习和 DDPG 算法

机器人与环境的交互过程可以被描述为马尔可夫决策过程(Markov decision process, MDP),包含了 $\langle S, A, P, R \rangle$ 四元组,其中 S 表示状态空间, A 表示动作空间, P 表示状态转移概率, R 表示回报函数。在强化学习中,智能体(agent)与其周围环境相互作用,以使累计回报最大化^[22]。对于学习过程的每一步 t , agent 获取当前状态 s_t , 执行动作 a_t , 得到当前回报 r_t , 到达下一状态 s_{t+1} 。强化学习的目标是在每一个状态 s_t 下执行最优动作 a^* , 以使未来的累积折扣回报 R_t 最大化:

$$R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k} \quad (9)$$

其中, $\gamma \in [0, 1]$ 表示折扣因子。在策略 π 下最优状态动作值函数 $Q^*(s_t, a_t)$ 可以表示为:

$$Q^*(s_t, a_t) = \max_{\pi} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] \quad (10)$$

根据贝尔曼方程可得:

$$Q^*(s_t, a_t) = \mathbb{E}[r_t + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})] \quad (11)$$

其中, a_{t+1} 表示在下一状态 s_{t+1} 下所采取的动作。

为了得到最优状态动作值函数,可以利用贝尔曼方程迭代更新状态动作值函数。但这种方法只适

用于低维离散状态的情况，因为贝尔曼方程在迭代更新过程中，需要根据每一组状态动作对逼近状态动作值函数。常用的解决方法是将非线性神经网络作为函数逼近器来估计高维空间中的状态动作值函数。为了基于状态动作对 (s_t, a_t) 估计状态动作值函数，并得到良好的泛化性能，定义状态动作值函数网络为 $Q(s_t, a_t; \theta_1)$ ，其中 θ_1 为网络参数。

此外，为了解决训练数据高度相关导致的强化学习不稳定的问题，基于 Q -learning^[23] 思想，引入经验重放和迭代更新机制打乱数据之间的相关性，通过最小化以下损失函数实现网络更新：

$$L(\theta_1) = \frac{1}{2} (r_t + \gamma Q'(s_{t+1}, a_{t+1}; \theta_1) - Q(s_t, a_t; \theta_1))^2 \quad (12)$$

其中， $Q(s_t, a_t; \theta_1)$ 是在网络参数 θ_1 下状态动作对 (s_t, a_t) 的状态动作值， $Q'(s_{t+1}, a_{t+1}; \theta_1)$ 是在网络参数 θ_1 下，对下一时刻状态动作对 (s_{t+1}, a_{t+1}) 的状态动作值的估计。

然而这种做法难以在高维连续动作空间中实现，为解决此问题，研究人员提出了确定性策略梯度 (deterministic policy gradient, DPG)^[24] 算法，使用一个动作函数 $\mu(s_t)$ 将状态 s_t 映射到一个确定动作 a_t 上，即使用状态动作值函数网络 $Q(s_t, a_t; \theta_1)$ 逼近状态动作值函数 $Q(s_t, a_t)$ ，使用动作网络 $\mu(s_t; \theta_2)$ 逼近动作函数 $\mu(s_t)$ ，通过梯度下降法来更新动作网络：

$$\begin{aligned} \nabla_{\theta_2} Q(s, a; \theta_1) \Big|_{s=s_t, a=\mu(s_t; \theta_2)} = \\ \nabla_a Q(s, a; \theta_1) \Big|_{s=s_t, a=\mu(s_t; \theta_2)} \nabla_{\theta_2} \mu(s; \theta_2) \Big|_{s=s_t} \end{aligned} \quad (13)$$

具有数千个参数的非线性神经网络在训练过程中通常会被证明是不稳定的，常用的解决方案包括引入重放缓冲区、批量标准化^[25]、引入目标神经网络等，这些方法可以应用于大型的非线性神经网络，使得训练更稳定，网络更容易收敛。在 DDPG^[26] 算法中，重放缓冲区用于存储先前遇到的经验，并通过重放随机选择的经验来训练神经网络。批量标准化用于标准化输入数据，其在适当的幅度内调整网络参数，从而稳定学习过程。目标神经网络旨在制作受训网络的副本，并使目标网络参数逐渐逼近受训网络，这样可以大幅提高训练过程的稳定性。

3.2 PER 机制

本文采用 DDPG 算法训练并输出六足机器人在平

面梅花桩环境中的运动策略。考虑到梅花桩分布的随机性以及奖励的稀疏性，标准的 DDPG 算法收敛速度慢，某些情况下难以采样到有效经验甚至无法收敛，本文采用 PER 机制加速算法收敛。

PER 机制使用索引来反映每个经验的重要性。本文使用经验的绝对时序差分 (temporal-difference, TD) 误差作为指标，衡量经验对于 agent 的重要程度。由于 TD 误差反映了 agent 从经验中学到多少，可以选择具有最大 TD 误差的经验来重放，即贪婪 TD 误差优先级。但使用贪婪优先级可能会遇到以下问题。一是 TD 误差的更新仅发生在被重放的经验上，这可以避免耗费时间成本扫描整个重放缓冲区。但在第一次重放中接收到的 TD 误差值很低的经验往往不会被重放多次，甚至在被丢弃之前永远不会被重放，这可能会忽略大量宝贵的可用经验。二是贪婪 TD 误差优先级仅在奖励是确定的情况下才有效。如果奖励是随机的，或者在训练过程中存在一些噪声干扰，则网络的训练效果可能变差。三是贪婪 TD 误差优先级往往只重放经验的一部分，当使用深度神经网络作为函数逼近器时，误差缓慢减小，这可能会降低所选经验的多样性，从而导致网络过拟合。

为了解决这些问题，笔者在选择经验的过程中引入了随机策略，从而在贪婪 TD 误差优先级和纯随机抽样之间进行权衡，该策略被称为随机优先级。在这种情况下，可以重放具有低值 TD 误差的经验，并且重放的概率与其优先级成比例。在对经验采样的过程中，本文利用 PER 机制^[27] 重放缓冲区中的经验，并根据其绝对 TD 误差进行排序。采样经验 j 的概率定义为：

$$P(j) = \frac{D_j^\alpha}{\sum_{i=1}^V D_i^\alpha} \quad (14)$$

其中， $\alpha \in [0, \infty)$ ， $D_j = 1/f(j) > 0$ ， $f(j)$ 代表重放缓冲区经验 j 的序号， V 代表重放缓冲区的大小。如果 $\alpha = 0$ ，则是一个完全随机的采样；若 $\alpha \rightarrow \infty$ ，则是贪婪优先采样。完全随机的采样可以使训练神经网络过程中的经验之间的相关性最小化，而具有高 TD 误差的采样策略将增加相关性，在选择经验的过程中添加适当的随机策略，以降低经验间的相关性。

对于许多强化学习训练过程来说，无偏的策略更新非常重要，PER 机制可能导致更新存在偏差，

影响训练过程的收敛。为了解决这个问题,在权重更新的计算中引入了重要性采样(importance sampling, IS)权重 $W_j^{[28]}$:

$$W_j = \frac{1}{V^\beta \cdot P(j)^\beta \cdot \max_i W_i} \quad (15)$$

其中, $\beta \in [0, 1]$ 。如果 $\beta = 0$, 则对于优先采样过程没有修正; 如果 $\beta = 1$, 则存在较强修正。每个经验分配的 IS 权重除以最大 IS 权重 $\max_i W_i (i = 1, 2, \dots, V)$, 可将梯度更新修正到适当的范围。在实验中, β 设定为从 0.5 线性增加到 1。为了便于对经验进行采样, 重放缓冲区被分为 K 段, 每段都具有相同的被选择概率, 其边界可以用经验 i 被选择的概率确定:

$$P(i) = \frac{D_i^\alpha}{\sum_{j=1}^V D_j^\alpha} = \frac{1}{\sum_{j=1}^V \left(\frac{r_j}{j}\right)^\alpha} \quad (16)$$

其中, $r = f(i)$ 代表经验 i 的序号。在每轮采样过程中, 每段区间仅采样一个经验, 以确保采样经验中始终包括高幅度、中等幅度和低幅度 TD 误差的经验。

完整的具有 PER 机制的 DDPG 算法如算法 1 所示。在选取动作时, 添加了一定的随机噪声来增大训练初期的探索概率。对于算法 1, 一些方法可以进一步提升算法性能, 如: ①优先级队列使用二进制堆来最小化搜索经验的额外时间成本, 在这种情况下, 搜索优先级最大的经验的时间复杂度为 $O(1)$, 更新所选取经验优先级的时间复杂度为 $O(\log N)$; ②为了使训练过程更稳定, 目标网络在每一步仅以一个较低的更新率 η 进行更新; ③新加入的经验被赋予最大的优先级, 与其 TD 误差无关, 从而确保每一个经验至少被重放一次, 进一步提高了所选取经验的多样性。

算法 1 具有 PER 机制的 DDPG 算法

步骤 1: 用随机权重 θ_1 、 θ_2 初始化评估网络 $Q(s_t, a_t; \theta_1)$ 和动作网络 $\mu(s_t; \theta_2)$;

步骤 2: 初始化重放缓冲区 B , 大小设置为 V ;

步骤 3: 设置目标网络 $Q'(s_t, a_t; \theta_1) = Q(s_t, a_t; \theta_1)$, $\mu'(s_t; \theta_2) = \mu(s_t; \theta_2)$;

步骤 4: 设置优先级 $D_1 = 1$, 指数 $\alpha = 0.7$ 、 $\beta = 0.5$, 最小批采样 $K = 32$, 学习率 $\tau = 10^{-4}$, 更新率 $\eta = 10^{-2}$, 最大训练回合数 $H = 600$, 每回合最大步数 $T = 300$;

步骤 5: for $h = 1, H$ do

步骤 6: 初始化一个随机探索过程 $\mathcal{R}$;

步骤 7: 获得初始状态 s_1 ;

步骤 8: for $t = 1, T$ do

步骤 9: 输出动作 $a_t = \mu(s_t; \theta_2) + \mathcal{R}_t$;

步骤 10: 获得回报 r_t 和新的状态 s_{t+1} ;

步骤 11: 存储 (s_t, a_t, r_t, s_{t+1}) 到重放缓冲区 B 中,

令 $D_t = \max_{i < t} D_i$;

步骤 12: if $t > V$ then

步骤 13: for $j = 1, K$ do

步骤 14: 用概率 $P(j) = \frac{D_j^\alpha}{\sum_i D_i^\alpha}$ 采集样本 j ;

步骤 15: 计算重要性采样权重 $W_j = (V^\beta \cdot P(j)^\beta \cdot \max_i W_i)^{-1}$;

步骤 16: 计算 TD 误差 $\delta_j = r_j + \gamma Q'(s_{j+1}, \mu'(s_{j+1})) - Q(s_j, a_j)$;

步骤 17: 根据绝对 TD 误差 $|\delta_j|$ 计算 $f(j)$, 更新 j 的优先级;

步骤 18: 更新 $D_j: D_j = \frac{1}{f(j)}$;

步骤 19: end for

步骤 20: 最小化损失函数来更新评估网络, $L = \frac{1}{2K} \sum_i W_i \delta_i^2$;

步骤 21: 计算策略梯度, $\nabla_{\theta_2} \mu \approx \frac{1}{K} \sum_i \nabla_a Q(s, a; \theta_1)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta_2} \mu(s; \theta_2)|_{s=s_i}$;

步骤 22: 用策略梯度更新动作网络, $\mu(s_t; \theta_2) = \mu(s_{t-1}; \theta_2) + \tau \cdot \nabla_{\theta_2} \mu$;

步骤 23: 更新目标评估网络, $Q'(s_t, a_t; \theta_1) = (1 - \eta) \cdot Q'(s_{t-1}, a_{t-1}; \theta_1) + \eta \cdot Q(s_t, a_t; \theta_1)$;

步骤 24: 更新目标动作网络, $\mu'(s_t; \theta_2) = (1 - \eta) \cdot \mu'(s_{t-1}; \theta_2) + \eta \cdot \mu(s_t; \theta_2)$;

步骤 25: end if

步骤 26: end for

步骤 27: end for

4 六足机器人在平面梅花桩环境中的运动规划

4.1 实验设置

平面梅花桩环境长 1 800 mm、宽 1 200 mm, 根据六足机器人足端运动空间分析可知, 在机身高度 $h = 80$ mm 的条件下, 六足机器人足端距离该腿第一个关节的最小距离 $L_{\min} \approx 116$ mm, 最大距离 $L_{\max} \approx 223$ mm,

最大距离与最小之差 $L_{\max} - L_{\min} = 107 \text{ mm}$ 。由此设置梅花桩在平面梅花桩环境的长宽范围内随机分布，相邻梅花桩在 x 轴与 y 轴方向上距离的均值为 100 mm ，梅花桩总数为 $18 \times 12 = 216$ ，并约束每个梅花桩与其相邻梅花桩之间的距离大于梅花桩的直径，防止两个相邻梅花桩相交，生成的梅花桩坐标以及六足机器人质心在平面梅花桩环境中的坐标默认已知。最终生成的梅花桩地形环境如图4所示。

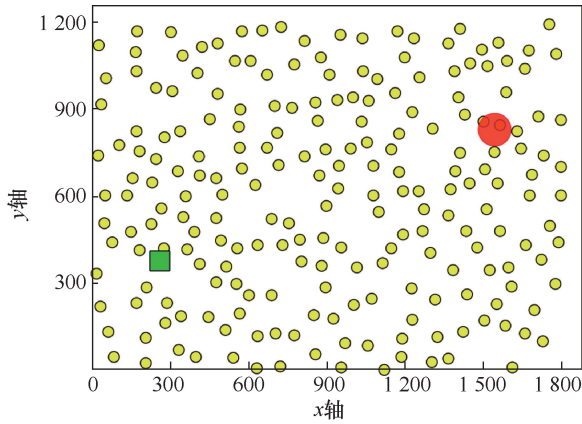


图4 梅花桩地形环境

其中，绿色方块表示六足机器人质心起始点，红圈以内的范围为目标区域，红圈半径为 60 mm ，黄点代表梅花桩。实验中随机设定绿色起始点与红色目标区域，并限制起始点与目标区域中心距离大于 600 mm ，六足机器人的任务是质心从起始点出发，在保证机身稳定的前提下，用最短的步态周期使质心到达红圈区域内。

4.2 运动策略生成

六足机器人在平面梅花桩环境中运动时采用三足步态行进，其机身高度 $h=80 \text{ mm}$ 。将六足机器人在平面梅花桩环境中的运动建模为有限马尔可夫决策过程，其状态空间 S 包含支撑相坐标、质心坐标以及目标区域中心点坐标，动作空间 A 包含六足机器人质心在平面内的运动方向。机器人质心全局坐标与各梅花桩的全局坐标默认为已知量，根据六足机器人摆动相与质心在平面内的运动空间范围，可以分别求出每一组摆动相可以落足的梅花桩与质心在某一方向上的最大运动距离。六足机器人在起始状态下，首先根据其质心坐标，由1、3、5号腿作为摆动相选取目标落足点，随后机器人在每一个运动周期内均完成两步操作：第一步，根据支撑相坐标与质心运动方向，计算六足机器人质心在稳定裕

度范围内能够运动的最大距离，并移动质心到目标点；第二步，根据新的质心坐标，选取摆动相目标落足点。以此循环，直到机器人质心到达目标区域内。支撑相三角形缩放示意如图5所示。

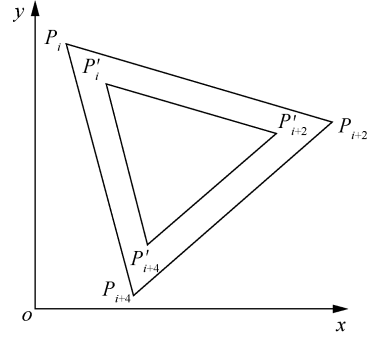


图5 支撑相三角形缩放示意

如图5所示，在每个运动周期内，为了保证六足机器人具备充分的稳定裕度，避免处于临界稳定状态，以支撑相坐标点 $P_i(x_i, y_i)$ 、 $P_{i+2}(x_{i+2}, y_{i+2})$ 、 $P_{i+4}(x_{i+4}, y_{i+4})$ 所构成的三角形 $PP_{i+2}P_{i+4}$ 的重心 P_c 为中心，将其按缩放比 $s=0.8$ 缩放为三角形 $P'_iP'_{i+2}P'_{i+4}$ ，其中：

$$P_c = (x_c, y_c) = \left(\frac{x_i + x_{i+2} + x_{i+4}}{3}, \frac{y_i + y_{i+2} + y_{i+4}}{3} \right) \quad (17)$$

$$P'_i = (sx_i + (1-s)x_c, sy_i + (1-s)y_c) \quad (18)$$

P'_{i+2} 、 P'_{i+4} 与 P'_i 的求法相同。

在已知质心运动方向的前提下，根据六足机器人质心运动空间，可以求得六足机器人质心在稳定裕度范围内能够运动的最大距离，并移动质心到目标点。随后，摆动相选择落足点 P_i 、 P_{i+2} 、 P_{i+4} ，每条腿可落足的梅花桩个数分别为 N_i 、 N_{i+2} 、 N_{i+4} ，3条腿足端全部可能的落足点组合共有 $N_i \times N_{i+2} \times N_{i+4}$ 种。对于每一组落足点所构成的如图5所示的三角形，为了避免六足机器人质心运动到三角形的边界处，导致机器人处于临界稳定状态，以三角形重心为中心将其按缩放比 $s=0.8$ 缩放为三角形 $P'_iP'_{i+2}P'_{i+4}$ ，如式(18)所示。接着判断六足机器人质心坐标是否在三角形 $P'_iP'_{i+2}P'_{i+4}$ 内，若是，则保留 $P'_iP'_{i+2}P'_{i+4}$ 作为备选落足点，否则舍弃。为了使每一条腿在作为支撑相时有较大的运动空间，在所有备用落足点中，对每一组备用落足点求每条腿落足点 P_i 和 P_{mid} 的距离 $D_{P_i-P_{\text{mid}}}$ ，其中：

$$P_{\text{mid}} = L_{\min} + \frac{L_{\max} - L_{\min}}{2} \quad (19)$$

最终选取 $D_{p_i-p_{mid}}$ 值最小的一组落足点作为当前摆动相的目标落足点。

六足机器人在平面梅花桩环境中运动时,采取的动作主要包括移动质心和选取摆动相落足点,其中移动质心时需要明确质心运动方向,进而计算该方向上稳定裕度范围内质心能够运动的最远距离。深度强化学习算法的任务是:输入各支撑相坐标、起始点坐标以及目标区域中心点坐标,输出质心运动方向,使得质心在稳定裕度范围内用最优的策略到达目标区域。

使用具有 PER 机制的 DDPG 算法训练六足机器人在平面梅花桩环境中的策略生成流程如图 6 所示。首先初始化算法参数,在每一个训练回合开始时,复位六足机器人质心到随机起始点,设定一个随机目标区域,并根据质心起始坐标点选取 1、3、5 号腿的落足点,完成机器人的复位。然后对于每一个训练回合内的每一步,由动作网络输出一个带有随机噪声的动作 a ($a \in [-1,1]$),六足机器人质心在平面内的运动方向 $\theta = a \cdot \pi$,计算质心在 θ 方向上能够运动的最远距离,移动质心到目标点。判断质心点坐标是否超出环境边界:如果超出边界,则移动质心到上一个坐标,给予-10 惩罚值,重新输出动作 a ;如果没有超出边界,则根据当前质心坐标选取摆动相落足点。如果不存在合适落足点,则返回质心到上一个坐标,给予-10 惩罚值,重新输出带有随机噪声的动作 a 。如果存在合适落足点,则令摆动相的 3 个足端移动到所选落足点,并判断

是否结束当前回合:若机器人质心坐标到达目标区域范围内,则给予 100 奖励值,并结束当前回合,若该回合内总步数大于设置的最大步数,也结束当前回合。若当前回合未结束,则将新的落足点坐标输入动作网络,并输出新的动作 a ,以此循环。对于回合内的每一步,给予-1 的惩罚值。当训练回合数等于预设的最大回合数时,结束训练。

图 7 为六足机器人在仿真平面梅花桩环境中的运动过程,其中绿色方块代表随机起始坐标,红色圆圈代表随机目标区域,蓝色线框代表机身,红色点代表六足机器人在运动过程中走过的梅花桩。图 7(a) 为开始训练时六足机器人的运动过程,随机生成的起始点坐标为(1 092, 900),目标区域中心坐标为(570, 407)。图 7(b) 为算法收敛时六足机器人的运动过程,随机生成的起始点坐标为(370, 372),目标区域中心坐标为(1 280, 761)。训练初期,六足机器人处于随机探索阶段,随着训练的进行,探索率降低,算法逐渐收敛,最终在给定起始点与目标区域的前提下,六足机器人能够在最少的步态周期内到达目标区域。

图 8 展示了采用具有 PER 机制的 DDPG 算法与采用标准 DDPG 算法平均回报曲线的收敛对比,其中深色实线为滑动平均滤波后的数据,浅色实线为原始未滤波的数据,平均回报的计算方式为回合总回报除以回合总步数。由于梅花桩分布的随机性以及奖励的稀疏性,使用标准的 DDPG 算法收敛速度慢,某些情况下难以采样到

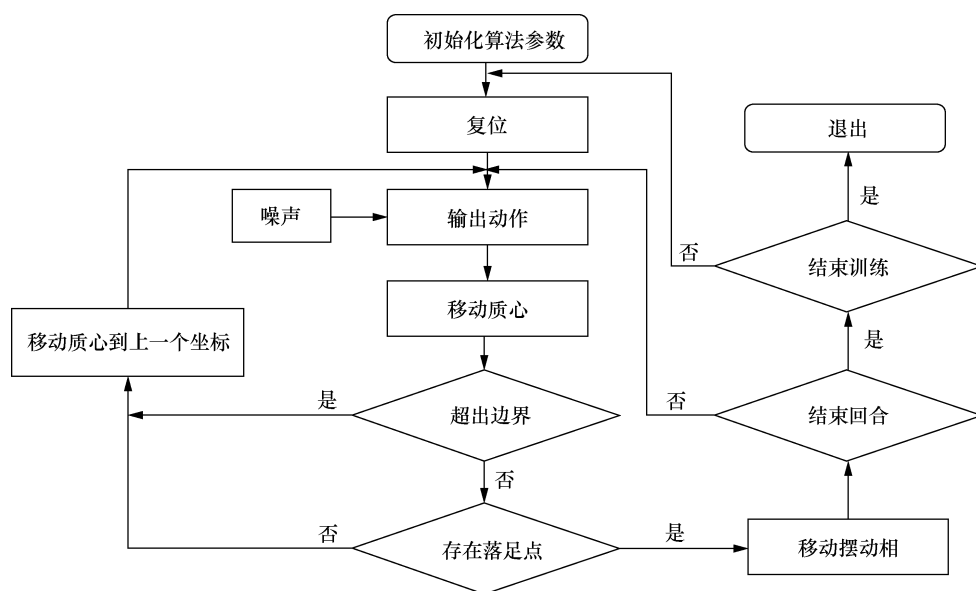


图 6 六足机器人在平面梅花桩环境中的策略生成流程

有效经验甚至无法收敛。从图 8 可以看出, 具有 PER 机制的 DDPG 算法在 150 回合左右收敛, 单个线程上用时约 3 min, 收敛速度明显快于标准 DDPG 算法。

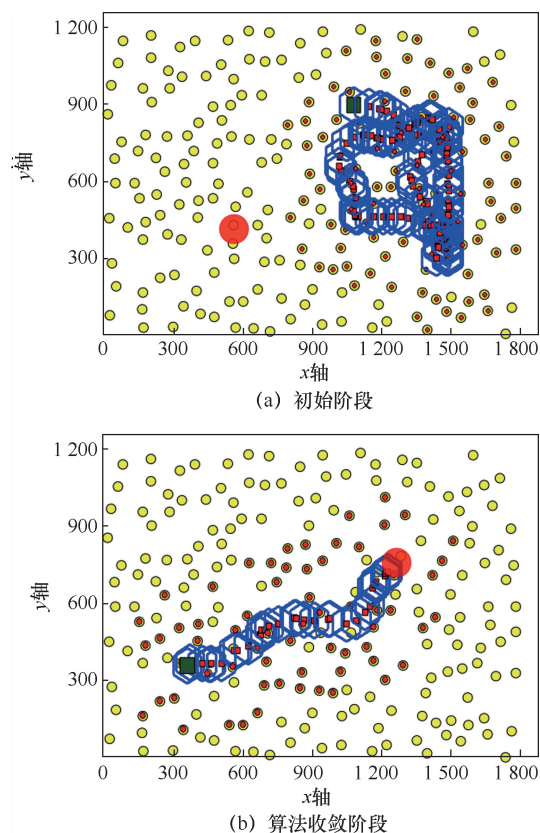


图7 六足机器人在仿真平面梅花桩环境中的运动过程

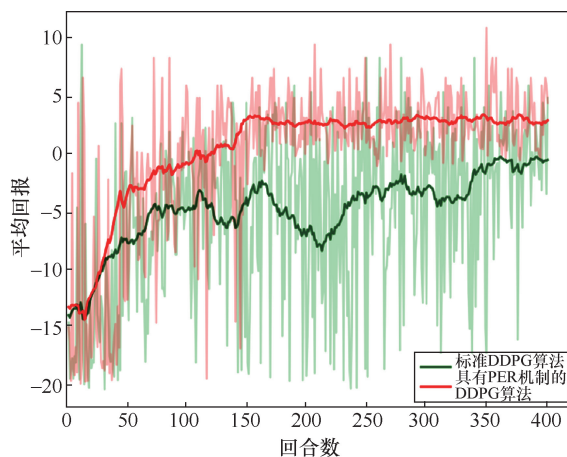


图8 平均回报收敛曲线

4.3 真实场景中的策略验证

为了进一步验证平面梅花桩环境下运动策略的正确性, 根据仿真环境与仿真样机, 等比例构建实际环境模型与实际样机模型进行策略验证。如图9所示,

其中绿色方块代表质心起始点, 红色圆圈代表目标区域, 半径为 60 mm, 黄色点为随机生成的半径为 30 mm 的梅花桩。

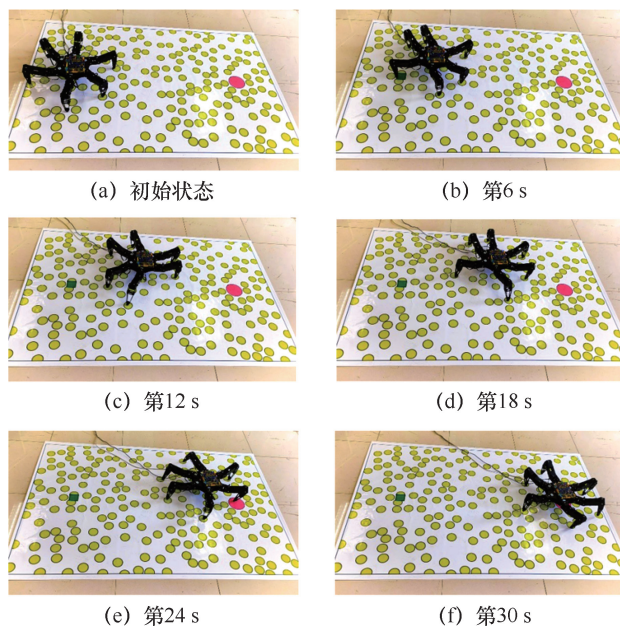


图9 六足机器人样机在实际平面梅花桩环境中的运动过程

根据仿真环境所得的六足机器人在平面梅花桩环境中的运动策略, 可以在任意给定六足机器人起始坐标与目标区域的前提下规划出足端与质心的运动轨迹。六足机器人在平面梅花桩环境中采用三足步态行走, 每一个步态周期内, 3 个支撑相移动质心到目标点, 同时 3 个摆动相分别运动到目标梅花桩。根据第 2.1 节所求得逆运动学计算式, 在已知六足机器人足端离散运动轨迹的前提下, 通过逆运动学计算式可以求得在每一个离散点处的各关节对应转角, 由此可以控制实际样机始终跟随目标轨迹运动。图 9 (a) ~ 图 9 (f) 展示了六足机器人从起始坐标 (300, 600) 运动到中心点坐标为 (1 500, 600) 的圆形区域的运动过程。整个过程用时约 30 s, 共经历 32 个步态周期。

图 10 展示了六足机器人运动过程中的各腿基节轴转角变化曲线, 由于运动规划始终在表 2 所设置的关节转角范围限制下进行, 因此基节轴转角始终处于 $\pm 45^\circ$ 之间, 进一步验证了所生成运动策略的正确性。

5 结束语

本文利用深度强化学习算法生成六足机器人在离散环境下的运动策略。针对六足机器人在平面

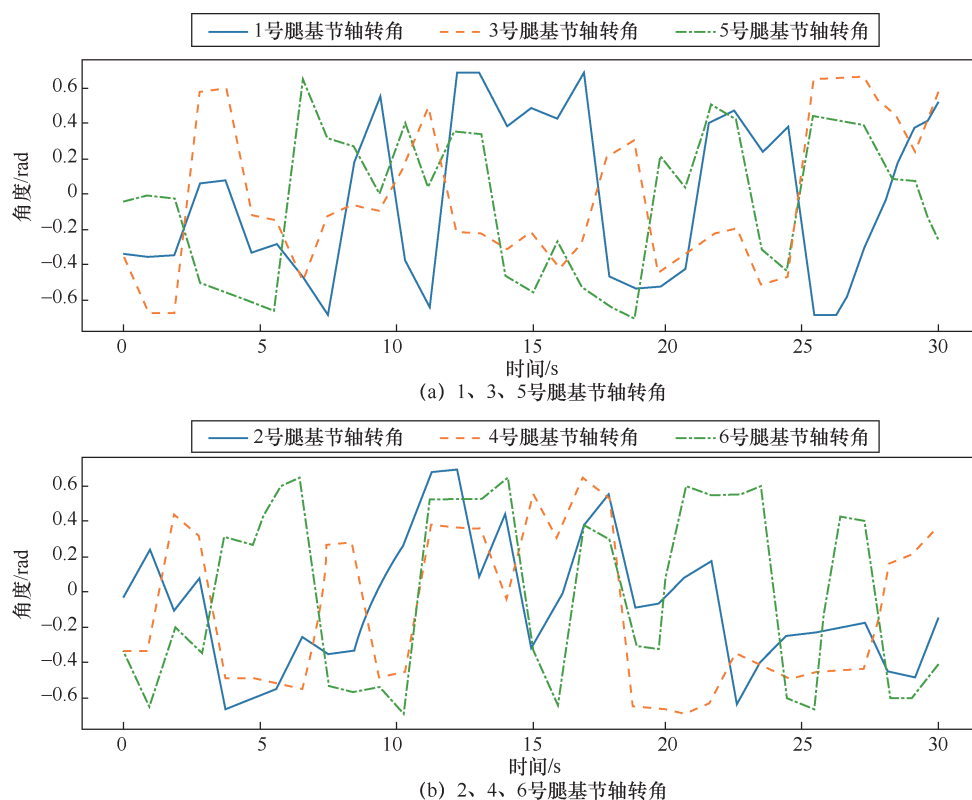


图10 基节轴转角变化曲线

随机梅花桩上行走的问题,对六足机器人进行了运动学分析,采用具有 PER 机制的 DDPG 算法,训练得到六足机器人在平面梅花桩环境中的运动策略,在随机给定起始坐标与目标区域的前提下,利用该运动策略可以规划得到六足机器人质心与足端的运动轨迹。最后等比例构建了真实平面梅花桩环境与六足机器人样机,利用规划得到的运动轨迹,六足机器人可以成功地从起始坐标运动到目标区域,进一步验证了运动策略的正确性。在未来的工作中,将针对六足机器人在三维随机梅花桩环境中的运动规划以及真实场景中的累计误差消除做进一步的研究。

参考文献:

- [1] 李满宏, 张明路, 张建华, 等. 六足机器人关键技术综述[J]. 机械设计, 2015, 32(10): 1-8.
LI M H, ZHANG M L, ZHANG J H, et al. Review on key technology of the hexapod robot[J]. Journal of Machine Design, 2015, 32(10): 1-8.
- [2] STELZER A, HIRSCHMÜLLER H, GÖRNER M. Stereo-vision-based navigation of a six-legged walking robot in unknown rough terrain[J]. The International Journal of Robotics Research, 2012, 31(4): 381-402.
- [3] HOMBERGER T, BJELONIC M, KOTTEGE N, et al. Terrain-dependant control of hexapod robots using vision[M]//International Symposium on Experimental Robotics. Cham: Springer, 2016: 92-102.
- [4] 胡勇. 六足机器人梅花桩行走步态研究[D]. 绵阳: 西南科技大学, 2019.
HU Y. Study on gait of hexapod robot walking on plum pile[D]. Mianyang: Southwest University of Science and Technology, 2019.
- [5] ČÍŽEK P, MASRI D, FAIGL J. Foothold placement planning with a hexapod crawling robot[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2017: 4096-4101.
- [6] MOSTAFA K, CHIANG J Y, WEI K T, et al. Image-based method for determining better walking strategies for hexapods[J]. International Journal of Advanced Robotic Systems, 2015, 12(5): 58.
- [7] 沈宇, 韩金朋, 李灵犀, 等. 游戏智能中的 AI——从多角色博弈到平行博弈[J]. 智能科学与技术学报, 2020, 2(3): 205-213.
SHEN Y, HAN J P, LI L X, et al. AI in game intelligence—from multi-role game to parallel game[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(3): 205-213.
- [8] 王飞跃, 曹东璞, 魏庆来. 强化学习: 迈向知行合一的智能机制与算法[J]. 智能科学与技术学报, 2020, 2(2): 101-106.
WANG F Y, CAO D P, WEI Q L. Reinforcement learning: toward action-knowledge merged intelligent mechanisms and algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(2): 101-106.
- [9] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.
- [10] 卜祥津. 基于深度强化学习的未知环境下机器人路径规划的研究[D]. 哈尔滨: 哈尔滨工业大学, 2018.

- BU X J. Research on robot path planning under unknown environment based on deep reinforcement learning[D]. Harbin: Harbin Institute of Technology, 2018.
- [11] 孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报, 2020(7): 1301-1312.
- SUN C Y, MU C X. Important scientific problems of multi-agent deep reinforcement learning[J]. Acta Automatica Sinica, 2020(7): 1301-1312.
- [12] GARIBALDI JONATHAN M, 陈虹宇, 李小双. 差异与学习: 模糊系统与模糊推理[J]. 智能科学与技术学报, 2019, 1(4): 319-326.
- GARIBALDI JONATHAN M, CHEN H Y, LI X S. Variation and learning: fuzzy system and fuzzy inference[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(4): 319-326.
- [13] 陈德旺, 蔡际杰, 黄允沛. 面向可解释性人工智能与大数据的模糊系统发展展望[J]. 智能科学与技术学报, 2019, 1(4): 327-334.
- CHEN D W, CAI J J, HUANG Y H. Development prospect of fuzzy system oriented to interpretable artificial intelligence and big data[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(4): 327-334.
- [14] WANG Z, CHEN C L, LI H X, et al. Incremental reinforcement learning with prioritized sweeping for dynamic environments[J]. IEEE/ASME Transactions on Mechatronics, 2019, 24(2): 621-632.
- [15] TSOUNIS V, ALGE M, LEE J, et al. Deepgait: planning and control of quadrupedal gaits using deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3699-3706.
- [16] SHAHRIARI M, KHAYYAT A A. Gait analysis of a six-legged walking robot using fuzzy reward reinforcement learning[C]//IEEE 13th Iranian Conference on Fuzzy Systems. Piscataway: IEEE Press, 2013: 1-4.
- [17] 唐开强. 基于迁移强化学习的六足机器人步态学习研究[D]. 南京: 南京大学, 2019.
- TANG K Q. Research on gait learning of hexapod robot based on transfer reinforcement learning[D]. Nanjing: Nanjing University, 2019.
- [18] HOU Y N, LIU L F, WEI Q, et al. A novel DDPG method with prioritized experience replay[C]//2017 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE Press, 2017: 316-321.
- [19] REN Z P, DONG D Y, LI H X, et al. Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2216-2226.
- [20] SHI W J, SONG S J, WU H, et al. Regularized anderson acceleration for off-policy deep reinforcement learning[J]. arXiv preprint, 2019, arXiv:1909.03245.
- [21] TING L H, BLICKHAN R, FULL R J. Dynamic and static stability in hexapedal runners[J]. Journal of Experimental Biology, 1994, 197(1): 251-269.
- [22] SUTTON R, BARTO A. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 1998.
- [23] WATKINS C. Learning from delayed rewards[J]. PhD thesis, University of Cambridge, 1989.
- [24] SILVER D, LEVER G, HEES N, et al. Deterministic policy gradient algorithms[C]//The 31st International Conference on Machine Learning. New York: ACM Press, 2014: 387-395.
- [25] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of Machine Learning Research. [S.l.:s.n.], 2015, 37: 448-456.
- [26] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint, 2015. arXiv:1509.02971.
- [27] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv preprint, 2015, arXiv:1511.05952.
- [28] MAHMOOD A R, HASSELT H P, SUTTON R S. Weighted importance sampling for off-policy learning with linear function approximation[C]//Conference and Workshop on Neural Information Processing Systems. New York: ACM Press, 2014: 3014-3022.

[作者简介]



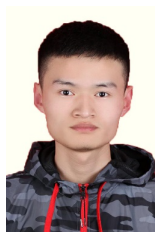
傅汇乔 (1996—), 男, 南京大学工程管理学院硕士生, 主要研究方向为机器人学、强化学习。



唐开强 (1992—), 男, 南京大学工程管理学院博士生, 主要研究方向为机器学习及随机优化等。



邓归洲 (1998—), 男, 西南科技大学制造科学与工程学院硕士生, 主要研究方向为机器视觉、机器人运动控制。



王鑫鹏 (1995—), 男, 西南科技大学制造科学与工程学院硕士生, 主要研究方向为深度学习、机器人运动控制。



陈春林 (1979—), 男, 博士, 南京大学工程管理学院教授, 主要研究方向为机器学习、智能机器人与量子控制等。