

四足机器人单目视觉循迹与强化学习步态控制方法研究

葛景仲¹ 关皓轩² 李思奇³

(1.北华航天工业学院, 河北 065000, 2.北华航天工业学院, 河北 06500, 3.北华航天工业学院, 河北 06500)

摘要:为解决四足机器人在结构化赛道环境下的自主循迹问题,提出一种“视觉循迹—强化学习步态”分层控制方法。上层以单目图像为输入,经颜色区间阈值分割得到赛道二值图,用三点模板逐行扫描左右边界并以行间连续性筛选有效行,取中线均值构造横向偏差,再由横移—转向复合比例控制律解算机体速度命令;同时以彩色路标掩膜像素数为触发量完成任务切换。下层将步态生成建模为马尔可夫决策过程,以本体感受与上层速度命令构成观测、以关节目标位置为动作,设计含速度跟踪与多项约束的加权奖励,采用近端策略优化在并行仿真中训练策略网络,再经域随机化迁移至实机以50Hz前向推理。两层以机体速度命令为唯一接口,实现感知语义与全身动力学的解耦。对现场采集的64幅图像离线复算表明:赛道中线提取成功率100%,平均有效中线点数136.2(上限150),横向偏差均值46.4px;5类彩色路标掩膜像素数最小值为3614,达触发阈值的7.2倍,检出率100%,判决裕度充分,所取控制增益与底层指令限幅相容。

关键词:四足机器人;单目视觉循迹;中线提取;强化学习;步态控制

基金项目:北华航天工业学院创新资助项目“基于四足机器人的复杂地形适应性优化及强化学习算法研究”(YKY-2026-037)

DOI:https://doi.org/10.68249/jstnei.2026.0015

0 引言

四足机器人以离散落足点获得优于轮式平台的地形通过能力,其本体构型与运动控制方法近年来发展迅速^[1]。传统方法依赖精确的动力学模型与人工设计的步态相位,参数整定繁琐且对模型误差敏感;以深度强化学习为代表的学习型方法把步态生成建模为序贯决策问题,通过大规模交互直接学习从本体状态到关节指令的映射^[2],在敏捷运动与非结构地形适应上取得了显著进展^[3-4]。

另一方面,在结构化赛道等半受控环境中沿指定路径自主行进,是四足平台进入巡检、导览等实际任务的基础能力。单目视觉是其中最经济的感知手段,但已有的视觉循迹研究多以轮式平台为对象^[5]:轮式底盘的运动学简单、指令维度低,视觉解算结果可直接作为底盘速度使用;四足平台则不同,其机体速度需由12个关节的协调运动实现,若把图像处理与步态生成放在同一控制回路中整体优化,不仅样本效率低,而且视觉参数与步态参数相互耦合,难以独立调试与复用。

针对上述问题,本文提出一种分层控制方法:上层用单目视觉解算机体速度命令,下层用强化学习策略把速度命令映射为关节指令,两层仅通过机体速度命令耦合。主要工作包括:①给出以速度命令为唯一接口的分层架构,使感知语义与全身动力学解耦;②提出基于三点模板行扫描与行间连续性筛选的赛道中线提取方法,并设计横移—转向复合比例控制律以利用四足平台的全向移动能力;③面向速度命令跟踪构造马尔可夫决策过程,给出观测、动作与加权奖励的设计以及近端策略优化的训练与部署框架;④在现场采集的图像集上对视觉层作离线复算,定量给出中线提取有效性与路标判决裕度。

1 分层控制系统总体设计

系统总体架构如图1所示,由感知层、视觉循迹层、强化学习步态层与执行层构成。视觉循迹层运行于板载工控机,以约50Hz的周期完成图像预处理、赛道分割、中线提取、路标检测与任务调度,输出机体速度命令 v_d ;强化学习步态层以 v_d 与本体感受(关节编码器、IMU)构成观测,由离线训练所得的策略网络推理出关节目标位置,再经关节PD环转为力矩下发。

作者简介:葛景仲(2002—),男,硕士研究生,研究方向为控制工程;

关皓轩(2002—),男,硕士研究生,研究方向为控制工程;

李思奇(1984—),女,讲师,硕士生导师,研究方向为控制工程。

通讯作者:葛景仲,邮箱:gejingzhong@stumail.nciae.edu.cn

分层的依据在于两者的时间尺度与信息类型不同：视觉回路处理的是米级、语义级信息，其带宽受成像与图像处理耗时限制；步态回路处理的是厘米级、动力学级信息，需要更高的更新率与更强的实时性。以机体速度命令作为唯一接口，可使任一层单独替换而不影响另一层——例如更换赛道颜色只需重新整定阈值，更换地形只需重新训练策略。层间与底层控制器之间的数据交换采用轻量通信与编组库 LCM[6]，其无中心节点、消息类型静态生成、编解码开销小，适合单机内高频率、低时延的控制数据流。

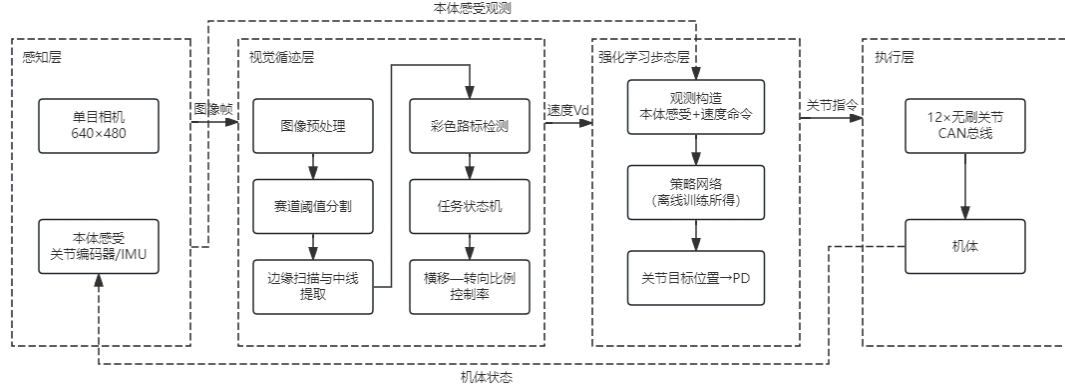


图 1 分层控制系统总体架构

2 单目视觉循迹算法

2.1 赛道阈值分割

赛道为深色地面上铺设的白色路面。原始图像先缩放至 $W \times H = 400 \times 300$ 以降低运算量，再作 5×5 中值滤波抑制织物褶皱与噪点，随后在 BGR 空间以颜色区间阈值分割为单通道二值图

$$B(u, v) = \begin{cases} 255, T_{\min, c} \leq I_c(u, v) \leq T_{\max, c}, \forall c \in \{B, G, R\} \\ 0, \text{其他} \end{cases} \quad (1)$$

式中 $I_c(u, v)$ 为像素 (u, v) 在通道 c 上的灰度， $T_{\min, c}$ 与 $T_{\max, c}$ 为该通道的下、上阈值。区间阈值的运算量为 $O(WH)$ 且无迭代，便于以固定周期运行；其代价是对光照敏感，因此系统预留了一条 UDP 旁路，可在线整定阈值并回传原图与二值图。

2.2 边缘扫描与中线提取

为抑制孤立噪点被误判为边界，采用三点模板：仅当当前像素属于背景、而沿扫描方向间隔 3 与 5 个像素处均属于路面时，才认定该点为边界（图 2）。左、右边界分别由自左向右与自右向左的首个满足点给出

$$\begin{aligned} L(v) &= \min\{u | B(v, u) = 0 \wedge B(v, u + 3) = B(v, u + 5) = 255\} \\ R(v) &= \max\{u | B(v, u) = 0 \wedge B(v, u - 3) = B(v, u - 5) = 255\} \end{aligned} \quad (2)$$

第 v 行的中线位置取两侧边界的中点

$$M(v) = \frac{L(v) + R(v)}{2} \quad (3)$$

由于透视投影下远端赛道所占像素少、噪声大，且机体前方近端区域对控制更为关键，仅取图像上半部参与计算；同时要求相邻两行的左右边界位置变化均小于 4 像素，以剔除因路标条带、阴影或分岔口造成的边界跳变。有效行集合 Ω 及横向偏差分别为

$$\Omega = \left\{ v \mid 0 \leq v < \frac{H}{2}, |L(v) - L(v + 1)| < 4, |R(v) - R(v + 1)| < 4, L(v)R(v) \neq 0 \right\} \quad (4)$$

$$\bar{u} = \frac{1}{N} \sum_{v \in \Omega} M(v), e = \bar{u} - u_d \quad (5)$$

式中 u_d 为图像水平中心附近的参考列。文献[5]在轮式小车上以阈值法提取引导线并辅以鲁棒线性回归拟合；本文同样以阈值分割为基础，但不做全局参数拟合，仅在行方向作一次线性扫描并以行间连续性作局部一致性检

验，单帧最坏情况下的像素比较次数约为 2.4×10^5 ，无需曲线拟合或神经网络推理，可与步态策略推理共享同一块无 GPU 的工控板。

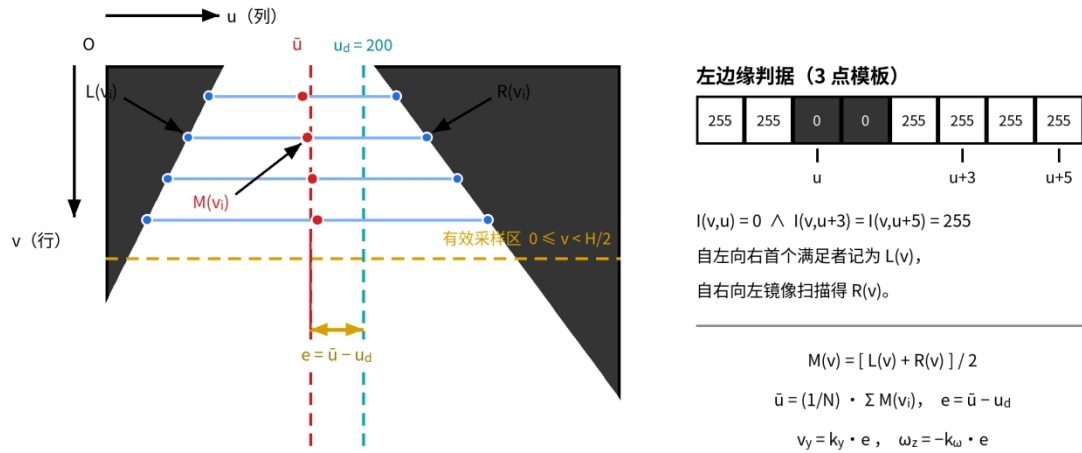


图 2 边缘扫描、中线提取与横向偏差量化原理

2.3 横移—转向复合比例控制律

四足机器人具备全向移动能力，其机体速度命令包含前进速度、横移速度与偏航角速度。本文对横向偏差同时施加横移与转向两路比例作用

$$\begin{aligned} v_{x,d} &= v_0 \\ v_{y,d} &= k_y e \\ \omega_{z,d} &= -k_\omega e \end{aligned} \quad (6)$$

式中 v_0 为巡航速度， k_y 、 k_ω 分别为横移与转向增益。横移分量在不改变机体朝向的前提下快速消除侧向位置误差，转向分量则调整航向使中线趋于对正，二者互补可在窄道与连续弯道上同时抑制超调与稳态偏置——这正是四足平台相对轮式平台的优势所在。实现中取 $v_0=0.2\text{m/s}$ 、 $k_y=0.001\text{m} \cdot \text{s}^{-1} \cdot \text{px}^{-1}$ 、 $k_\omega=0.008\text{rad} \cdot \text{s}^{-1} \cdot \text{px}^{-1}$ ，并对输出限幅。

2.4 彩色路标检测与任务切换

赛道任务由横铺于路面的彩色胶带条指示。对预处理后的图像按目标颜色阈值区间生成掩膜 B_c ，统计其前景像素数并与阈值比较，即得触发判据

$$N_c = \sum_{v=0}^{H-1} \sum_{u=0}^{W-1} \frac{B_c(u,v)}{255}, N_c \geq N_{th} \quad (7)$$

该判据只依赖像素计数，不涉及连通域分析或形状匹配，计算量与分割本身同阶。任务被触发后进入相应的瞬态子过程，由独立定时线程按预设时长推进子过程序号，主循环据此改写速度命令与站高、姿态偏置；子过程结束后自动返回循迹状态。各路标对应的任务映射见表 1。

表 1 彩色路标与任务动作映射

路标颜色	目标任务	机体动作	瞬态时长/s
蓝	限高通过	站高偏置 $-0.10\text{m} \rightarrow -0.15\text{m}$ 下蹲通过后复位	9
绿	分岔左转	先沿中线对正，再叠加 $\omega_z = +0.15\text{rad/s}$	4
红	分岔右转	先沿中线对正，再叠加 $\omega_z = -0.15\text{rad/s}$	4
棕	右侧交互	减速至 $0.06\text{m/s} \rightarrow$ 横滚 $-0.4\text{rad} \rightarrow$ 复位循迹	4
紫	左侧交互	减速至 $0.06\text{m/s} \rightarrow$ 横滚 $+0.4\text{rad} \rightarrow$ 复位循迹	7

3 基于强化学习的步态控制策略

3.1 马尔可夫决策过程建模

把步态生成建模为马尔可夫决策过程。观测由本体感受与上层速度命令拼接而成

$$o_t = [\omega_b, g_b, v_d, q - q_0, \dot{q}, a_{t-1}]^T \quad (8)$$

式中 ω_b 与 g_b 分别为机体系角速度与重力方向的机体系投影, v_d 为视觉循迹层输出的速度命令, q 与 $\dot{q}$ 为关节位置与速度, q_0 为默认站立关节角。将速度命令纳入观测, 使同一策略能够跟踪任意合法的速度指令, 而不必为每种任务动作单独训练策略; 采用本体感受而非外感受构成观测, 则可避免视觉噪声直接进入高频步态回路^[4]。动作 a_t 取 12 个关节的目标位置偏移量, 经关节 PD 环转为力矩

$$\tau_t = k_p(\sigma a_t + q_0 - q_t) - k_d \dot{q}_t \quad (9)$$

其中 σ 为动作缩放系数, k_p 、 k_d 为关节刚度与阻尼。以关节目标位置而非力矩作为动作, 相当于在策略与执行器之间引入一层低通结构, 可显著降低策略输出的高频抖动。MDP 各要素的组成与维度见表 2。

表 2 强化学习步态策略的马尔可夫决策过程要素

要素	组成	维度
观测 o_t	机体角速度、重力方向投影、速度命令、关节位置偏差、关节速度、上一步动作	3+3+3+12+12+12 = 45
动作 a_t	12 个关节的目标位置偏移量, 经式(9)转为关节力矩	12
奖励项 Φ	线速度跟踪、偏航角速度跟踪、机体姿态、关节力矩、足端滑移、动作平滑	6 项加权
策略输出频率	与视觉层速度命令同步更新, 策略以固定周期前向推理	50Hz

3.2 奖励函数设计

奖励由速度跟踪主项与若干正则项加权构成

$$r_t = \sum_{i \in \Phi} w_i r_i, r_v = \exp\left(-\frac{\|v_{xy,d} - v_{xy}\|^2}{\sigma_v^2}\right) \quad (10)$$

式中 Φ 为奖励项集合, w_i 为权重。主项 r_v 与偏航跟踪项均取指数型, 使奖励在命令附近平滑饱和而不鼓励过冲; 正则项分别惩罚机体姿态倾斜、关节力矩、足端滑移与相邻两步动作差分, 用于抑制训练早期常见的高频振荡与滑步。速度命令在训练中按均匀分布随机采样, 其范围覆盖 2.3 节控制律的限幅区间, 从而保证策略在部署时始终工作在训练分布之内。

3.3 策略训练与实机部署

采用近端策略优化算法^[7]训练策略网络与价值网络, 其截断代理目标为

$$L^{\text{CLIP}}(\theta) = E_t[\min(\rho_t(\theta)\hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t)], \rho_t(\theta) = \frac{\pi_\theta(a_t|o_t)}{\pi_{\theta_{\text{old}}}(a_t|o_t)} \quad (11)$$

式中 ρ_t 为新旧策略在同一观测下的概率比, 式中带帽符号为广义优势估计, 由时序差分残差按指数加权得到, ε 为限制单次更新幅度的截断系数。训练在并行仿真环境中进行, 大量并行实例可将策略训练时间压缩至分钟量级^[8]; 同时对地面摩擦系数、机体质量与惯量、关节增益、观测噪声与控制时延作域随机化, 以缩小仿真与实机的动力学差异。训练完成后冻结策略参数并导出, 在板载工控机上以与视觉层相同的周期作前向推理, 整体训练与部署框架如图 3 所示。

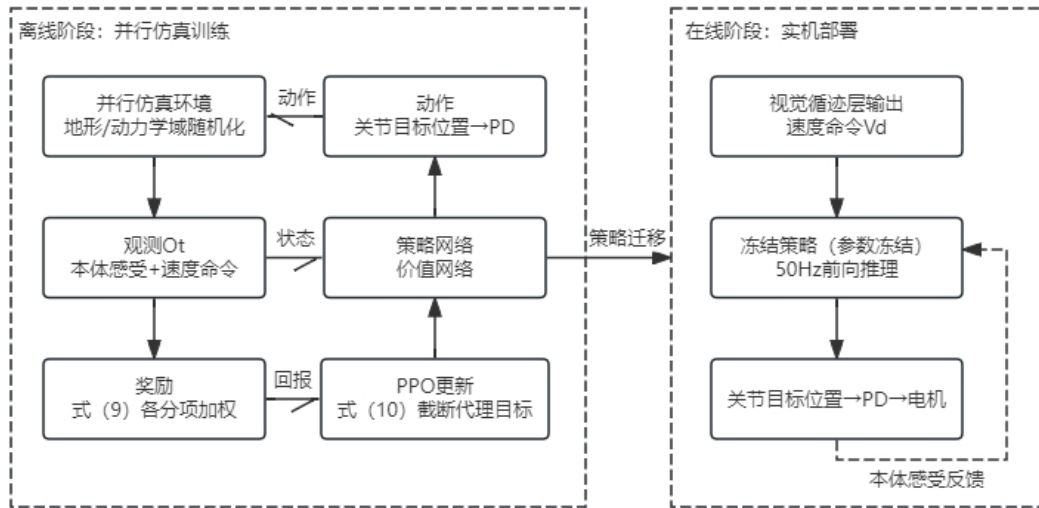


图 3 强化学习步态策略的训练与部署框架

4 算法总体流程

系统的完整运行流程如图 4 所示。主循环依次完成图像采集与预处理、任务瞬态判定、彩色路标检测、赛道分割与中线提取、比例控制律解算、策略推理与关节指令下发；处于任务瞬态或检出路标时，速度命令改由状态机按子程序序号给出，其余环节不变。由于策略始终以速度命令为输入，任务动作的插入不需要改动步态层，这正是分层接口带来的直接收益。

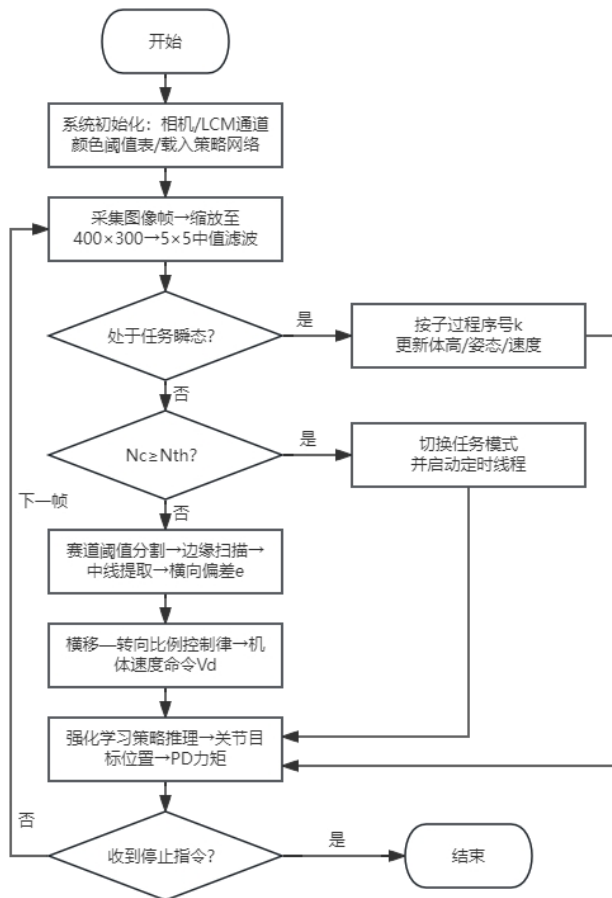


图 4 视觉循迹与运动控制总流程

5 实验与结果分析

实验在铺设白色路面的室内赛道上进行,沿途横铺蓝、绿、红、棕、紫五种颜色的胶带条作为任务路标。由机器人在实际行进过程中采集图像并同步保存对应的分割二值图,共获得赛道图像 36 幅、彩色路标图像 28 幅,典型场景如图 5 所示。将全部 64 幅二值图重采样至 400×300 的处理尺寸,按 2.2 与 2.4 节的算法离线复算,以评估视觉层输出的有效性与判决裕度。

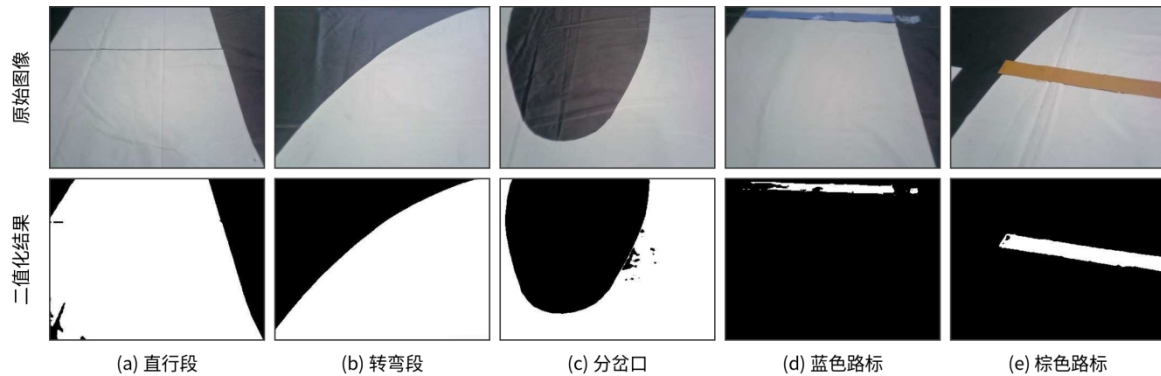


图 5 典型场景下的赛道分割与彩色路标检测结果

在 36 幅赛道图像上,中线提取的成功率为 100%,每帧有效行数 N 的均值为 136.2、中位数为 146、最小值为 79 (理论上限 150),表明三点模板与行间连续性筛选在弯道、分岔口及织物褶皱造成的边界不规则处仍能保留足够的中线样本;横向偏差 $|e|$ 的均值为 46.4px、最大值为 127.8px,代入式(6)对应的横移量与偏航角速度最大值分别为 0.128m/s 与 1.022rad/s,均落在限幅范围之内,说明所取增益与速度命令的可执行区间相容,不会把步态层推出训练分布。

彩色路标的检测结果见表 3。五类路标掩膜像素数的最小值为 3614,为触发阈值 $N_{th}=500$ 的 7.2 倍,全部 28 帧均被正确触发,检出率 100%。如此大的裕度意味着阈值可在较宽范围内取值而不影响判决结果,系统对路标距离与拍摄角度的变化具有较好容忍度。需要说明的是,上述结果是在离线图像集上对视觉层感知与解算环节的复现,用以刻画其输出的有效性与判决裕度,并不等价于闭环跟踪误差;步态层策略的训练收敛性与实机跟踪精度的定量评估将另文给出。

表 3 彩色路标掩膜像素数统计 (处理尺寸 400×300 , $N_{th} = 500$)

路标颜色	帧数	N_c 最小值	N_c 均值	N_c 最大值	裕度 N_{cmin}/N_{th}	检出率/%
蓝	8	3 636	8 247	12 918	7.3	100
绿	4	3 614	9 101	16 321	7.2	100
红	4	4 366	9 721	14 620	8.7	100
棕	7	6 026	11 231	15 905	12.1	100
紫	5	4 614	10 445	17 277	9.2	100
合计	28	3 614	9 749	17 277	7.2	100

6 结论

本文提出并实现了四足机器人“视觉循迹—强化学习步态”分层控制方法。上层给出了三点模板边缘扫描与行间连续性筛选的中线提取算法及横移—转向复合比例控制律,下层给出了以速度命令为条件的马尔可夫决策过程建模、加权奖励设计与近端策略优化训练部署框架,两层仅通过机体速度命令耦合,使感知语义与全身动力学得以解耦。在现场采集的 64 幅图像上离线复算表明,中线提取成功率 100%、平均有效行数 136.2,五类路标检出率 100%、最小判决裕度达 7.2 倍,视觉层输出与底层指令区间相容。后续工作将补充策略训练收敛曲线与实机闭环跟踪精度的定量评估,并把当前由定时线程驱动的任务瞬态过程替换为基于本体状态反馈的闭环判据。

参考文献:

- [1] 王旭.四足机器人运动及稳定控制关键技术综述[J].吉林大学学报(工学版),2025,55(5):1483-1496.
- [2] 唐力强,韦宇轩,胡海东,等.基于深度强化学习的月面四足机器人控制方法综述[J].空间控制技术与应用,2025,51(5):1-15.
- [3] HWANGBO J, LEE J, DOSOVITSKIY A, et al. Learning agile and dynamic motor skills for legged robots[J].Science Robotics,2019,4(26):eaau5872.
- [4] MIKI T, LEE J, HWANGBO J, et al. Learning robust perceptive locomotion for quadrupedal robots in the wild[J].Science Robotics,2022,7(62):eabk2822.
- [5] 岳俊峰,李秀梅.基于机器视觉的智能小车自动循迹及避障系统[J].杭州师范大学学报(自然科学版),2020,19(2):200-207.
- [6] HUANG A S, OLSON E, MOORE D C. LCM: lightweight communications and marshalling[C]//2010 IEEE/RSJ International Conference on Intelligent Robots and Systems.Taipei (Taiwan Region, China):IEEE,2010:4057-4062.
- [7] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL].(2017-08-28)[2026-08-18].<https://arxiv.org/abs/1707.06347>.
- [8] RUDIN N, HOELLER D, REIST P, et al. Learning to walk in minutes using massively parallel deep reinforcement learning[C]//Proceedings of the 5th Conference on Robot Learning.London:PMLR,2022:91-100.

Monocular Vision-Based Line Following and Reinforcement Learning Gait Control for a Quadruped Robot

Jingzhong Ge¹, Haoxuan Guan², Siqi Li³

1 School of Electronic and Control Engineering, North China Institute of Aerospace Engineering, Langfang 065000, China

2 School of Electronic and Control Engineering, North China Institute of Aerospace Engineering, Langfang 065000, China

3 School of Electronic and Control Engineering, North China Institute of Aerospace Engineering, Langfang 065000, China

Abstract: A hierarchical "visual line following - reinforcement learning gait" control method is proposed for autonomous line following of a quadruped robot in a structured track environment. In the upper layer, a monocular image is segmented by a colour-interval threshold; the left and right track borders are located by a three-point template row scan, valid rows are selected by inter-row continuity, and the mean of the extracted centre line gives the lateral deviation, which drives a combined lateral-shift and yaw proportional control law that outputs a body velocity command; the pixel count of each colour landmark mask triggers task switching. In the lower layer, gait generation is formulated as a Markov decision process whose observation consists of proprioception and the velocity command and whose action is the joint target position; a weighted reward combining velocity tracking with several regularisation terms is designed, and the policy network is trained by proximal policy optimisation in parallel simulation and transferred to the real robot with domain randomisation for 50 Hz forward inference. The two layers are coupled only through the body velocity command, which decouples perceptual semantics from whole-body dynamics. Offline reproduction on 64 images captured on site gives a 100% centre-line extraction rate, 136.2 valid centre points per frame on average, a mean lateral deviation of 46.4 px, and a minimum landmark mask area of 3614 pixels, i.e. 7.2 times the trigger threshold, yielding a 100% detection rate with sufficient decision margin.

Keywords: quadruped robot; monocular visual line following; centre line extraction; reinforcement learning; gait control