文章编号:1672-6553-2025-23(10)-035-010

DOI:10.6052/1672-6553-2025-050

基于强化学习的仿蠕虫机器人驱动排布优化研究*

俞云潇　张舒†

(同济大学 航空航天与力学学院,上海　200092)

摘要　针对多单元仿蠕虫机器人驱动优化问题,本文提出了一种基于强化学习的智能配置方法.首先建立多单元仿蠕虫机器人的动力学模型,并将驱动排布问题形式化为马尔可夫决策过程.通过设计多重离散动作空间,显著降低了计算成本.提出融合运动速度和能耗约束的奖励函数,有效平衡了探索与开发矛盾.针对驱动器受限条件,提出的动作掩蔽机制实现了硬性约束下的高效策略搜索.研究发现:(1)全驱动时中部对称激活最优;(2)约束条件下呈现“后部优先、向心聚集”的排布规律.

关键词　仿蠕虫机器人,　驱动优化,　强化学习,　动作掩码,　多单元系统

中图分类号:TP242　　**文献标志码**:A

Optimization of Actuation Configuration in Earthworm-like Robots via Reinforcement Learning*

Yu Yunxiao　Zhang Shu†

(School of Aerospace Engineering and Applied Mechanics, Tongji University, Shanghai　200092, China)

Abstract　This study presents a reinforcement learning-based intelligent method for optimizing the actuation of configuration multi-segment earthworm-like robots. A dynamic model of the multi-segment robotic system is first established, and the actuator arrangement problem is formulated as a Markov decision process. By designing a multi-discrete action space, computational costs are significantly reduced. A reward function integrating locomotion speed and energy consumption constraints is proposed to effectively balance exploration and exploitation. For actuator-limited conditions, an action masking mechanism enables efficient policy search under hard constraints. Key findings include: (1) Midline-symmetric actuation yields optimal performance under full-drive conditions; (2) A “posterior-priority, centripetal-clustering” distribution pattern emerges under constrained actuation.

Key words　earthworm-like robot,　actuation optimization,　reinforcement learning,　action masking,　multi-segment system

引言

仿生机器人模仿自然界中生物的外部形状、运动原理和行为方式,能从事与生物特点相衬的工作.近年来,蠕虫型生物的蚯蚓由于在平衡可塑性和力量方面的优异表现,逐渐受到了学术界和工业

2025-04-23 收到第 1 稿,2025-05-13 收到修改稿.

* 国家自然科学基金资助项目(12472022), National Natural Science Foundation of China(12472022).

† 通信作者 E-mail:zhangshu@tongji.edu.cn

界的关注[1-3]. 蚯蚓的生理结构主要有三个特点:身体单元的分节、单元内环纵肌肉拮抗协调和刚毛[4]. 其分节特征使得每个单元都可独立工作,即每个单元的变形不会影响相邻单元的状态. 当蚯蚓运动时,每个单元内环向和纵向肌肉交替收缩,不同单元间交替伸长和缩短,这种拮抗协调机制使得蚯蚓形成蠕动波[5,6],此时单元内刚毛交替向外伸出,保持与外界环境的锚固,使得蚯蚓实现向前运动[7]. 由此,研究人员模仿蚯蚓生物学特征和运动机理开发了仿蠕虫移动机器人,其结构简单、控制性良好、微型化潜力较大,可以通过灵活改变自身形状进行运动,有望满足工业领域对在非结构化和不可预测的环境中进行作业的要求,实现管道的检测与维修、地震废墟探索搜救、灾后地形侦察与信息搜集等[8-13].

对于仿蠕虫移动机器人,不同的研究团队在力学建模、单元及驱动器设计、材料选用等方面都存在差异,所以会针对不同的移动机器人模型选择合适的优化方法来优化系统的运动性能. Fang 等[14]以机器人稳态平均速度和正则化平均速度为步态优化目标函数,利用 KKT 条件求解该问题,得到对应于不同单元数目的最大平均速度和最大正则化平均速度,并归纳出相应最优步态. 该团队还针对机器人相位排布模式进行研究,利用对称群理论得到机器人所有可能的对称性相位排布模式,通过优化各种模式得到最优相位差模式. 研究发现恒等相位差排布模式,与最优相位排布模式相比,对稳态平均速度影响非常小,但可以实现驱动降维也有利于实验的进行和工程实践. Bhovad 等[15]根据 Kresling 折纸段设计并制作了两单元仿蠕虫折纸机器人,并利用 Kresling 折纸中的多重稳定性生成只需一个驱动器的驱动循环. 研究使用优化算法计算用于该折纸机器人的 Kresling 设计参数,目标函数为机器人总应变能,设定相关运动学约束,寻找合适的折纸段长度使得总应变能局部最小. 最终实现仅由一个驱动模块,来生成确定的变形序列驱动两单元仿蠕虫折纸机器人,实现减少机器人系统中的驱动器数目而不影响机器人运动这一目标. Jiang 等[16]利用两个方波驱动仿蠕虫机器人运动,研究机器人运动性能与一些重要参数(波峰距离、波宽等)之间的最佳关系. 使用准静态模型描述类蠕虫运动,研究其在一定时间内的稳态平均速度与参数之间的关系. 并根据平均速度表达式和运动约束,利用 KT 条件优化各种参数间的最优关系,后续通过数值方法和实验验证了优化结果的有效性.

当前针对分节式仿蠕虫移动机器人的驱动优化研究主要存在以下几点局限性:首先,现有研究大多聚焦于稳态平均速度或相位协调等单一性能指标的优化;其次,研究对象通常局限于单元数目较少的机器人系统(一般不超过 7 个单元),对于高维情况下的多单元机器人(如 11 个单元),特别是在驱动器数量受限条件下的优化问题,相关研究仍较为缺乏;此外,现有研究多着眼于全驱动条件下的移动速度优化,而较少探讨机器人的功能性优化问题,如驱动资源受限场景下的优化配置问题,特别是驱动器部分失效或激活数量受限时的最优排布策略.

值得注意的是,强化学习方法[17,18]在处理此类高维优化问题时具有独特优势. 相较于传统启发式算法,强化学习通过深度神经网络的表征能力,能够有效处理高维稀疏搜索空间的问题. 具体而言,启发式算法在高维空间中往往面临局部最优解数量激增和样本稀疏性导致的搜索效率低下问题;但强化学习的在线学习特性使其能够动态评估和调整策略,而启发式算法在参数变更时需要重新进行迭代计算,计算成本显著增加. 鉴于实际物理实验难以穷尽所有可能的驱动配置方案,本研究采用理论建模与仿真验证相结合的方法:首先建立动力学方程构建仿真环境,如 Zhang[19]建立的仿蠕虫机器人多单元动力学模型,进而通过强化学习算法在虚拟环境中实现驱动排布策略的优化.

基于上述分析,本研究重点开展以下工作:首先建立 11 单元仿蠕虫机器人的动力学模型,并将驱动排布优化问题形式化为马尔可夫决策过程;随后提出基于动作掩蔽机制的约束优化方法,实现受限条件下的最优驱动排布. 这一研究框架为解决多单元仿生机器人的驱动优化问题提供了新的思路和方法.

1 仿蠕虫机器人动力学建模

为能使用强化学习算法在虚拟环境中实现驱动排布的优化,本节构建了多单元仿蠕虫机器人的动力学模型. 蚯蚓的定向运动依赖于其体节的拮抗变形与各向异性摩擦的协同作用,具体表现为:(1)

分节式体节结构——通过环向与纵向肌肉的交替收缩产生拮抗变形;(2)动态摩擦调控——刚毛在体节膨胀时伸出以增大摩擦,收缩时收回以减小摩擦;(3)运动波传导——神经信号控制体节激活的相位滞后,形成沿体长传播的蠕动波.基于此,可以构建出相应物理构型.

如图 1 所示,仿蠕虫机器人由 n 个同构运动单元串联而成,各单元参数相同,其中质量为 m,刚度为 k,阻尼为 c,原长为 L_0.每个单元包含集中质量块,相邻单元间通过并联的弹性结构与阻尼器连接,构成拮抗变形结构.驱动器输出的周期性驱动力与弹性元件的恢复力交替作用,使单元产生拮抗变形,并通过接触界面摩擦力的调制,将驱动能量转化为净推进力.为保证蠕动波传导的完整性,系统末端附加质量为 m 的刚性配重模块.定义 $x_i \in \boldsymbol{R}$ 为第 i 个单元质心在全局坐标系下的绝对位移 $(i=1,2,\cdots,n)$,附加刚体的绝对位移为 x_{n+1}.

对各个单元进行受力分析(如图 2 所示)并建立机器人的动力学方程,对于第 i 个单元:

$$m\ddot{x}_i = F_{\mathrm{d}}^{(i)} - F_{\mathrm{d}}^{(i-1)} - k(\delta_i - \delta_{i-1}) - c(\dot{\delta}_i - \dot{\delta}_{i-1}) - F_{\mathrm{f}}^{(i)} \tag{1}$$

其中,$F_{\mathrm{d}}^{(i)} = \tau_i(t)$ 为单元驱动力,且 $F_{\mathrm{d}}^{(0)} = F_{\mathrm{d}}^{(n+1)} = 0$;$\delta_i = x_i - x_{i-1} - L_0$ 为弹簧变形量;$F_{\mathrm{f}}^{(i)}$ 为第 i 单元与接触面的摩擦力,且 $F_{\mathrm{f}}^{(n+1)} = 0$.那么,可以得到 n 单元仿蠕虫移动机器人的动力学方程组:

$$\begin{cases} m\ddot{x}_1 = F_{\mathrm{d}}^{(1)} - k\delta_1 - c\dot{\delta}_1 - F_{\mathrm{f}}^{(1)} \\ m\ddot{x}_i = F_{\mathrm{d}}^{(i)} - F_{\mathrm{d}}^{(i-1)} - k(\delta_i - \delta_{i-1}) - c(\dot{\delta}_i - \dot{\delta}_{i-1}) - F_{\mathrm{f}}^{(i)}, \ i = 2,\cdots,n \\ m\ddot{x}_{n+1} = -F_{\mathrm{d}}^{(n)} + k\delta_n + c\dot{\delta}_n \end{cases} \tag{2}$$

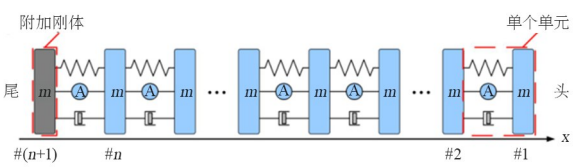


图 1　多单元仿蠕虫机器人力学模型简图

Fig. 1　Schematic diagram of the multi-segment worm-like robot's mechanical model

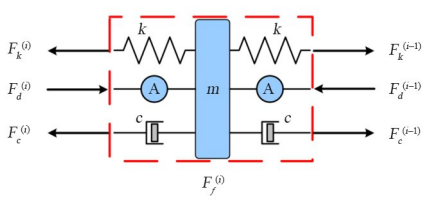


图 2　单单元受力分析示意图

Fig. 2　Force analysis schematic of the individual robotic segment

为实现类蚯蚓蠕动波的传播,定义第 i 单元的理想变形函数:

$$L_i(t) = L_0 + al_0\sin(\omega t - \phi_i),\ i = 1,\cdots,n \tag{3}$$

其中,a 为单元变形系数,ω 为角速度,ϕ_i 为第 i 个单元的初相位.本文使用恒等相位差控制模式来生成蠕动波,此种模式下,单元间的相位差恒为 $\Delta\phi$,则 ϕ_i 可表示为:

$$\phi_i = (i-1)\Delta\phi,\ i = 1,\cdots,n \tag{4}$$

机器人在移动时单元持续与外界接触,我们以库伦干摩擦为基础设计如下摩擦力形式:

$$F_{\mathrm{f}}^{(i)} = S_i^{\beta}\mu mg\,\mathrm{sgn}(\dot{x}_i),\ i = 1,\cdots,n \tag{5}$$

$$S_i(t) = \frac{L_0}{x_i - x_{i+1}},\ i = 1,\cdots,n \tag{6}$$

其中,μ 为摩擦系数,S_i 表示单元径向变形对摩擦力的影响,上标 β 为径向变形系数,当 $\beta=0$ 时,表示单元的变形不会影响摩擦力的大小.当 $\beta>0$ 时,若 $S_i>1$,单元膨胀导致受到更大摩擦,若 $S_i<1$,单元回缩导致受到更小摩擦.

采用 PD 控制实现对运动的调控:

$$\begin{aligned} &\tau_i(t) = K_{\mathrm{p}}e_i(t) + K_{\mathrm{d}}\dot{e}_i(t) \\ &e_i(t) = L_{\mathrm{the},i}(t) - L_{\mathrm{act},i}(t),\ i = 1,\cdots,n \\ &\dot{e}_i(t) = \dot{L}_{\mathrm{the},i}(t) - \dot{L}_{\mathrm{act},i}(t) \end{aligned} \tag{7}$$

其中,K_{p} 为比例系数,K_{d} 为微分系数,$e_i(t)$ 和 $\dot{e}_i(t)$ 为位移偏差及其导数,$L_{\mathrm{the},i}(t)$ 和 $\dot{L}_{\mathrm{the},i}(t)$ 为第 i 个单元的理论变形及其导数,$L_{\mathrm{act},i}(t)$ 和 $\dot{L}_{\mathrm{act},i}(t)$ 为第 i 个单元的真实变形及其导数.

2　基于强化学习的驱动器优化方法

2.1　马尔可夫决策过程建模

仿蠕虫机器人的驱动器排布优化问题可形式化为马尔可夫决策过程 $\langle \boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R}\rangle$,其中 $\boldsymbol{S}$(状态空间)一般表示机器人所有可能的状态,$\boldsymbol{A}$(动作空间)表示机器人对驱动器排布方式的可能选择,$\boldsymbol{P}$ 根据当前状态和动作确定下一个状态的状态转移概率,$\boldsymbol{R}$(奖励函数)表示采取动作后从环境中获得的奖励.智能体(仿蠕虫机器人)在离散时间步 t 观测状态 $\boldsymbol{s}_t \in \boldsymbol{S}$,执行动作 $\boldsymbol{a}_t \in \boldsymbol{A}$ 后,依据状态转移概率 $\boldsymbol{P}(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)$ 迁移至新状态 $\boldsymbol{s}_{t+1}$,并获得即时奖励 $\boldsymbol{r}_t = \boldsymbol{R}(\boldsymbol{s}_t,\boldsymbol{a}_t,\boldsymbol{s}_{t+1})$.该过程满足马尔可夫性质,即 $\boldsymbol{P}(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t) = \boldsymbol{P}(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_0,\boldsymbol{a}_0,\cdots,\boldsymbol{s}_t,\boldsymbol{a}_t)$,确

保系统动态仅由当前状态－动作对完全决定. 接下来,我们将具体定义状态空间 $\boldsymbol{S}$、动作空间 $\boldsymbol{A}$ 和奖励函数 $\boldsymbol{R}$.

2.1.1 状态空间

文中状态空间 $\boldsymbol{S}$ 被构造为包含局部运动特征和全局运动性能的复合观测向量,其数学表达为:

$$\boldsymbol{s}_t=[\boldsymbol{s}_{\text{units}}^{\mathrm{T}},v_{\text{com}}]^{\mathrm{T}}\in\mathbb{R}^{24+1} \tag{8}$$

其中 $\boldsymbol{s}_{\text{untis}}=[x_1,\dot{x}_1,x_2,\dot{x}_2,\cdots,x_{12},\dot{x}_{12}]^{\mathrm{T}}\in\mathbb{R}^{24}$,表征 11 个仿蠕虫运动单元及 1 个尾部配重刚体的位移与速度,v_{com} 表征机器人的质心速度. 这种状态空间设计能够全面捕捉仿蠕虫机器人的局部和全局运动特性,为智能体提供足够的环境信息以学习最优策略.

2.2.2 动作空间

动作空间被设计为 $\boldsymbol{A}=\{0,1\}^{11}$ 的多维离散空间,其中每个元素 $a_i\in\{0,1\}$ 表征第 i 个驱动器的激活状态,其中 0 表示休眠,1 表示激活. 直接枚举 11 个单元的所有可能驱动器排布将导致动作空间维度呈指数爆炸($2^{11}=2048$ 种组合),而本方案通过分解独立决策将维度压缩至 11 维,有效解决了传统编码方式的维数灾难问题. 任意动作向量 $\boldsymbol{a}=(a_1,a_2,\cdots,a_{11})^{\mathrm{T}}$ 唯一确定驱动器的空间激活模式,例如当执行动作 $\boldsymbol{a}=(1,0,1,0,0,1,0,0,1,0,1)^{\mathrm{T}}$ 时,第 1、3、6、9、11 号单元驱动器被激活.

2.2.3 奖励设置

在强化学习框架中,奖励函数的设计对策略优化具有决定性作用. 机器人的最终目标是找到使稳态平均速度最大化的最优排布. 然而,这一目标奖励具有明显的稀疏性,仅在找到最优排布后才能获得显著反馈. 为了缓解稀疏奖励问题,我们设置了两个即时奖励分量,前向速度奖励 v_t 和驱动器数量奖励 n_t,表示为:

$$R(s_t,a_t)=\alpha\cdot v_t+\beta\cdot n_t \tag{9}$$

其中,v_t 表示时间步 t 的前向速度,n_t 表示当前激活的驱动器数量,α 为速度权重,β 为数量权重. 本文的奖励设置中,速度奖励的权重 α 需要相对大以保证移动性能的主导地位,而数量奖励的权重 β 则应相对较小,主要起到打破局部最优的辅助作用. 在后续强化学习训练中,$\alpha=10$,$\beta=0.33$,这种设计使得智能体在初期能够广泛探索不同驱动配置,随着学习深入则逐步聚焦于优化移动性能.

2.2 强化学习算法

2.2.1 神经网络

基于上述状态观测空间和问题描述,本文采用标准的 Actor-Critic 强化学习架构,由策略网络(Actor)和价值网络(Critic)2 个核心组件构成. 如图 3 所示,策略网络采用多重离散动作空间输出结构,其架构包含 1 个共享的特征提取层和 11 个独立的输出层. 特征提取层为单隐藏层全连接结构,输入为 25 维状态向量,包含机器人各单元的运动状态及整体质心速度,隐藏层包含 128 个神经元并使用 ReLU 激活函数. 每个输出层对应 1 个机器人单元,为了后续对算法施加动作掩蔽,输出层处不使用激活函数,直接输出 2 维原始逻辑值,分别表示该单元驱动器激活和未激活的倾向性.

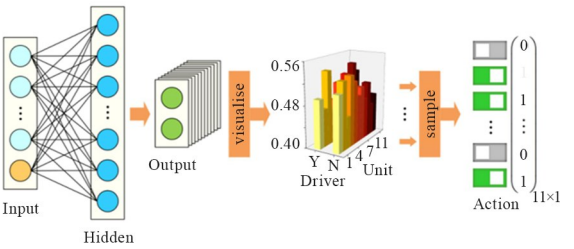


图 3 神经网络结构设计

Fig. 3 Neural network architecture design

对于由 11 个独立二进制决策组成的复合动作 a_t,其联合动作的对数概率[20]可以分解为各维度对数概率的加权平均:

$$\log\pi(a_t\mid s_t)=\frac{1}{11}\log\prod_{c=1}^{11}\pi^c(a_t^c\mid s_t) \tag{10}$$

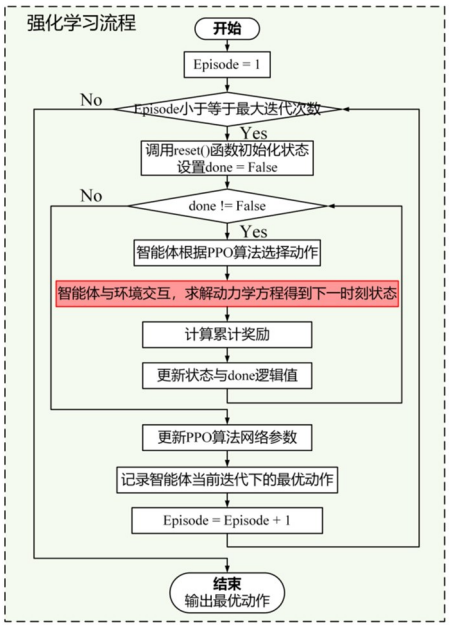
这一动作设计使得强化学习算法能够对每个单元的驱动器状态进行独立决策. 通过从各输出层的概率分布中采样,最终组合形成完整的 11 单元驱动排布方案.

价值网络采用与策略网络相同的特征提取层结构,并与策略网络共享层中参数,包括 128 个神经元的单隐藏层(ReLU 激活). 输出层为单神经元线性结构,用于输出状态价值估计以计算优势函数.

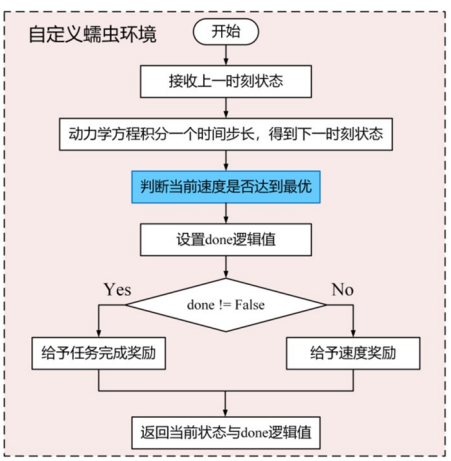
2.2.2 强化学习流程

本文采用近端策略优化(PPO)算法作为核心训练框架,该算法通过引入信任域约束和重要性采样机制,在保证策略更新稳定性的同时显著提高了

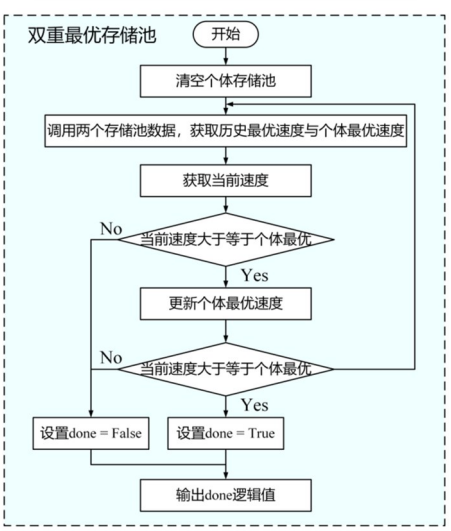
样本利用效率. 如图 4 所示,构建了一个闭环的“智能体－动力学环境”交互学习架构,其中智能体通过策略网络生成动作指令,动力学环境反馈状态转移信息,这种紧密耦合的设计特别适用于仿蠕虫机器人这类具有连续状态空间和离散动作空间的复杂控制问题.



(a) 强化学习算法流程图
(a) Flowchart of reinforcement learning algorithm



(b) 自定义蠕虫环境流程图
(b) Flowchart of customized earthworm-like robot environment



(c) 双重最优存储池流程图
(c) Flowchart of dual-optimal memory pool architecture

图 4　仿蠕虫机器人强化学习算法流程图
Fig. 4　Flowchart of the reinforcement learning algorithm for earthworm-like robots

与经典强化学习框架不同,本文将仿蠕虫机器人的非线性动力学模型直接整合到环境交互环节中. 在每个时间步 t,智能体根据当前状态 $\boldsymbol{s}_t \in \mathbb{R}^{25}$ 生成动作 $\boldsymbol{a}_t \in \{0,1\}^{11}$ 后,该动作输入至蠕虫动力学模型,通过采用变步长龙格－库塔数值积分方法求解一个时间步长,获得新状态 $\boldsymbol{s}_{t+1}$ 并计算即时奖励 $\boldsymbol{R}_t$. 智能体利用这些交互数据进行策略优化,经过充分学习后输出使平均运动速度最大化的最优排布策略,从而完成强化学习的过程.

在终止条件设计方面,本文采用双重最优存储机制:构建个体最优速度池记录当前训练周期内的最佳性能,构建历史最优速度池保存全局最优解. 训练过程分为两个阶段:初期鼓励智能体着重探索,直至迭代较长时间步时终止;后期引入双重收敛准则,要求当前策略产生的运动速度需同时超越当前训练周期内的个体最优速度及全局历史最优速度,方可终止训练.

3　驱动器配置优化分析

3.1　仿真环境

本文基于 OpenAI Gym 框架搭建了仿蠕虫机器人运动训练环境,使用 Pytorch 框架实现强化学习. 在此仿真环境中,蠕虫机器人将进行直线运动,运动状态受动力学方程控制. 在运动过程中,机器

表 1　仿蠕虫机器人动力学环境参数设置
Table 1　Parameter configuration for the dynamic environment of earthworm-like robots

Parameter	Value	Unit
L_0	0.0385	m
β	5	—
m	0.07	kg
g	9.8	m/s^2
k	205	N/m
c	0.1	N·s/m
μ	0.4	—
K_p	2043	N/m
K_d	158.9	N·s/m

人会随着时间步的增加改变单元中驱动器的激活情况，通过强化学习模型评估机器人当前稳态平均速度的优劣并给予相对奖励，目标是使得机器人以最高稳态平均速度移动. 仿真中使用的动力学环境参数与 PPO 算法超参数设置如表 1 与表 2 所示，其中动力学环境参数主要来源于 Zhang 等[19]的工作.

表 2 PPO 算法超参数设置

Table 2 Hyperparameter configuration of the PPO algorithm

Hyperparameter	Description	Value
actor_lr	Policy learning rate	1×10^{-3}
critic_lr	Value learning rate	1×10^{-2}
γ	Discount factor	0.98
λ	GAE parameter	0.95
ε	Policy update threshold	0.2

3.2 驱动器全局优化分析

对于具有 n 个单元数的蠕虫机器人来说，对移动性能影响最大的因素是驱动器激活的个数与激活方式，即驱动器的优化排布方式. 驱动器排布方式的全局优化结果如图 5、图 6 所示.

可以看到，强化学习训练过程展现出显著的阶段性特征. 图 6 显示，训练奖励折线在前 15 个训练周期呈现明显的探索期特征，随后在 15～20 个训练周期实现快速提升，最终稳定在高奖励平台期. 这一演化过程表明，智能体成功实现了从随机探索到稳定策略的转变. 图 5 展示了训练步数随迭代次数的变化(柱状图)，进一步揭示了策略优化的动态特性：在前 15 个周期，智能体迭代变换排布至最大时间步也难以找到较优解，而在训练中后期(大于 20 个训练周期)，该数值显著降低. 这种搜索效率的跃升说明智能体已经有效学习到环境动力学特征与最优排布方式之间的映射关系.

图 7 的最终排布统计显示，在训练中后期的所有实验中，最优策略均表现为 11 个驱动器全部激活的状态(出现频率 100%). 这一发现直接回答了本研究的关键问题：在给定动力学参数条件下(见表 1)，不添加额外限制条件，最大化仿蠕虫机器人运动速度的最优驱动策略为全驱动器激活模式. 该策略的优越性也符合一般物理认知：通过协调所有单元的拮抗变形，实现最大程度的蠕动波传导，同时全激活状态产生的连续推力分布，也有效克服了地面摩擦.

为了探索蠕虫机器人功能化实现的过程中可能出现的驱动资源受限问题，我们限制驱动器的激活个数，探究在一定驱动器激活个数下，机器人具有最大移动速度的驱动排布情况，如图 7 所示.

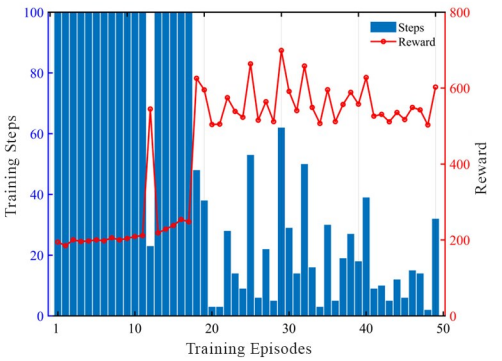


图 5 奖励与迭代步数图

Fig. 5 Reward vs training iterations diagram

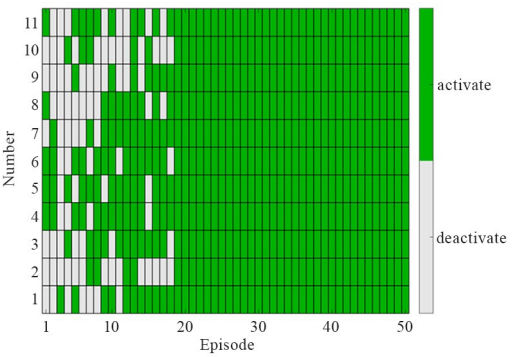


图 6 迭代周期对应排布图

Fig. 6 Activation pattern vs training epochs diagram

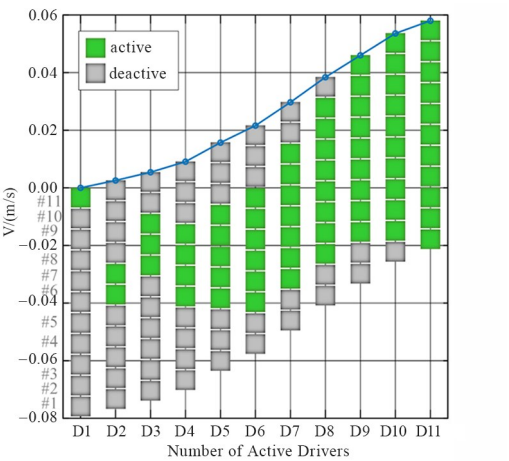


图 7 不同驱动器个数对应的最优排布

Fig. 7 Optimal actuator configurations for varying numbers of active drivers

在限定驱动器激活数量的条件下，智能体展现出具有明确空间规律的优化策略. 强化学习结果表明，最优驱动器排布遵循“中部优先，对称扩展”的

原则.当激活单元数小于等于 6 时,移动速度随激活单元向中部集中而显著提升;当激活单元数大于 6 时,速度增加趋于平缓,此时系统开始向两端对称扩展激活单元.

3.3　驱动器的动作掩蔽优化分析

3.3.1 动作掩蔽

为了进一步探究在驱动资源受限条件下的最优排布策略问题,本文提出一种基于动作掩码的约束优化方法.传统基于惩罚函数的方法,如对非法动作施加负奖励的方法,存在两点局限性:其一,无法完全避免智能体探索非法动作空间,导致部分训练周期浪费在无效尝试上;其二,会显著延缓算法收敛速度.为此,本文采用动作掩码机制,来消除非法动作的选取可能性.如图 8 所示,该方法通过在策略网络的输出层引入可编程动作掩码,实现对动作空间的智能约束:在神经网络输出动作概率分布后,对非法动作进行硬性屏蔽.本文重点考察此类典型约束条件:强制屏蔽中部单元的激活权限,研究对称性被破坏情况下对机器人运动性能的影响.

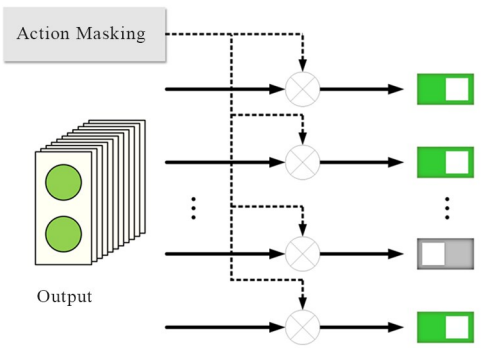


图 8　神经网络施加动作掩蔽示意图

Fig. 8　Schematic of action masking in neural networks

在技术实现层面,策略网络输出层输出 11 个独立的二维概率分布 $\pi^c \in \mathbb{R}^2$($c=1,\cdots,11$),每个分布对应一个驱动单元的激活决策:

$$\pi^c = [\pi_0^c, \pi_1^c], \quad \pi_0^c + \pi_1^c = 1 \tag{11}$$

其中,π_0^c 代表单元 c 未激活的概率,π_1^c 代表单元 c 被激活的概率.

对于指定的掩码单元集合 D(如 $D=\{6,7\}$,代表单元 6 和单元 7 的驱动器将不能被激活),执行以下操作:

$$\pi_1^c = \begin{cases} -\infty & \text{若 } c \in D \\ \pi_1^c & \text{否则} \end{cases} \tag{12}$$

随后使用 Softmax 进行重新归一化,当 $c \in D$ 时:

$$\begin{aligned} \pi^c &= \mathrm{softmax}[\pi_0^c, \pi_1^c] \\ &= \left[\frac{e^{\pi_0^c}}{e^{\pi_0^c}+e^{\pi_1^c}}, \frac{e^{\pi_1^c}}{e^{\pi_0^c}+e^{\pi_1^c}}\right] \\ &= \left[\frac{e^{\pi_0^c}}{e^{\pi_0^c}+0}, \frac{0}{e^{\pi_0^c}+0}\right] \\ &= [1,0] \end{aligned} \tag{13}$$

此时,单元 c 未激活的概率恒为 1,可被激活的概率为 0,这确保采样时被掩蔽单元只能输出未激活状态,即被掩蔽单元只能输出动作'0'(驱动器未激活).而由于被掩蔽单元的强制归零,在梯度回传过程中,这些被掩蔽位置的概率分布值恒为常数,从而不会产生任何参数更新信号.

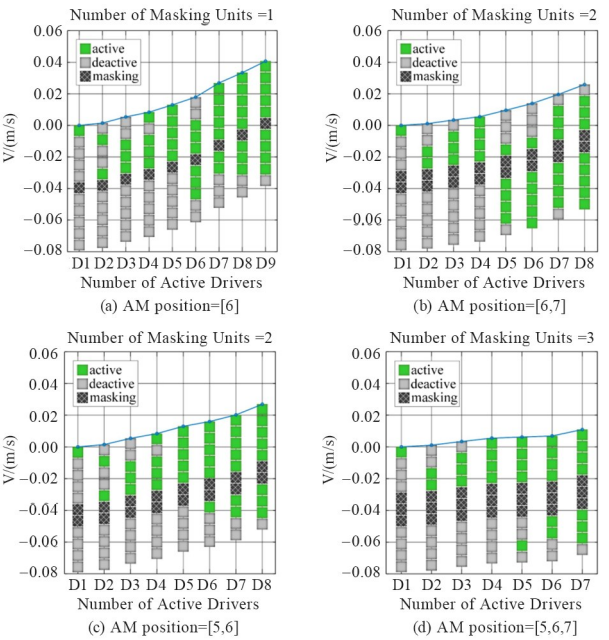
3.3.2 掩蔽后的优化分析

在仿真中,我们分别设置掩码位置从中部向两端延伸,将不同驱动器激活数下的最优排布方式按质心速度进行升序排列,如图 9 所示.

仿真结果表明,当施加动作掩蔽后,驱动器激活策略呈现显著的空间偏好特征(图 9).具体表现为:

(1) 后部单元优先激活现象.在屏蔽第 6 单元(中部)后,当可用驱动器数量小于等于 5 时(即不足总单元数的一半),最优排布全部集中于机器人后半段(第 7～11 单元),且激活单元位置呈现向中部靠拢的趋势.这一现象表明后部单元在非对称条件下对移动速度具有更大的贡献.

(2) 分段填充规律.当驱动器数量大于 5 时,系统首先填满后半段可用单元,再向前半段扩展.



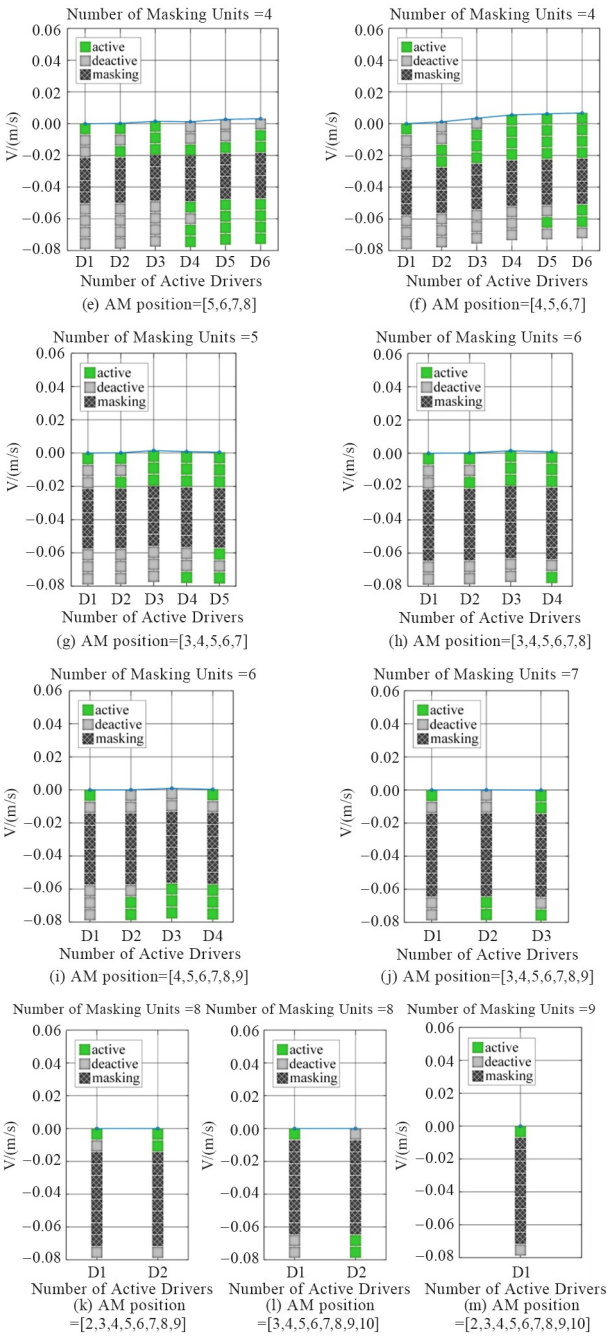


图 9 施加动作掩蔽后不同驱动器个数对应的最优排布

Fig. 9 Optimal actuator configurations under action masking constraints

(3) 中部偏向性. 无论在何种掩码条件下,当激活单元数小于当前可用后半单元数时,驱动器始终优先激活最靠近中部的可用单元.

(4) 前后段不对称响应. 当掩码导致前半段可用单元多于后半段时(如屏蔽 6、7 单元),最优策略反而转向优先激活前半段单元,但依然保持向中部聚集的特性. 这一反直觉现象揭示出:驱动器的空间价值排序为后部单元>前部单元>中部单元(在对称破坏条件下).

上述发现证实,在对称性破坏的驱动配置下,后部单元的激活优先级显著高于前部单元. 这一规律可能与仿蠕虫机器人的动力学特性有关:后部单元驱动比前部单元驱动能更有效地形成推力累积. 该结论为驱动器激活受限状态下的机器人容错控制提供了重要指导——当部分驱动器失效时,应优先保障后部近中区域的驱动能力.

4 结论

本文针对多单元仿蠕虫机器人的驱动优化问题,提出了一套基于强化学习的智能驱动配置方法,主要取得以下研究进展:

首先,通过建立 11 单元仿蠕虫机器人的动力学模型并将其形式化为马尔可夫决策过程,构建了融合动力学模型与强化学习的优化框架. 基于该框架的优化研究发现,最大移动速度对应的驱动排布为全驱动器激活模式,符合一般物理认知,验证了算法的有效性.

其次,研究发现驱动器排布优化存在一定的空间规律——在无约束条件下呈现“中部优先,对称扩展”的特征. 当激活单元数小于等于 6 时,运动速度随中部单元激活显著提升;激活单元数大于 6 时,运动速度的增加随单元激活递减.

最后,在驱动器受限条件下揭示的反直觉现象具有重要科学价值. 对称破坏时后部单元激活优先级高于前部;存在“分段填充”的排布序列规律;始终表现出向心聚集特性. 这些发现可以说明,后部单元更利于推力累积.

本研究发现的驱动排布规律对仿蠕虫机器人的实际设计具有明确指导意义:针对驱动器布局设计,根据“中部优先,对称扩展”规律,建议将高功率驱动器集中配置在机器人中后部,以最大化推力效率;针对驱动器故障场景,可基于“分段填充”规律调整激活序列——优先保持后部单元激活,再补充中部单元;在能源受限时,采用“向心聚集”策略,即先激活中后部 3 个核心单元再逐步向两端扩展.

本研究的主要贡献在于:开发了支持动作掩蔽且结合动力学模型的强化学习方法;发现了驱动资源受限下的驱动排布空间优化规律. 研究成果不仅为多单元仿生机器人的驱动优化提供了解决方案,更为复杂系统中的资源受限优化问题研究开辟了新思路.

参考文献

[1] MAEDA K, SHINODA H, TSUMORI F. Miniaturization of worm-type soft robot actuated by magnetic field [J]. Japanese Journal of Applied Physics, 2020, 59: SIIL04.

[2] CHEN Y H, CHEN Y, WANG J G. A novel worm drive for rehabilitation robot joint [C]//2013 IEEE 9th International Conference on Mobile Ad-hoc and Sensor Networks. New York: IEEE, 2013: 586－590.

[3] MANWELL T, VÍTEK T, RANZANI T, et al. Elastic mesh braided worm robot for locomotive endoscopy [C]//2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. New York: IEEE, 2014: 848－851.

[4] BLANCO-CANQUI H. Cover crops and soil ecosystem engineers [J]. Agronomy Journal, 2022, 114(6): 3096－3117.

[5] WANG Y F, PANDIT P, KANDHARI A, et al. Rapidly exploring random tree algorithm-based path planning for worm-like robot [J]. Biomimetics, 2020, 5(2): 26.

[6] ZHAO J, ZHANG Y H, ZHU Y H, et al. Locomotion control of modular self-reconfigurable robot worm-like structure [J]. Journal of Southeast University. Natural Science Edition, 2007, 37(3): 409－413.

[7] XAVIER M S, FLEMING A J, YONG Y K. Experimental characterisation of hydraulic fiber-reinforced soft actuators for worm-like robots [C]//2019 7th International Conference on Control, Mechatronics and Automation (ICCMA). New York: IEEE, 2019: 204－209.

[8] MANWELL T, GUO B J, BACK J, et al. Bioinspired setae for soft worm robot locomotion [C]//2018 IEEE International Conference on Soft Robotics (RoboSoft). New York: IEEE, 2018: 54－59.

[9] XU F P, ZHOU Y, ZHAO Z C. Research on walking principle of a worm pipe robot. In: Wang Y, Mechanical science and engineering IV [C]//4th International Conference on Mechanical Science and Technology (ICMSE 2014). Baech, Switzerland: Trans Tech Publications Ltd, 2014: 115－119.

[10] ZHANG K T, QIU C, DAI J S. Helical kirigami-inspired centimeter-scale worm robot with shape-memory-alloy actuators [C]//Proceedings of the ASME Design Engineering Technical Conferences and Computers and Information in Engineering Conference (DETC). New York: ASME, 2014.

[11] 方虹斌，彭海军. 先进机器人中的动力学与控制专刊序[J]. 动力学与控制学报，2023，21(12)：1－4.
FANG H B, PENG H J. Preface to the special issue: dynamics and control of advanced robots [J]. Journal of Dynamics and Control, 2023, 21(12): 1－4. (in Chinese)

[12] 刁斌斌，徐鉴，何健锋，等. 一款仿蚯蚓机器人的纤维驱动特性建模与辨识[J]. 动力学与控制学报，2023，21(2)：1－11.
DIAO B B, XU J, HE J F, et al. Actuation feature modeling and identification of a fiber-driven earthworm-like robot [J]. Journal of Dynamics and Control, 2023, 21(2): 1－11. (in Chinese)

[13] 周柏李，方虹斌，徐鉴. 模块化可重构机器人动力学研究进展[J]. 动力学与控制学报，2023，21(1)：1－17.
ZHOU B L, FANG H B, XU J. Advances in dynamics of modular reconfigurable robots [J]. Journal of Dynamics and Control, 2023, 21(1): 1－17. (in Chinese)

[14] FANG H B, WANG C H, LI S Y, et al. A comprehensive study on the locomotion characteristics of a metameric earthworm-like robot [J]. Multibody System Dynamics, 2015, 35(2): 153－177.

[15] BHOVAD P, KAUFMANN J, LI S Y. Peristaltic locomotion without digital controllers: exploiting multi-stability in origami to coordinate robotic motion [J]. Extreme Mechanics Letters, 2019, 32: 100552.

[16] JIANG Z W, XU J. The optimal locomotion of a self-propelled worm actuated by two square waves [J]. Micromachines, 2017, 8(12): 364.

[17] 刘丰瑞，颜格，张晓龙，等. 基于深度强化学习的动基座双自由度系统动力学控制方法[J]. 动力学与控制学报，2023，21(10)：26－33.
LIU F R, YAN G, ZHANG X L, et al. Dynamic control method for dual-degree-of-freedom systems with moving base by deep reinforcement learning [J]. Journal of Dynamics and Control, 2023, 21(10): 26－33. (in Chinese)

[18] 柯恺宸，金士博，高博扬，等. 基于强化学习的机器人多接触交互任务控制[J]. 动力学与控制学报，

2023，21(12)：53－69.
KE K C，JIN S B，GAO B Y，et al. A survey of robot intelligent control method in contact-rich tasks [J]. Journal of Dynamics and Control，2023，21(12)：53－69. (in Chinese)

[19] ZHANG G，ZHANG S. Actuation coordination of the earthworm-like locomotion robot based on dynamic model [J]. Chinese quarterly of mechanics，2022，43(4)：791－802.

[20] NIKOLAIDIS S，RAMAKRISHNAN R，GU K R，et al. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks [C]//Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. New York：ACM，2015：189－196.