王胤清, 陈弘毅, 徐济濛,等. 基于多智能体强化学习的双足机器人行走学习研究 [J]. 智能计算机与应用,2026,16(6):1-9. DOI:10.20169/j.issn.2095-2163.24100601

基于多智能体强化学习的双足机器人行走学习研究

王胤清[1,2], 陈弘毅[1,2], 徐济濛[1], 薛 杰[1,2], 李清都[1]
(1 上海理工大学 机器智能研究院, 上海 200093; 2 上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: 双足机器人由于其高维动力学和高动态性,实现全向行走面临着巨大挑战。为应对这些挑战,本研究以自主设计的伸缩腿双足机器人为模型,基于多智能体强化学习,提出了一种专门针对双足机器人训练的算法。为提高训练效率和准确性,研究设计了一种经验回放机制,使智能体能够更有效地利用先前的训练经验。此外,引入课程学习方法,并根据伸缩腿机器人的行走特点,设计了一套周期性摆腿曲线。实验结果表明,所提出的方法能够更有效地实现双足机器人在侧向、前向及转向等全向步态的稳定行走。

关键词: 双足机器人; 伸缩腿设计; 多智能体强化学习; 经验回放机制; 全向行走
中图分类号: TP242 **文献标志码**: A **文章编号**: 2095-2163(2026)06-0001-09

Research on walking learning of bipedal robots based on multi-agent reinforcement learning

WANG Yinqing[1,2], CHEN Hongyi[1,2], XU Jimeng[1], XUE Jie[1,2], LI Qingdu[1]
(1 Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Bipedal robots face great challenges in achieving omnidirectional walking due to their high-dimensional dynamics and high dynamics. To address these challenges, this study takes a self-designed telescopic-legged bipedal robot as a model and proposes an algorithm specifically for bipedal robot training based on multi-agent reinforcement learning. To improve the efficiency and accuracy of training, the paper designs an experience replay mechanism that enables the agent to more effectively utilize previous training experience. In addition, a curriculum learning method is introduced, and a set of periodic leg swing curves are designed based on the walking characteristics of the telescopic-legged robot. Experimental results show that the proposed method can more effectively achieve stable walking of bipedal robots in omnidirectional gaits such as sideways, forward and turning.

Key words: bipedal robot; telescopic leg design; multi-agent reinforcement learning; experience replay mechanism; omnidirectional walking

0 引 言

双足机器人,也称仿人机器人,由于其结构与人类相似,在生活环境、生产建设以及家庭服务方面可以更好地进行工作,不需要再为其打造特殊的工作环境,因此能够更好地适应各种地形环境。然而,由于人类行走的复杂性以及对于动态平衡的要求,双足机器人的行走一直是一个具有挑战性的问题。

由于双足机器人结构复杂,非线性程度较高,所以一般采用简化模型的方式来进行步态规划并运用到机器人上[1]。Domínguez 等学者[2]提出了一种利用线性倒立摆模型(LIPM)和逆几何模型(IGM)为 12 自由度双足机器人生成步态周期的方法。Caron 等学者[3]提出了一种将线性倒立摆(LIP)模型的可捕获性分析推广到可变高度倒立摆(VHIP)模型的方法,这使得机器人能够在不平整地形上进行三维步行。Guan 等学者[4]提出了一种虚拟质量椭球倒立摆模型,并将其应用到 3 位双足机器人在不平整

基金项目: 国家自然科学基金(92048205); 东方学者计划(TP2019064)。
作者简介: 王胤清(1998—),男,硕士研究生。主要研究方向:机器人运动控制,强化学习; 陈弘毅(2000—),男,硕士研究生。主要研究方向:参数辨识,强化学习; 徐济濛(2001—),男,硕士研究生。主要研究方向:强化学习,机器人运动控制; 薛 杰(1997—),男,博士研究生。主要研究方向:强化学习,机器人运动控制。
通信作者: 李清都(1980—),男,博士,教授。主要研究方向:仿生机器人理论与技术,复杂系统的动力学与控制。Email:liqd@usst.edu.cn。
收稿日期: 2024-10-06

地面的运动控制。Dong 等学者[5]在机器人的平衡控制中,采用了模糊零力矩点(ZMP)控制器来维持机器人的平衡和稳定其质心,同时通过姿势控制器和阈值增加的躯干位置合规性控制器来增加稳定性边界。林知微等学者[6]提出的基于中枢模式发生器(CPG)的六足机器人运动控制方案,通过改进 Hopf 振荡器模型和设计合理的 CPG 网络拓扑结构,能够实现对六足机器人不同运动步态的稳定控制。Gibson 等学者[7]运用角动量线性倒立摆模型与模型预测控制(MPC)相结合,使用虚拟约束来增强机器人的敏捷性。以上的传统方法多年来一直被用于双足机器人的步态控制,然而,这些传统方法往往表现出计算复杂度较高、通用性差、鲁棒性低等特点。如对地面的要求过高、对外部抗干扰的能力有限,需要耗费大量的时间进行手动参数优化,往往会产生低于最优值的结果。

为了解决这些问题,近年来双足机器人的控制领域开始广泛采用强化学习(RL)方法[8-9]。强化学习方法在解决双足机器人的行走方面具有许多优点,包括自主学习、感知能力强、泛化能力强以及端到端训练等。这些优点使得强化学习成为解决双足机器人行走问题的强大工具,有助于提高机器人的行走稳定性、适应性和智能水平。万宇航等学者[10]提出了一种改进的近端优化算法的机器人臂动态规划方法,该方法解决了非结构化环境中环境不确定性增加和建模困难的挑战。刘晓德等学者[11]提出了一种基于强化学习和脉冲神经网络的智能运动控制算法,以提高多自由度连续运动控制的鲁棒性。Li 等学者[12]提出了使用多阶段训练框架来开发目标条件跳跃策略的观点,同时还使用了多层感知器(MLP)和一维卷积神经网络(1DCNN)来表示控制策略。在这个框架中,通过 3 个阶段的训练,包括单目标训练、多目标训练和动力学参数随机化,来提高机器人控制策略的泛化性,最终得到了高动态性能的跳跃策略。Duan 等学者[13]同样使用了卷积神经网络(CNN)来构建模型,将高维机器人状态映射到一系列步长命令的步长误差,以预测下一个着陆点的可达性。Yu 等学者[14]使用了长短期记忆网络(LSTM)应用于双足站立运动控制的策略架构,通过使用经过优化的单刚体模型轨迹生成的参考数据,训练循环策略来执行四步、90°转弯,再通过使用结尾奖励来学习特定行为,并成功转移到 Cassie 上。Singh 等学者[15]设计了一种遵循 ROS 框架中给定步骤序列的学习策略,可以实现机器人在平面及立体环境的全方位行走。Li 等学者[16]研究利用域随机化技术来促进策略学习适应系统动力学的变化,从而实现鲁棒行为。Rodriguez 等学者[17]提出了一种实现双足机器人步态的新方法。研究将 RL 与课程学习结合起来方法,逐步增加步态训练任务的难度,这种整合使得在各个方向上学习稳健和通用的步态成为可能。Kasaura 等学者[18]提出用于机器人控制的动作约束强化学习算法,确保学习系统采取的行动符合特定的约束,这对现实世界的可行性和安全性至关重要。Ou 等学者[19]提出了一个使用深度强化学习的框架,用于在仿真环境中学习控制策略并将其传输到实物机器人上。

1 多智能体强化学习算法设计

多智能体强化学习相较于传统强化学习方法,其优势在于每个智能体都可与其他智能体及环境进行交互,通过感知环境的状态再进行选择动作,并且多智能体可以通过协同合作来达到最终目标,而不是仅仅关注单个智能体的奖励最大化,使得系统表现出更为复杂和动态的行为模式。

1.1 马尔可夫决策过程建模

研究中将机器人的运动控制描述为一个马尔可夫过程(MDP):在每个时间步 t 内,机器人感知环境,获得状态 S_t,然后策略网络通过当前时刻的状态 S_t 推理出动作的分布 $\pi(A_t|S_t)$,并从中采样出动作 A_t。随后与环境交互,进入下一个时间步 $t+1$,并接收到奖励 r_{total}。本算法的目的是通过一种创新的强化学习算法,找到最优策略 π^* 使状态值函数的贝尔曼方程满足:

$$V^*(s)=\max_a E_\pi[R_{t+1}+\gamma V^*(S_{t+1}) \mid S_t=s, A_t=a] \tag{1}$$

其中,R_{t+1} 表示从状态 s 采取动作后得到的即时奖励;γ 表示折扣因子,$0\leqslant\gamma\leqslant1$,用于削弱未来奖励的影响;$E_\pi$ 表示在策略 π 下的期望奖励。

1.1.1 状态空间

本文讨论的策略网络输入是一个 58 维的状态 $S_t\in R^{58}$。该状态空间主要包括以下部分:机器人质心的加速度 $(a_x,a_y,a_z)\in R^3$,机器人质心的角速度 $(\omega_x,\omega_y,\omega_z)\in R^3$,每个关节的关节位置 $q_t\in R^6$,上一步动作 $A_{t-1}\in R^6$ 以及对应的 4 步历史数据共 48 维,机器人目标控制速度,包括了机器人前向速度以及机器人偏航角速度 $v_{\text{cmd}}, v_{\omega\text{cmd}}$ 以及左右脚的目标控制足底高度 h_{cmd}。具体有:

$$S_t = [a_t\omega_t q_t A_t \cdots q_{t-3}A_{t-3}v_{\mathrm{cmd}}v_{\omega\mathrm{cmd}}h_{\mathrm{cmd}}] \quad (2)$$

1.1.2　动作空间

本文所使用的伸缩腿双足机器人模型为欠驱动系统。在欠驱动系统中，PD 控制器无法直接控制所有自由度，通过传统的 PD 控制直接施加在关节转矩上可能会引起系统振荡和不稳定。为应对这一问题，使用位置增量的控制方法，对机器人的关节位置增量进行计算和调整，从而优化扭矩输出。这样既可以保持机器人的稳定状态，又可以增加其移动时的灵活性和精确性。计算方式如下：

$$q_t = q_{t-1} + \delta q_t \quad (3)$$

其中，q_t 表示 t 时刻关节角度位置；q_{t-1} 表示上一时刻目标关节位置；δq_t 表示关节位置增量值。

PD 控制器施加到关节扭矩 τ 的计算方式如下：

$$\tau = k_p(q_{\mathrm{des}} - q) + k_d(\dot{q}_{\mathrm{des}} - \dot{q}) \quad (4)$$

其中，k_p、k_d 表示 PD 增益；q_{des}、$\dot{q}_{\mathrm{des}}$ 分别表示所期望的关节目标位置及速度；q、$\dot{q}$ 分别表示关节当前位置及速度。

1.2　算法设计

1.2.1　演员评论家算法框架

在多智能体强化学习算法中，使用了演员评论家(Actor-Critic)算法框架。Actor-Critic 算法是一种强化学习方法，结合了值函数估计和策略优化两种方法，用于解决连续动作空间的问题。演员根据当前状态和策略 $\pi(a_t, s_t, \theta)$ 生成动作 a_t，其中 θ 表示策略网络的参数。评论家根据当前状态和行动估计动作值函数 $Q(s_t, a_t, \theta_q)$ 或优势函数 $A(s_t, a_t, \theta_a)$，其中，θ_q 和 θ_a 表示评论家网络的参数。算法的目标是使总回报 $J(\theta) = E_\pi[R]$ 最大化，其中 R 是回报值。为了实现这一目标，行动者利用策略梯度方法通过最大化策略值函数 $V(s_t, \theta)$ 的期望值来更新参数 θ：

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{t=1}^{N}\nabla_\theta \log \pi(a_t \mid s_t, \theta)A(s_t, a_t, \theta_a) \quad (5)$$

其中，N 表示样本的数量。通过迭代更新参数 θ，参与者优化策略以选择产生更高回报的行为。

批评家则根据所选择的行为和真实收益来估计行为价值函数或优势函数：

$$L(\theta_q) = \frac{1}{2}E_\pi[(Q(s_t, a_t, \theta_q) - R_t)^2] \quad (6)$$

$$L(\theta_a) = \frac{1}{2}E_\pi[(A(s_t, a_t, \theta_q) - R_t)^2] \quad (7)$$

其中，R_t 表示真实收益。通过梯度下降，更新评论家网络的参数 θ_q 和 θ_a，以提高对动作值函数或优势函数的估计精度。通过不断地更新演员及评论家模型，可以使得模型在较为复杂的条件下找到良好的策略。

多智能体联合策略可表示为 $\pi = \{\pi^1, \pi^2, \cdots, \pi^N\}$，所有智能体合作的主要目标是找到可以满足以下条件的最佳联合策略 π^*：

$$\pi^* = \arg\max_\pi \vartheta(\pi) \quad (8)$$

其中，$\vartheta(\pi)$ 表示所期望的折扣奖励。定义公式为：

$$\vartheta(\pi) = E_{(s_t, a_t)\sim\rho_\pi}\left[\sum_{t=0}^{+\infty} r(s_t, a_t)\right] \quad (9)$$

通过在算法中引入策略熵，可以显著提升多智能体系统的学习效率和解决复杂任务的能力：

$$\pi^* = \arg\max_\pi E_{(s_t, a_t)\sim\rho_\pi}\left[\sum_{t=0}^{\infty} r(s_t, a_t) + \sigma\varepsilon(\pi(\cdot \mid s_t))\right] \quad (10)$$

其中，σ 表示熵系数；$\varepsilon(\pi(\cdot \mid s_t))$ 的定义公式如下：

$$\varepsilon(\pi(\cdot \mid s_t)) = -\log(\pi(a_t \mid s_t)) \quad (11)$$

策略熵越高，其对未知环境的探索能力就越强，为了平衡探索及利用环境的能力，引入了熵系数的优化目标函数：

$$J(\sigma) = E_{a_t\sim\pi}[-\sigma\log \pi(a_t \mid s_t) - \sigma_\varphi] \quad (12)$$

其中，φ 表示主体的行动维度。

经验重放机制在强化学习中具有提高样本效率和稳定训练过程的技术。设置一个经验过滤单元，并设定一个阈值 v，将相似度小于 v 的经验样本放入低池，相似度高于 v 的经验样本放入高池。首先在初始阶段，训练产生的经验存储在经验过滤单元 $F = \{\phi_1, \phi_2, \cdots, \phi_n\}$ 中，并建立一个初始经验样本作为基线经验样本。使用欧氏距离计算新经验样本与基线经验样本之间的相似度，将经验样本添加到集合中，并根据相似度阈值将其分类到低池或高池中，异构策略回放机制如图 1 所示。



图 1　异构策略回放机制

Fig. 1　Heterogeneous strategy replay mechanism

1.2.2 奖励函数的设计

机器人的奖励函数被设计成一个包含多个部分的综合指标,其目的是激励机器人遵循指定的命令,保持身体平稳,实现稳定的运动。

首先设计奖励函数的结构,大部分的奖励函数是由当前的机器人状态与生成或设置的参考值对比得出的,即一种状态跟踪的奖励函数。定义一种指数距离函数:

$$r(u,v)=\exp(-\|u-v\|_2^2) \tag{13}$$

其中,u,v 定义了当前机器人的某一状态和需要跟踪的值。基于定义的指数距离函数设计奖励函数。

结合已知的状态信息,设计了如下的奖励函数:

$$r_{\text{total}}=r_{\text{vel}}+r_o+r_{\text{fheight}}+r_A+r_{\text{fpos}}+r_{\text{co}}+r_{\text{agm}}+r_{\text{alive}} \tag{14}$$

表 1 列出了各个奖励的设计细节。

表 1 奖励函数设计

Table 1 Reward function design

奖励设计	说明
$r_{\text{vel}}=r(v,v_{\text{cmd}})$	速度跟踪
$r_o=r(\omega_x,0)+r(\omega_y,0)$	保持平衡
$r_{\text{fheight}}=\prod_{i=0}^{1}r\left(\frac{h_{\text{tar}}^{\text{Ffoot}_i}+h_{\text{tar}}^{\text{Bfoot}_i}}{2},h_{\text{rel}}^{\text{foot}_i}\right)$	周期性步态跟踪
$r_A=r(A_t,A_{t-1})$	动作平稳
$r_{\text{fpos}}=\prod_{i=0}^{1}r\left(\frac{p_{\text{tar}}^{\text{Ffoot}_i}+p_{\text{tar}}^{\text{Bfoot}_i}}{2},p_{\text{rel}}^{\text{foot}_i}\right)$	落脚点学习
$r_{\text{co}}=\prod_{i=0}^{1}r(h_{\text{tar}}^{\text{Bfoot}_i},h_{\text{rel}}^{\text{Ffoot}_i})$	脚底一致性
$r_{\text{agm}}=\prod_{i=0}^{2}r(L_i,0)$	防止身体晃动
$r_{\text{alive}}=\begin{cases}-1, & \text{if stop condition}\\ 0, & \text{else}\end{cases}$	运动终止

表 1 中,r_{vel} 控制了速度方向与控制速度方向对齐,v_{cmd} 表示速度控制向量,包括了前向速度与转向速度;r_o 对 ω_x,ω_y 限制以防止机器人前倾和侧向晃动;r_{fheight} 跟踪了一个周期性的步态;$h_{\text{rel}}^{\text{foot}_i}$ 是由左右脚模仿的抬脚高度曲线输出的目标高度;$h_{\text{tar}}^{\text{Ffoot}_i}$ 和 $h_{\text{tar}}^{\text{Bfoot}_i}$ 是前脚掌和后脚掌的实际高度;r_A 规定了每次的动作幅度不能太大;$p_{\text{tar}}^{\text{foot}}$ 是经过 ROS 脚步规划器生成的落脚点,$p_{\text{rel}}^{\text{foot}}$ 是实际的落脚点,r_{fpos} 意在让机器人学习脚步规划器的策略;r_{co} 规定了前后脚掌不可以距离过远;r_{agm} 对机器人的角动量惩罚,角动量过大会导致行走途中剧烈晃动,设置该奖励函数有助于机器人的稳定行走。

1.3 课程学习框架设计

设计了小贝机器人的全向行走步态的课程学习框架,即侧向行走、前向行走、转向行走。在训练双足机器人的过程中,通过构建一个课程参数集合 α 来实现。这个集合中包含了各种任务难度的排序,能够更好地了解任务的挑战程度。根据这个集合中任务难度的分布,可以生成新的训练任务。这种方法使得机器人可以根据自身的训练状态来选择合适的训练任务,从而提高了训练效率并优化了行走步态。通过这种方式,能够更精确地调整训练过程,使机器人能够更快地学习和适应各种复杂的行走环境,进而提高其整体性能。由此得到的公式为:

$$\alpha=\{l\mid 0\leqslant l\leqslant l_{\max}\} \tag{15}$$

其中,$l_{\max}$ 表示课程学习中的最大难度级别。并且推得:

$$S_l=\begin{cases}1, & \text{taskcompleted}\\ 0, & \text{otherwise}\end{cases} \tag{16}$$

在课程设计中,将任务完成率作为一个最为关键的指标,并特别设计了一个名为 S_l 的指标函数,通过指标函数和课程参数的结合得到的结果,可以用来实时追踪任务完成情况。一旦任务完成率达到设定的标准阈值,这意味着机器人已经训练好并且有能力晋升至下一个学习难度级别。

在课程难度 l 的范围内,定义了课程学习任务完成率的公式为:

$$\Pr(l)=\frac{\sum_{i=0}^{N}S_l(i)}{N} \tag{17}$$

设定一参数 k,即定义一个任务完成的及格线,若 $\Pr(l)>k$,则表明机器人的训练难度可以晋升到下一难度级别。

图 2 为多智能体训练架构。课程学习算法步骤如下:

1. 初始化课程难度
2. 收集数据并使用公式 (α) 计算课程是否完成
3. 计算课程完成率
4. If: $\Pr(l)>k$
5. If: 到达最高难度
6. 课程难度随机化
7. Else:
8. 增加课程难度
9. Else:
10. 保持课程难度
11. 返回步骤 2

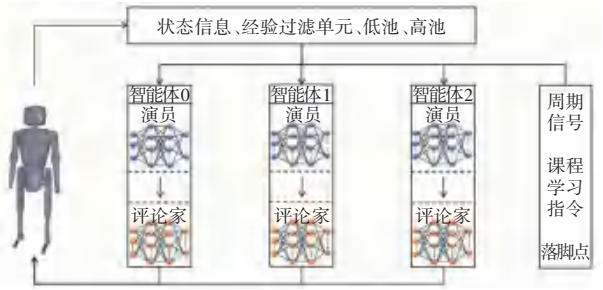


图 2　多智能体训练架构

Fig. 2　Architecture of multi-agent training

2　机器人行走设计

2.1　全向步态路径规划

实验中使用了双足机器人支撑脚的世界坐标系的位置和方向来确定机器人的位姿信息，位姿信息由双足机器人一侧的髋关节上附带的传感器提供，并且设计了一个 θ 用来表示方向。实验中，以坐标 (x,y,θ) 作为输入输出，其中 θ 表示当前方向和正方向之间的角度差。

以初始姿态的信息作为原点，然后使用 ROS 包中的足迹规划器来生成一条包含航向信息的二维路径 (x,y,θ)，以指导双足机器人的行走，在空白图上从目标范围 $(0,-1,-\frac{\pi}{2})$ 到 $(0,1,\frac{\pi}{2})$ 随机抽取，生成了 800 个目标着地点，用来预测机器人的落脚点。通过预测机器人的落脚点来实现机器人行走步态路径的规划。

2.2　摆腿曲线设计

机器人设计时，左右腿是相对的，因此在走路时，同样也具有对称性。机器人行走时，有支撑态和摆动态两种状态。在支撑态时，有双腿支撑阶段，以及一条腿处于摆动态时另一条腿的单腿支撑阶段。因此设计出了一种正弦曲线来模拟 2 条腿交替抬起放下的过程，作为机器人的摆腿曲线。摆腿曲线如下：

$$I_{\mathrm{ref1}}=\begin{cases}\dfrac{1-\cos(2\pi\times t)}{T}\times(0.5\times I_{\max}), & 0\leqslant t<\omega T\\ 0, & t\geqslant\omega T\end{cases}\tag{18}$$

$$I_{\mathrm{ref2}}=\begin{cases}0, & 0\leqslant t<\omega T\\ \dfrac{1-\cos(2\pi\times(t+\omega\times T))}{T}\times(0.5\times I_{\max}), & t\geqslant\omega T\end{cases}\tag{19}$$

其中，I_{ref1} 和 I_{ref2} 分别表示左右脚的摆腿模仿曲线；t 表示一个 step，直到增加到步行周期 T；$I_{\max}$ 表示双足机器人的最大步长；ω 表示双足机器人脚部支撑时间和摆动时间的占比。

2.3　周期性运动分析

机器人的行走步态具有明显的周期性运动。因此训练一步作为一个周期并采集一个周期结束后的双脚落地位置作为下一个周期的起点重复运动。在实验中，成功训练出了自主行走的机器人步态，以及一个用于学习摆腿模仿曲线的周期信号 P，周期信号如下：

$$P=[\sin(2\pi\times t)/T,\cos(2\pi\times t)/T]\tag{20}$$

3　实验

仿真实验中，操作系统使用 Ubuntu20.04LTS 版本，使用 Pytorch 深度学习框架，基于 AMD Ryzen Thread Ripper 3990X CPU 和 Nvidia GeForce RTX 3090 (24 GB) GPU 进行模型的训练。MuJoCo(Multi-Joint dynamics with Contact)是一个通用的物理引擎，旨在促进机器人学、生物力学、图形和动画、机器学习等领域的研究和开发，并提供了高度优化的物理仿真引擎，能够快速准确地模拟多关节系统的运动，所以使用 MuJoCo 仿真平台为机器人的行走仿真提供支撑。

3.1　机器人平台介绍

小贝机器人模型如图 3 所示，在模拟环境中构建了小贝机器人模型，以进行双足行走研究。作为一款伸缩腿直腿机器人，包含 17 个关节，这款机器人的一大亮点在于“双滑块”腿部伸缩机制设计，借助这种双滑块机制，成功地模拟了膝关节和踝关节的灵活运动。



图 3　小贝机器人模型

Fig. 3　Xiaobei robot model

小贝机器人关节信息见表 2，分别显示了各个关节的最大扭矩及关节运动范围。为了获取机器人的位姿信息，在髋部上方安装了惯性测量单元(IMU)。通过获取 IMU 的横滚角、俯仰角及偏航

角,从而得知机器人的行走状态。

表 2 小贝机器人关节主要信息

Table 2 Main information of Xiaobei robot joints

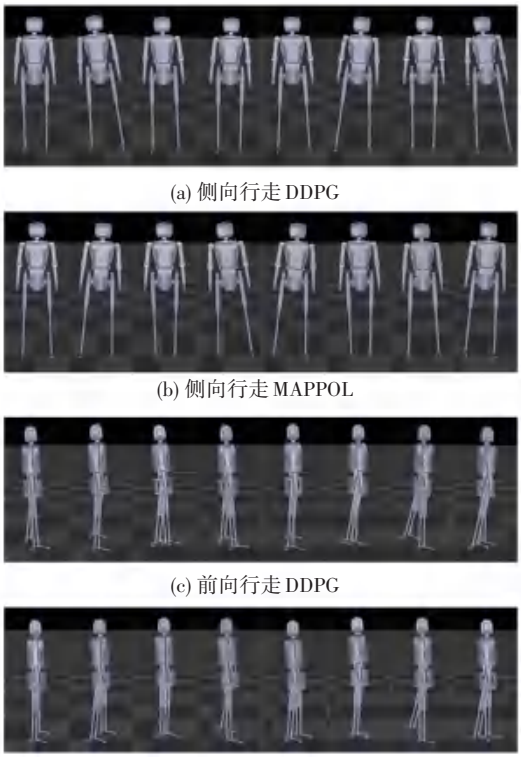
关节名称	最大扭矩/(N·m)	关节运动范围
头部偏航	64	−90°~90°
头部俯仰	64	−45°~45°
肩部俯仰	64	−45°~180°
肩部横滚	64	−0°~90°
肘部俯仰	64	−0°~135°
髋部偏航	1 000	−15°~15°
髋部横滚	500	−10°~10°
髋部俯仰	100	−90°~90°
伸缩腿	500	−200 mm~0 mm

此外,在仿真的实验中,使用改进的 PPO 方法(Multi - Agent Proximal Policy Optimization with Learning,以下称 MAPPOL)及 DDPG 算法[20]进行对比,在保证其他参数一致的情况下观察机器人的行走状态,从而验证改进算法的优势。为更好地观察机器人仿真效果,实验获取了机器人行走过程中的抬脚高度、质心位置及 ZMP 位置图,通过这些数据可以随时了解机器人行走所处的状态。奖励图可以很好地观察机器人的训练情况。

3.2 仿真实验

图 4 分别为侧向行走、前向行走的机器人状态,可以通过对比其抬脚高度、COM 位置及 ZMP 位置来判断其行走稳定性。

图 5、图 6 分别为小贝机器人侧向行走和前向行走的抬脚高度、COM 位置、ZMP 及双脚位置图。



(a) 侧向行走 DDPG

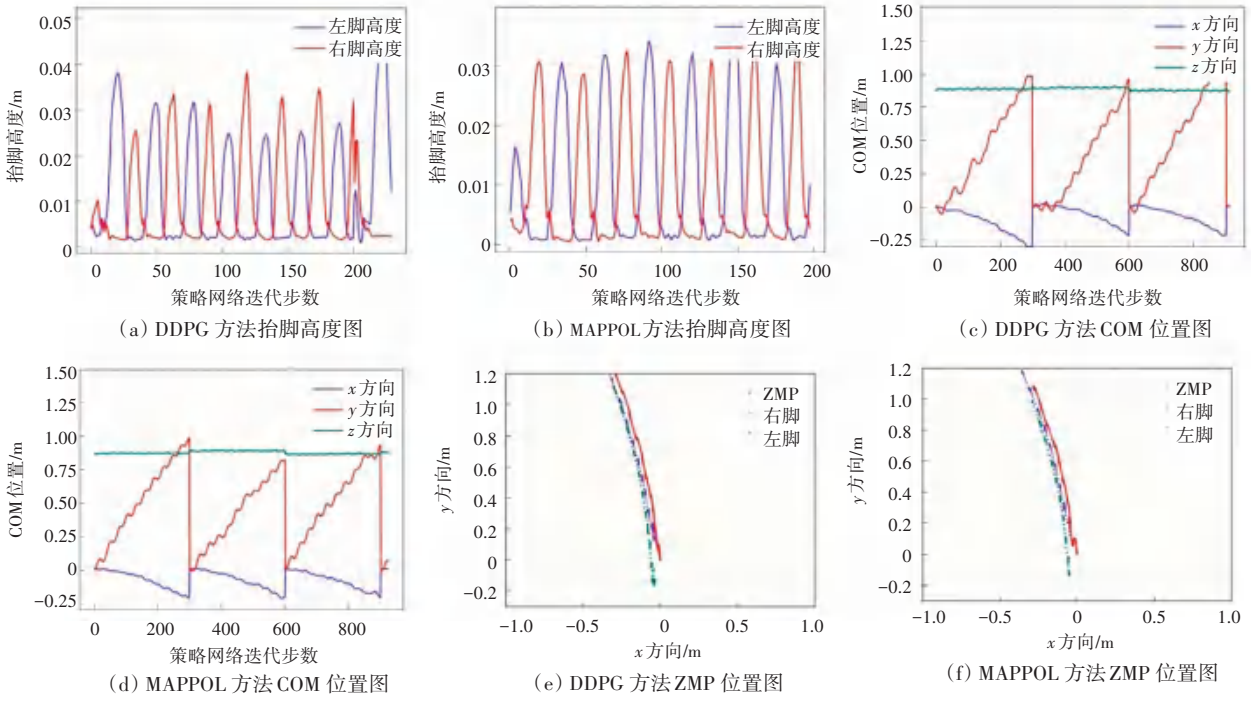
(b) 侧向行走 MAPPOL

(c) 前向行走 DDPG

(d) 前向行走 MAPPOL

图 4 小贝机器人行走状态图

Fig. 4 Diagram of the walking status of the Xiaobei robot



(a) DDPG 方法抬脚高度图

(b) MAPPOL 方法抬脚高度图

(c) DDPG 方法 COM 位置图

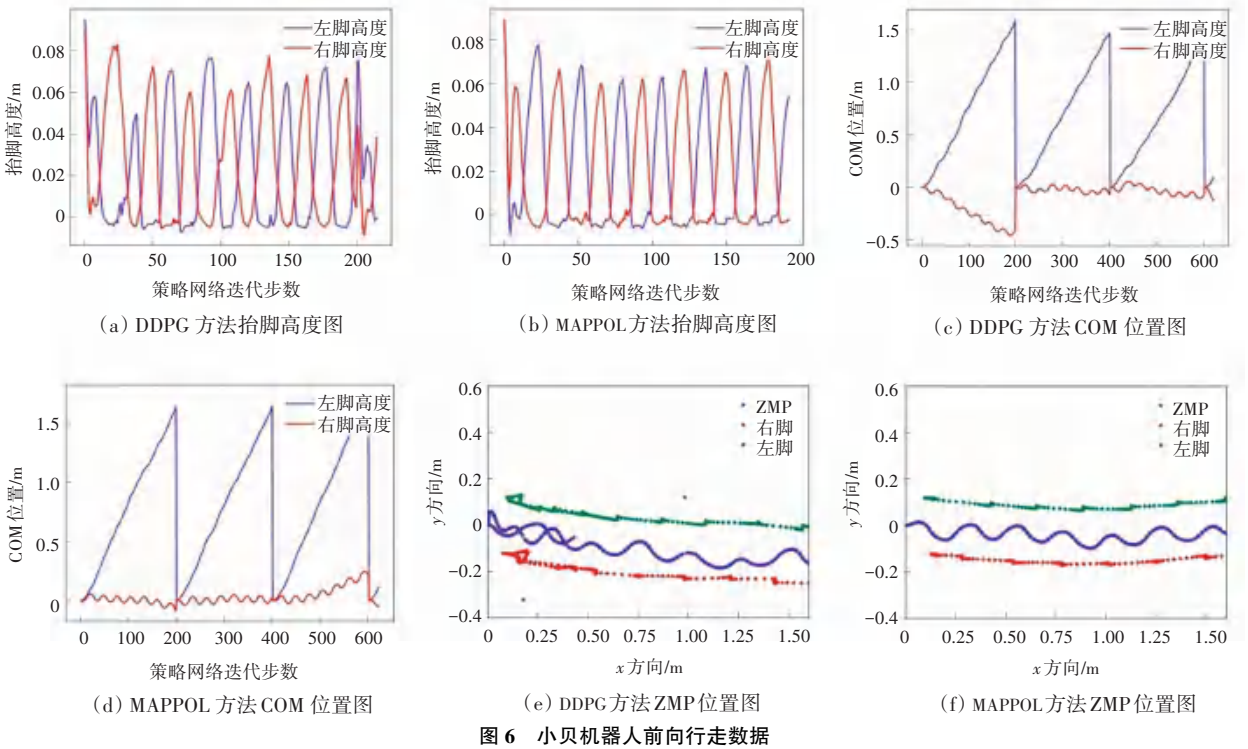
(d) MAPPOL 方法 COM 位置图

(e) DDPG 方法 ZMP 位置图

(f) MAPPOL 方法 ZMP 位置图

图 5 小贝机器人侧向行走数据

Fig. 5 Lateral walking data of Xiaobei robot



(a) DDPG 方法抬脚高度图　(b) MAPPOL方法抬脚高度图　(c) DDPG 方法 COM 位置图

(d) MAPPOL 方法 COM 位置图　(e) DDPG方法ZMP位置图　(f) MAPPOL方法ZMP位置图

图 6　小贝机器人前向行走数据

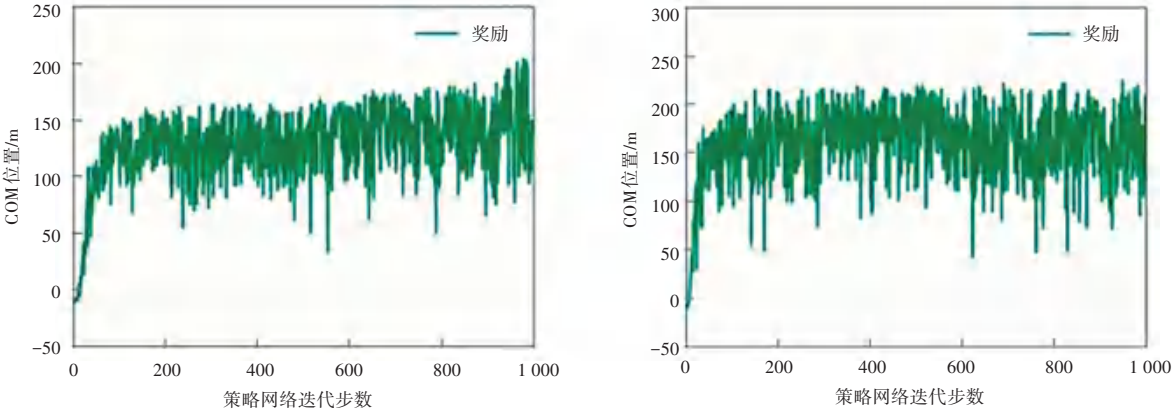
Fig. 6　Forward walking data of Xiaobei robot

就侧向行走而言，DDPG 方法两脚抬脚高度均差为 0.005 7 m，改进后的 PPO 算法两脚抬脚高度均差为 0.001 3 m；ZMP 与双脚中心线差值分别为 0.002 1 m 和 0.000 2 m。对比 DDPG 训练后方法，改进后的 PPO 方法训练的小贝机器人其抬脚高度较为一致，并且可以稳定地将 ZMP 位置保持在两脚之间，呈现出明显周期性特征，对于小贝机器人的训练稳定性有了大幅提升，使得行走更为平稳。

小贝机器人前向行走奖励如图 7 所示。由图 7 看出小贝机器人在前向、侧向行走任务中的总奖励随着步数的增加而变化。X 轴表示策略网络迭代步数。Y 轴表示总奖励值，范围从 0 到约 200。在训练一定步数后，奖励开始收敛至稳定。

由于 DDPG 方法无法实现小贝机器人的转向步态，本实验仅测试 MAPPOL 方法，图 8、图 9 为转向行走图及其他状态图。

通过图 9 中 COM 位置图及 ZMP 转向图可以看出，其在 X 轴和 Y 轴的移动相对平滑，说明 MAPPOL 方法对小贝机器人的转向训练效果良好，可以有效地实现机器人精确的转向动作。



(a) 侧向奖励图　(b) 前向奖励图

图 7　小贝机器人前向行走奖励

Fig. 7　Forward walking reward of Xiaobei robot

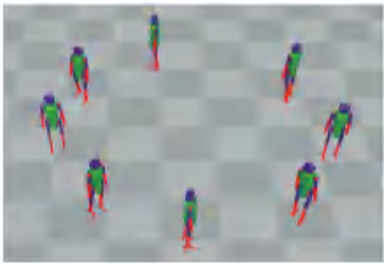


图 8 小贝机器人转向行走数据

Fig. 8 Steering walking data of Xiaobei robot

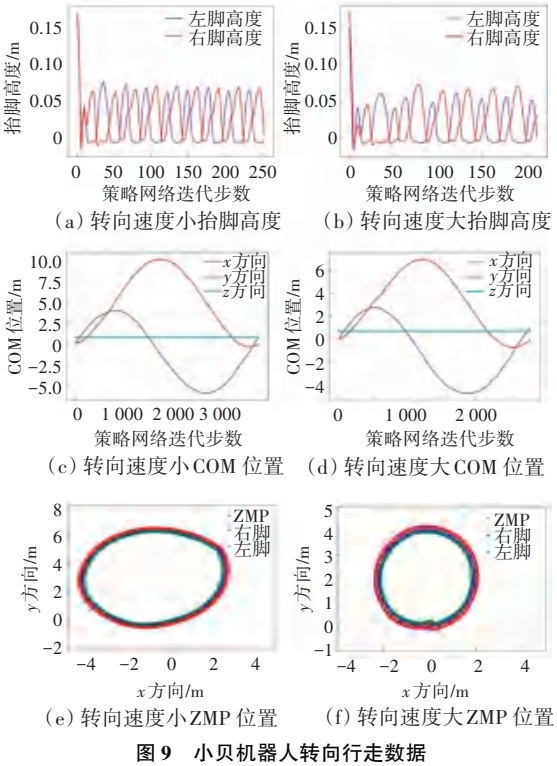


图 9 小贝机器人转向行走数据

Fig. 9 Steering walking data of Xiaobei robot

小贝机器人转向行走奖励如图 10 所示。观察转向行走累计奖励图,在训练较少的步数后就可以稳定在奖励较高的水平,并且未出现显著的趋势变动,仅有细微的波动,说明 MAPPOL 方法可以在短时间内训练至稳定状态。

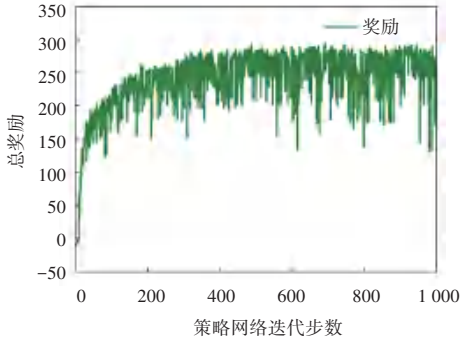


图 10 小贝机器人转向行走奖励图

Fig. 10 Chart of Xiaobei robot steering walking reward

4 结束语

本文提出了一种全新的多智能体强化学习框架,并且融合了课程学习的方法,有序地组织学习内容,逐步提高学习难度,从而实现高效的学习,提出的异构策略回放机制有效地加快了对于双足机器人步态的训练速度。完成了对于双足机器人周期性步态规划的设计,可以使得小贝机器人很好地实现稳定及平滑的行走,进一步验证了本文强化学习算法与双足机器人相结合的优势。

参考文献

[1] GUAN K, YAMAMOTO K, NAKAMURA Y. Push recovery by angular momentum control during 3d bipedal walking based on virtual - mass - ellipsoid inverted pendulum model [C]// Proceedings of 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids). Piscataway, NJ: IEEE, 2019: 120-125.

[2] DOMÍNGUEZ C A M, SÁNCHEZ E M, SOTO G I P, et al. Gait synthesis for a 12-degree-of-freedom biped robot by the lipm [C]//Proceedings of 2018 XX Congreso Mexicano de Robótica (COMRob). Piscataway, NJ: IEEE, 2018: 1-5.

[3] CARON S, ESCANDE A , LANARI L ,et al. Capturability-based pattern generation for walking with variable height [J]. IEEE Transactions on Robotics, 2020,36(2): 517-536.

[4] GUAN K, YAMAMOTO K, NAKAMURA Y. Virtual-mass-ellipsoid inverted pendulum model and its applications to 3d bipedal locomotion on uneven terrains [C]//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2019: 1401-1406.

[5] DONG C, YU Z , CHEN X ,et al. Adaptability control towards complex ground based on fuzzy logic for humanoid robots [J]. IEEE Transactions on Fuzzy Systems, 2022, 30(6): 1574-1584.

[6] 林知微, 林翰文, 刘宗朋. 基于 CPG 的六足机器人运动步态控制方法[J]. 计算机应用研究, 2020, 37(9): 2749-2753.

[7] GIBSON G, DOSUNMU-OGUNBI O, GONG Y, et al. Terrain-adaptive, alip-based bipedal locomotion controller via model predictive control and virtual constraints [C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2022: 6724-6731.

[8] BERANEK R, KARIMI M, AHMADI M. A behavior-based reinforcement learning approach to control walking bipedal robots under unknown disturbances [J]. IEEE/ASME Transactions on Mechatronics, 2022, 27(5): 2710-2720.

[9] YUAN Yuxia, LI Zhijun, ZHAO Ting, et al. Dmp-based motion generation for a walking exoskeleton robot using reinforcement learning[J]. IEEE Transactions on Inducstrial Electronics, 2020, 67(5): 3830-3839.

[10] 万宇航, 朱子璐, 钟春富, 等. 基于改进 PPO 算法的机械臂动态路径规划[J]. 系统仿真学报, 2025, 37(6): 1462-1473.

[11] 刘晓德, 郭宇飞, 陈元培, 等. 基于脉冲强化学习的连续运动控制仿真与优化[J]. 系统仿真学报, 2025, 37(10): 2662-2671.

[12] LI Zhongyu, PENG Xuebin, ABBEEL P, et al. Robust and

versatile bipedal jumping control through reinforcement learning [J]. arXiv preprint arXiv, 2306.13199, 2023.

[13] DUAN Helei, MALIK M S, DAO G J, et al. Sim-to-real learning of footstep-constrained bipedal dynamic walking[C]// Proceedings of 2022 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2022: 10428-10434.

[14] YU Fangzhou, BATKE R, DAO J, et al. Dynamic bipedal turning through sim-to-real reinforcement learning [C]// Proceedings of 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids). Piscataway, NJ: IEEE, 2022: 903-910.

[15] SINGH R P, BENALLEGUE M, MORISAWA M, et al. Learning bipedal walking on planned footsteps for humanoid robots [C]//Proceedings of 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids). Piscataway, NJ: IEEE, 2022: 686-693.

[16] LI Zhongyu, CHENG Xuxin, PENG Xuebin, et al. Reinforcement learning for robust parameterized locomotion control of bipedal robots [C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2021: 2811-2817.

[17] RODRIGUEZ D, BEHNKE S. Deepwalk: omnidirectional bipedal gait by deep reinforcement learning[C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2021: 3033-3039.

[18] KASAURA K, MIURA S, KOZUNO T, et al. Benchmarking actor-critic deep reinforcement learning algorithms for robotics control with action constraints[J]. arXiv preprint arXiv, 2304.08743, 2023.

[19] OU Yafei, TAVAKOLI M. Sim-to-real surgical robot learning and autonomous planning for internal tissue points manipulation using reinforcement learning[J]. IEEE Robotics and Automation Letters, 2023, 8(5): 2502-2509.

[20] 李敏, 张森, 曾祥光, 等. 基于深度强化学习的四足机器人单腿越障轨迹规划[J]. 系统仿真学报, 2025, 37(4): 895-909.