doi: 10.7690/bgzdh.2025.02.018

基于多智能体强化学习的履带机器人摆臂控制方法

张洪川[1,2]，任君凯[2]，潘海南[2]，梅　勇[1]，卢惠民[2]
（1. 中国兵器装备集团自动化研究所有限公司特种计算机事业部，四川 绵阳 621000；
2. 国防科技大学智能科学学院，长沙 410073）

摘要：为解决摆臂式履带机器人在 3 维环境下实现自主摆臂控制面临的挑战，提出一种基于多智能体强化学习的摆臂控制方法。将机器人的每个摆臂视为一个独立智能体，设计一套兼顾底盘稳定性和摆臂动作的奖励函数，采用多智能体强化学习训练各个摆臂运动；将所提方法部署在基于 Isaac Sim 搭建的 3 维仿真环境中，通过向每个智能体输入局部高程图和机器人状态，输出摆臂转角。实验结果表明：该方法能实现多种地形下的摆臂自主控制，在机器人自主越障方面相对于单智能体强化学习有显著提升。

关键词：多智能体强化学习；履带机器人；自主越障；摆臂自主控制
中图分类号：TP242.6　　**文献标志码：**A

Articulated Flipper Control Method for Tracked Robots Based on Multi-agent Reinforcement Learning

Zhang Hongchuan[1,2], Ren Junkai[2], Pan Hainan[2], Mei Yong[1], Lu Huimin[2]
(1. *Department of Special Computer*, *Automation Research Institute Co*., *Ltd*. *of China South Industries Group Corporation*, *Mianyang* 621000, *China*;
2. *College of Intelligence Science and Technology*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: To address the challenges faced by flipper tracked robots in achieving flipper autonomous control in a 3D environment, a flipper control method based on multi-agent reinforcement learning is proposed. Consider each flipper of the robot as an independent intelligent agent, design a reward function that balances chassis stability and flipper movements, and use multi-agent reinforcement learning to train the movements of each flipper; Deploy the proposed method in a 3D simulation environment based on Isaac Sim, and output the flipper angle by inputting local elevation maps and robot states to each agent. The experimental results show that this method can achieve autonomous control of the flipper in various terrains, and has significant improvement in robot autonomous obstacle crossing compared to single agent reinforcement learning.

Keywords: multi-agent reinforcement learning; tracked robot; autonomous traversal; flipper autonomous control

0　引言

带有摆臂结构的履带式机器人（简称摆臂履带机器人）是特种机器人中较为常见的类型。国防科技大学自主研发设计的 NuBot 四摆臂履带机器人[1]在左右差动控制的双履带底盘基础上，增加了可主动控制转动的摆臂。这种设计不仅继承了履带底盘的大触地面积和强负载能力等优点，而且使机器人在遇到崎岖地形时可以通过转动摆臂改变自身构型和地面支撑点。相比其他形式的机器人，基于摆臂履带结构的特种机器人越障能力最强，且结构简单可靠，但其运动控制自由度高，对操作手在视距外进行遥操作提出了较高要求；因此，降低摆臂履带机器人的操作难度，实现摆臂自主运动控制成为一个亟待解决的问题。

基于此，笔者将多智能体强化学习与摆臂机器人相结合，旨在解决履带机器人在越障过程中摆臂运动的自主控制问题，从而降低机器人操作难度。

1　方法

笔者采用多智能体强化学习的方法解决摆臂机器人自主越障的问题，整体框架如图 1 所示。下面将详述每个部分的设计细节。

1.1　状态空间和动作空间设计

每个智能体的局部观测空间 o^t 可以表示为：

$$o^t=\{h_{\text{map}}, \theta_{\text{orient}}, v_{\text{cmd}}, \theta_{\text{flipper}}, \omega\}。\tag{1}$$

式中：h_{map} 为基于车辆中心的高程地图，范围为 2.5 m×1.2 m；$\theta_{\text{orient}}=\{\theta_{\text{roll}}, \theta_{\text{pitch}}, \theta_{\text{yaw}}\}$ 为摆臂机器人在世界坐标系下的方向；v_{cmd} 为车辆前进速度的指

收稿日期：2024-07-24；**修回日期：**2024-08-24
基金项目：国家自然科学基金资助项目（62203460, U22A2059）；国防科技大学自主创新科学基金（24-ZZCX-GZZ-11）
第一作者：张洪川（2000—），男，四川人，硕士。

令值；$\theta_{\text{flipper}}=\{\theta_{\text{fl}},\theta_{\text{fr}},\theta_{\text{rl}},\theta_{\text{rr}}\}$ 为机器人 4 个摆臂的角度；ω 是从 IMU 中获取的机器人的角速度值。

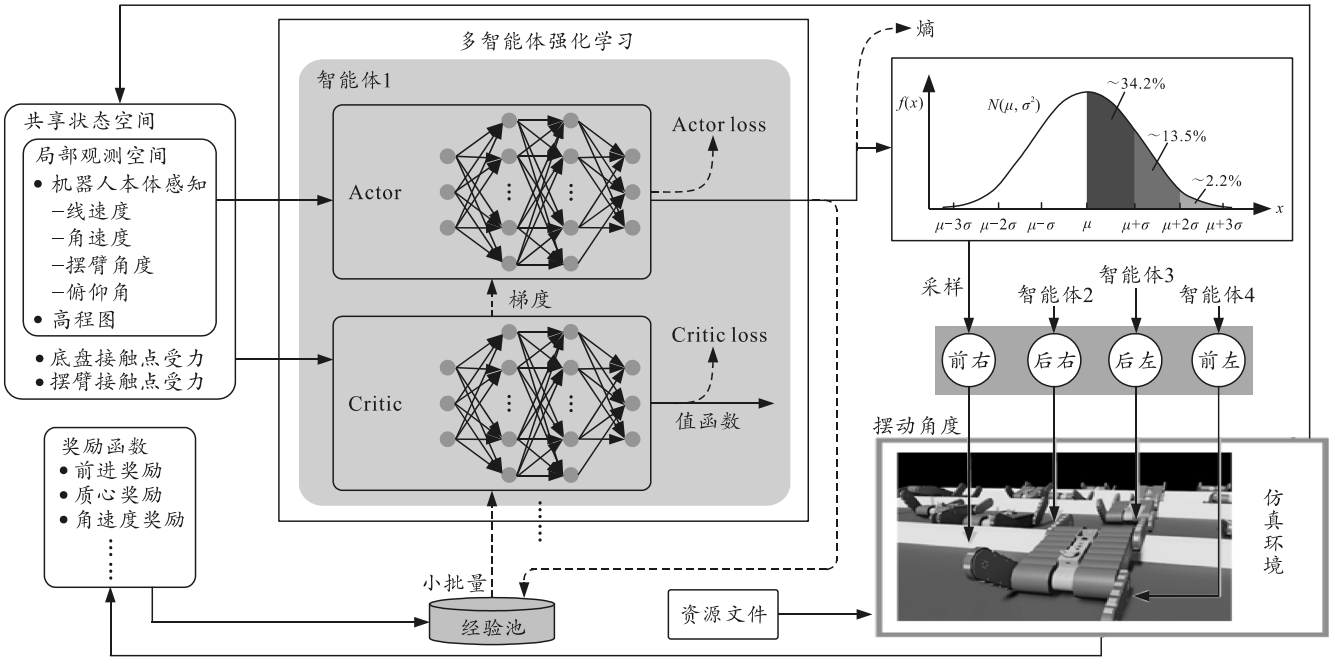

图 1　基于多智能体强化学习的摆臂履带机器人运动自主规划框架

除了可观测到的局部状态，还有许多不能观测的状态，但却与评价该状态的价值密切相关。所有智能体共享的状态空间 $\boldsymbol{S}^t$ 可以表示为：

$$s^t=\{o^t, F_{\text{contact}}, h_{\text{flipper}}\}。\tag{2}$$

式中：F_{contact} 为摆臂机器人与地面接触点支撑力的大小；h_{flipper} 为 4 个摆臂的末端到地面的距离。

图 1 右侧展示了每个智能体控制摆臂的示意，每个智能体的动作 a^t 表示为：

$$a^t=\{\Delta\theta_i\},\ i=\{rl, fr, rl, rr\}\tag{3}$$

每个智能体对应的控制摆臂应该摆动的角度，范围为[−4°, 4°]。

1.2　奖励函数设计

为了使摆臂机器人能够以尽可能平稳的姿态越过障碍物，笔者设计了 5 个不同的奖励函数。

1) 摆臂机器人前进奖励 r_{track}。

摆臂机器人以越趋近于 v_{cmd} 的速度前进获得的奖励越大。该奖励的表达式如下：

$$r_{\text{track}}=\exp\{-32(v_x-v_{\text{cmd}})^2\}。\tag{4}$$

式中 v_x 为摆臂机器人实际前进的速度。

2) 摆臂机器人停止惩罚 r_{shutdown}。

当摆臂机器人出现停止状态时，应给予一个惩罚。该奖励函数的表达式为：

$$r_{\text{shutdown}}=\begin{cases}0, & \text{if}\sum_{t=0}^{N}\Delta d_t<0\\ \sum_{t=0}^{N}\Delta d_t<0, & \text{else}\end{cases}。\tag{5}$$

式中：Δd_t 为摆臂机器人在时间步 t−1 到时间步 t 前进的距离；N 为要统计的过去时间步的个数。

3) 摆臂机器人越障平稳性奖励 r_{stable}。

在文献[2−4]的影响下，笔者也认为摆臂机器人质心相对于地面的高度对越障平稳性有着重要的影响。然而，笔者仅在明确判定为不稳定状态时对摆臂机器人给予负奖励，即当质心高度超过地形高度时对其进行惩罚。该奖励函数可以表示为：

$$r_{\text{height}}=-\max(0, p_z-\max(h_{\text{map}}))。\tag{6}$$

式中 p_z 为摆臂机器人底盘中心在世界坐标 z 轴的值。

4) 危险姿态惩罚 $r_{\text{excess\_pitch}}$ 和 $r_{\text{excess\_roll}}$。

在越障过程中，摆臂机器人并不总是以平稳的方式越过障碍物，有时甚至会出现翻车的情况，为此笔者设计了 $r_{\text{excess\_pitch}}$ 和 $r_{\text{excess\_roll}}$ 2 个奖励函数来限制底盘的位姿。当 θ_{pitch} 和 θ_{roll} 的值超过一定的阈值则认为这是一个接近危险的状态，并且给予负的奖励，这 2 个奖励可以表示为：

$$r_{\text{excess\_i}}=\begin{cases}-(\theta_i)^2, & \text{if}\theta_i>\varepsilon_i\\ 0, & \text{otherwise}\end{cases},\ i=\{\text{pitch, roll}\}。\tag{7}$$

式中 ε_i 为常数，代表机器人位姿安全的阈值。

5) 摆臂机器人角速度惩罚 $r_{\text{ang\_vel}}$。

笔者认为在越障过程中，摆臂机器人在 yz 平面内的角速度 ω_{yz} 与越障表现好坏密切相关，当 ω_{yz} 越大时意味着摆臂机器人越不稳定，应当给予机器

人惩罚；因此，$r_{\text{ang\_vel}}$ 可以表示为：

$$r_{\text{ang\_vel}}=-(\omega_{yz})^2。\tag{8}$$

1.3 策略训练方法

采用多智能体强化学习(multi-agent proximal policy optimization，MAPPO)算法[5]进行策略优化。在 MAPPO 中，每个智能体 i 都会包含策略 π_{θ_i} 和值函数 V_{ϑ_i}，同时使用中心评价函数 V_{tot} 去评价所有智能体的总预期回报。在训练期间，MAPPO 算法使用以下目标函数更新智能体策略：

$$L(\theta_i)=E_{(s,a)\sim\pi_{\theta_{\text{old}}}}\left[\min(r(\theta_i),\text{clip}(r(\theta_i),1-\varepsilon,1+\varepsilon)A^{\text{MAPPO}}(s,a)\right]。\tag{9}$$

式中：$r(\theta_i)=\pi_{\theta_i}(a_i|s)/\pi_{\theta_{\text{old}}}(a_i|s)$ 为新旧策略的概率比值；$A^{\text{MAPPO}}(\boldsymbol{s},\ \boldsymbol{a})$为多智能体优势函数，表达式为：

$$A^{\text{MAPPO}}(s,a)=r(s,a)+\gamma V_{\text{tot}}(s',a')-V_{\text{tot}}(s,a)。\tag{10}$$

式中：$r(s,a)$为即时奖励；γ 为折扣因子；s' 和 a' 分别为下一状态和联合动作。为了能准确评估在某个状态下的值函数，使用以下目标函数去更新值函数网络：

$$L(\vartheta_i)=\text{huber}(r(s,a)+\gamma V_{\text{tot}},V_{\text{tot}}(s,a))。\tag{11}$$

式中采用 huber 函数来计算损失函数。

完整的目标函数可以表示为：

$$\text{Loss}=L(\theta_i)+L(\vartheta_i)-c_{\text{ent}}\mathcal{H}(\pi_{\theta_i}(\cdot|\boldsymbol{s}))。\tag{12}$$

式中：c_{ent} 为一个超参数，用于控制策略探索和利用之间的平衡；$\mathcal{H}(\pi_{\theta_i}(\cdot|\boldsymbol{s}))$ 为策略 π_{θ_i} 的熵。

2 摆臂机器人越障仿真实验分析

2.1 实验设置

笔者在 Isaac Sim 中搭建了一个具有各种地形的世界地图。在越障开始时，摆臂机器人会随机选择一个范围为[0.2, 0.3]m/s 的初始速度，并保持这一初始速度直到任务完成。为了使学习出的策略具有泛化性，为摆臂机器人的局部观测值添加高斯噪声，环境噪声参数如表 1 所示。

表 1 环境噪声参数

字段名称	高斯噪声方差
高程图/m	0.20
3 维方向/rad	0.02
摆臂角度/rad	0.01
IMU 角速度/(rad/s)	0.20

2.2 仿真训练结果分析

2.2.1 超参数 c_{ent} 分析

超参数 c_{ent} 是用于调节熵对策略影响的系数，通过调整 c_{ent} 的值，可以控制策略在探索和利用之间的平衡。图 2 展示了不同 c_{ent} 取值对获得平均奖励的影响，从图中可以看出：c_{ent}=0.005 时获得的平均奖励更高，相较于 c_{ent}=0 时平均奖励提升了 91%，因此，在后续的实验中均使用 c_{ent}=0.005。

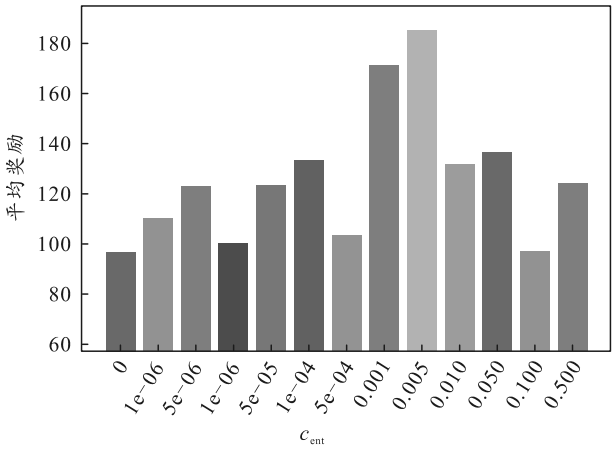

图 2 c_{ent} 不同取值获得平均奖励

2.2.2 MAPPO 训练结果分析

为了证明笔者所提方法的有效性，选用单智能体强化学习算法 PPO 作为对比方法。图 3 展示了 MAPPO 和 PPO 算法获得平均奖励函数随时间步变化的曲线。可以看出，最终收敛时 MAPPO 的平均奖励相对于 PPO 提升了约 60%。

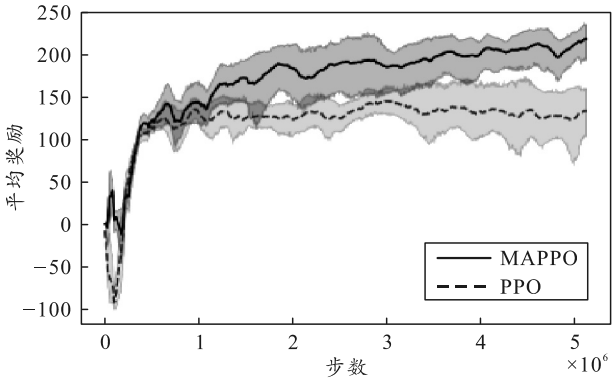

图 3 MAPPO 和 PPO 获得平均奖励函数变化曲线

2.3 仿真定量分析

为了验证笔者所提方法的有效性，在 Isaac Sim 仿真平台上进行了实验。在仿真中搭建了 40 cm 高度台阶、33°楼梯、45°楼梯和梅花桩的地形，用于对所提方法进行定量分析。台阶高度是摆臂机器人自身高度的 2 倍；33°楼梯的台阶宽度为 30 cm，高

度为 20 cm，共 5 级；45°楼梯的台阶宽度为 20 cm，高度为 20 cm，共 5 级。测试任务是让摆臂机器人以固定的速度匀速越过障碍物，实验从障碍物边缘前 1 m 处开始，到障碍物边缘后 1 m 处结束，重复进行 10 次。

为了体现笔者提出算法的效果，使用 2 个量化指标来定量分析实验结果，采用文献[6]使用的平均俯仰摆动(average pitch swing，APS)和最大摇晃角速度(maximum of angular swing，MAS)来分析越障的效果。这 2 个指标分别由下式计算：

$$\mathrm{APS}=\frac{1}{n}\sum\left|\dot{\theta}_{\mathrm{pitch}}\right|; \tag{13}$$

$$\mathrm{MAS}=\max(|\dot{\theta}_{\mathrm{pitch}}|+|\dot{\theta}_{\mathrm{roll}}|+|\dot{\theta}_{\mathrm{yaw}}|)。 \tag{14}$$

式中：n 为整个越障过程的采样次数总和；$\dot{\theta}_{\mathrm{pitch}}$、$\dot{\theta}_{\mathrm{roll}}$ 和 $\dot{\theta}_{\mathrm{yaw}}$ 为在 2 次采样期间的平均变化率，在本实验中所使用的采样频率是 10 Hz。

4 种地形的量化指标结果如表 1 所示，加粗的结果表示在同一实验条件下，不同方法中的最佳表现。表中将笔者提出的方法与 PPO 算法的结果进行了对比，展示了 10 组测试数据的平均值和方差。结果表明：笔者提出的方法在平均值上显著优于 PPO 算法，且方差更小，说明笔者提出的算法不仅效果更好，而且在部署上更具稳定性。

表 1　不同地形下仿真实验定量分析结果

测试地形	算法	APS/(rad/s)	MAS/(rad/s)	耗时/s
40 cm 台阶	Ours	**0.20±0.03**	**1.81±0.20**	**12.1±0.1**
	PPO	0.24±0.03	2.20±0.31	12.6±0.4
33°楼梯	Ours	**0.22±0.02**	**2.34±0.39**	**18.9±0.4**
	PPO	0.29±0.03	3.35±2.58	19.5±2.6
45°楼梯	Ours	**0.24±0.02**	**3.17±1.05**	**19.8±1.9**
	PPO	失败	失败	失败
梅花桩	Ours	**0.27±0.01**	**3.19±0.9**	**15.7±0.3**
	PPO	0.29±0.03	4.6±1.5	15.8±1.4

3　结论

为了解决摆臂机器人在复杂场景中的自主越障问题，笔者提出一种基于多智能体强化学习的履带机器人摆臂控制方法，使用 4 个智能体分别控制单个摆臂。同时，设计了一套兼顾底盘稳定性和摆臂动作的奖励函数，采用 MAPPO 算法进行训练。实验结果表明：与单智能体强化学习算法相比，该方法的性能提升显著，可以帮助机器人在多类复杂地形下自主调整摆臂姿态，从而实现自主越障。

参考文献：

[1] 陈柏良, 黄开宏, 潘海南, 等. 智能搜救机器人在障碍地形的自主构型规划[J]. 国防科技大学学报, 2023, 45(6): 132−142.

[2] MITRIAKOV A, PAPADAKIS P, KERDREUX J, et al. Reinforcement learning based, staircase negotiation learning: Simulation and transfer to reality for articulated tracked robots[J]. IEEE Robotics & Automation Magazine, 2021, 28(4): 10−20.

[3] MITRIAKOV A, PAPADAKIS P, GARLATTI S. Staircase traversal via reinforcement learning for active reconfiguration of assistive robots[C]//2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2020: 1−8.

[4] MITRIAKOV A, PAPADAKIS P, GARLATTI S. Staircase negotiation learning for articulated tracked robots with varying degrees of freedom[C]//2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE, 2020: 394−400.

[5] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of ppo in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611−24624.

[6] CHEN B, HUANG K, PAN H, et al. Geometry‐based flipper motion planning for articulated tracked robots traversing rough terrain in real‐time[J]. Journal of Field Robotics, 2023, 40(8): 2010−2029.