文章编号:1671-4598(2020)01-0105-06 DOI:10.16526/j.cnki.11—4762/tp.2020.01.023 中图分类号:TP391 文献标识码:A

基于分层学习的四足机器人运动自适应控制模型

崔俊文,刘自红,石 磊,刘福强,乐 玉
(西南科技大学 制造科学与工程学院,四川 绵阳 621010)

摘要:针对四足机器人面对腿部损伤无法继续有效自主运作的问题,提出一种基于分层学习的自适应控制模型;该模型结构由上层状态策略控制器(SDC)和下层基础运动控制器(BDC)组成;SDC对机器人腿部及姿态进行决策并选择运动子策略,BDC子运动策略表达该状态下机器人的运动行为;在Unity3D中构建反关节多自由度的四足机器人,训练多种腿部受损状况的BDC子运动策略,BDC成熟后20 s周期随机腿部受损并训练SDC;该模型控制流程为SDC监测机器人状态,激活BDC策略,BDC输出期望关节角度,最后由PD控制器进行速度控制;其实现机器人在腿部受损后自我适应继续保持运作;仿真与实验结果表明,该控制模型能在机器人损伤后能自我快速、稳定调整运动策略,并保证运动的连贯性及柔和性。

关键词:分层学习;深度强化学习;四足机器人;部分马尔可夫决策;步态控制;机构失效

A Quadruped Robot Motion Adaptive Model Based on Hierarchical Learning

Cui Junwen, Liu Zihong, Shi Lei, Liu Fuqiang, Yue Yu
(Southwest Unversity of Technology and Science, Mianyang 621010, China)

Abstract: Aiming at the problem that the quadruped robot can not continue to operate effectively and independently after the leg accidently be damaged, the adaptive control model based on hierarchical learning is proposed. The model structure consists of an upper state policy controller (SDC) and a lower base motion controller (BDC). The SDC estimates the expected motion sub-strategy of the robot's legs and posture, and the BDC sub-motion strategy is activated to control the robot to express the athletic behavior. Damage to the robot is manifested in the complete loss of athletic ability in any leg. The adaptive control of the model is reflected in the robot's self-adjusting strategy after the leg fails. In Unity3D, a four-legged robot with anti-joint multi-degree of freedom is built. The SDC monitors the state of the robot and adjusts the strategy. The BDC output gives the joint PD controller speed control. Simulation and experimental results show that the model shows a fast and stable effect on the robot's self-adjusting motion strategy after the legfails.

Keywords: hierarchical learning; deep reinforcement learning; quadruped robot; partially observable markov decision prosess; gait control; institutional failure

0 引言

在自然界中,多足生物可以在腿部受伤失效后,快速调整姿态和运动步态,继续保持一定的运动状态前进,如图1。因此,在实际的应用环境中也要求机器人具有应对自我损伤,改变原有的运动方式的能力。

以四足机器人为例,正常运动时,可以有多种步态控制方法实现前进、转向、跳跃等多种运动技能。然而在单腿失效后,其腿部运动布局变化及重心偏移的影响导致控制策略复杂性提升,快速适应及调整的控制是较为复杂的。在文献[1]中,提出依据行为空间得到价值直觉,指导试错学习算法实现局部关节失效的六足机器人和机械臂恢复运动技能。Bongard在文献[2]中提出连续自我建模恢复算法,产生替代步态的方式。

图1 单腿失效狗和四足机器人

近年来,深度强化学习(Deep Reinforcement Learning)在无经验的条件下,在虚拟器中让机器人学习各种运动技巧取得了巨大的进步[3-5]。在文献[6]中,迪士尼的研究人员构建了在实际环境中训练单腿或多腿机器人系统。

收稿日期:2019-06-17; **修回日期**:2019-07-05。
基金项目:四川省大学生创新创业训练项目基金(S201910619035)。
作者简介:崔俊文(1997-),男,安徽阜阳人,本科,主要从事机器人智能控制方向的研究。

然而其结果仍是较为简单的单一策略。对于一些复杂的多要求的运动控制，则不能满足。分层学习将控制策略结构化，使得任务复杂度降低。目前的分层学习，仍没有可通用的模型，一般依据任务特点采用人工分层方式。在文献[7]中使用高低级的控制器有效控制双足类人模型的行走于小道和踢足球。分层目的将基本运动中枢和调节器分开，达到较好运动策略。然而，对于状态空间因腿部失效维度降低，采用串级结构则是失效的。文献[8]中，Google 团队将直接从图像像素到空间机械臂运动抓取进行端到端训练，CNN 和电机控制输出层网络相合，达到三维空间中抓取多形态物体的目的。文献[9]中提出分层 DQN，让高级策略决定内在目标，低级策略满足给定目标完成 Atari 游戏。

本文提出采用深度强化学习（Deep Reinforcement Learning）对四足机器人单一向前运动进行连续控制，以此为子策略构建分层学习框架形成自适应控制模型。框架分为状态评估控制器（Status Decision Controller）和基础运动控制器（Basic Dynamic Controller），分别采用 Double－DQN（Double Deep Q－Network）[10] 和 PPO（Proximal Policy Optimization）[3] 算法实现。状态评估策略监测机器人状态，相应激活子策略进行运动控制，该过程不受状态空间部分信息失效而影响。该方法成功实现机器人在腿部受损失效后，仍保持有效运作，并保证整个过程快速响应和稳定。

1 分层学习

1.1 概述

根据四足机器人要求在状态空间维度变化的条件下实现自适应控制问题，提出由状态判断控制器（Status Decision Controller）和多个基础运动控制器（Basic Dynamic Controller）组合而成的分层学习结构，系统如图 2 所示。SDC 要求每个时间步都进行状态决策，并以 3 Hz 向子策略激活单元发送最优的运动方式。由于构建的机器人及电机运动模型要求 BDC 的控制频率也为 3 Hz，相同决策频率保证机器人实现基本运动切换连贯性。SDC 状态决策后估计最优 BDC，BDC 控制机器人以期望目标运作。

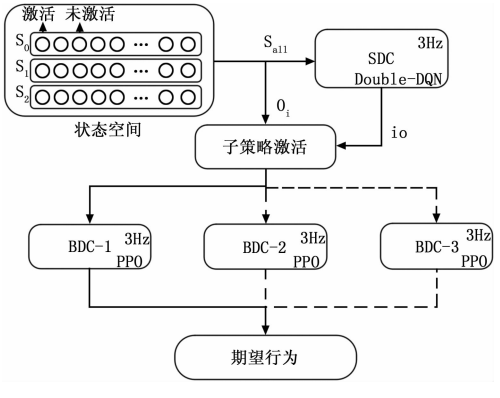

图 2 分层学习系统结构

SDC 功能特性只要求其离散决策，采用 Double－DQN 算法优化策略。SDC 输入为全部状态单元 S_{all}，未激活的状态单元则为 $o_x = 0, o_x \in S_{all}$。将 SDC 策略输出行为最高估值的索引 i_o 传递给策略激活单元。策略激活单元将索引在固定序列上匹配到合适的子策略，SDC 以完成相对最大化 Q 值这一过程作为收敛目标。

BDC 作为基础运动控制器，主导基础运动控制。由于基本运动控制要求的连续性，采用近端策略优化（PPO）算法，其在文献[3]证明其在连续运动控制上所达到的较好的效果。由状态决策单元获取目标索引 i_o，整理状态空间 $O_i, i_o \in 0,1,2$，输出行为 a_t 作为腿部运动关节的 PD 控制器输入值。BDC 各个单一策略提前进行训练，满足期望累积奖励后，再进行 SDC 训练。使用该种分层学习模型，将复杂运动策略结构化，简化神经网络深度及策略收敛的复杂性，摆脱腿部失效后状态空间维度变化所带来的影响，同时各部分相互独立，不受单一策略变动而影响整体。

1.2 状态判断控制器（Status Decision Controller）

假设构建机器人所有可能运动状态所需的状态空间为 S_{all}，并依据所有 BDC 策略要求的维度确定 A_s，构建价值函数为 $Q(s,a;\theta_t)$，其中 $s \in S_{all}, a \in A_s$，θ 表示策略参数集。当前奖励 r_{t+1} 子运动决策在环境中交互通过公用的奖励策略式（7）所述 R_{t+1}，即：

$$r_{t+1} = \left(R_{t+1} - \frac{R_{\max}}{2}\right)w_s \tag{1}$$

其中：w_s 为奖励权重，实验测试时 $w_s = 0.8$，$R_{\max}$ 为 BDC 策略成熟后单个时间步可获得最大直接奖励。由此可得 Double－DQN 的单时间步损失为 $L(\theta_t)$：

$$Y_t \equiv r_{t+1} + \gamma Q(S_{t+1}, \operatorname{argmax}_a Q(S_{t+1}, a;\theta_t);\theta'_t) \tag{2}$$

$$L(\theta_t) = (Y_t - Q(S_t, a;\theta_t))^2 \tag{3}$$

其中：$\gamma \in [0,1]$ 为折扣因子，经实验选取为 0.9。学习速率设定值为 0.001。SDC 策略 φ 则存在最优函数，需最大化这一过程：

$$Q^*(s,a) = \max_\varphi Q^\varphi(s,a) \tag{4}$$

最佳策略是由行动的最高估值即：

$$\varphi(s) = \operatorname{argmax}_a(Q^*(s,a)) \tag{5}$$

在该问题中策略激活单元的最佳行为正是最高估值的索引 $i_o = \varphi(s)$，i_o 则代表 BDC 激活的对应序列值。该过程使用神经网络表示 SDC 策略 φ 输出最高估值。SDC 的策略训练较 BDC 在时序上落后，其需要有足够的经验累积，每个时间步不断刷新经验累积内容。

1.3 基础运动控制器（Basic Dynamic Controller）

1.3.1 部分马尔科夫决策（POMDP）

BDC 为多个运动策略单元组成。每个单元代表某种状态下的控制策略，并将控制过程制定为部分马尔科夫决策（POMDP）。受损其被表述为组员 $(S, A_B, \tau, R, \Omega, \gamma)$，其中 S 为状态空间，A_B 行为空间，τ 代表系统动力，R 为奖励函数，Ω 代表概率观察函数，$\gamma \in [0,1]$ 为折扣因子。

由于状态部分可观测，有状态可观测集 O，而非 S，可得 $o \in O$。采用 PPO 算法优化 BDC 策略 $\vartheta: O \rightarrow A_B$，由

此可以计算价值策略如下：

$$V^{\vartheta}(o) = E[\sum_{t=0}^{T}\gamma^{t}R(o_t,a_t) \mid \vartheta] \tag{6}$$

其中：T 表示总时间步。$R(o_t,a_t)$ 表示训练体在给定目标下执行动作获取的反馈函数。该函数由人为依据训练目的而设定为（7）。

BDC 所可能包含的部分状态空间 O 及行为空间 A_B．如表 1 所示。每个关节运动范围及方向如图 3。以正常策略为例，包含身体的欧拉角，四组腿关节的相对角度，共 11 个信息。由于底层采用 PD 控制，为简化复杂性，未涉及速度和加速度信息。行为空间 A_B 作为四足机器人关节的运动期望，PD 进行速度控制，在关节角度上进行范围限制。在单腿失效后，失效腿部异侧引入摆动关节以调节身体平衡，失效腿状态单元不作为输入量，此时的状态量维度为 10。

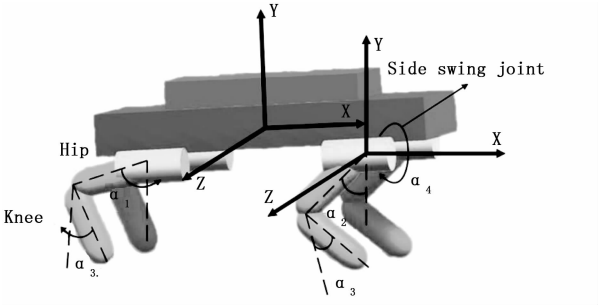

图 3　四足机器人关节运动示意图

表 1　状态及行为空间范围

名称	范围(deg)
前腿 Hip (α_2)	$-50^{\circ}\sim-10^{\circ}$
后腿 Hip (α_1)	$-80^{\circ}\sim-30^{\circ}$
摆动关节(Side swing joint) (α_4)	$-30^{\circ}\sim30^{\circ}$
前腿 Knee (α_3)	$60^{\circ}\sim100^{\circ}$
后腿 Knee (α_3)	$60^{\circ}\sim100^{\circ}$
行为空间 A_B	$-25.30^{\circ}\sim25.30^{\circ}$

奖励策略 R 的构建，以机器人快速向前运动为运作目标。构建机器人坐标系如图 3。以四足机器人在 x 轴运动方向速度 V_x 为参考的主要部分，以机器人身体在水平面的稳定性为次要部分，摄入参数为身体的欧拉角 $\alpha_x,\alpha_y,\alpha_z$。构建奖励函数为

$$R(o_t,a_t) = w_1V_x - \mid w_2\alpha_x \mid - \mid w_3\alpha_y \mid - \mid w_4\alpha_z \mid \tag{7}$$

其中：w_1,w_2,w_3,w_4 为各参量权重。测试时设定速度经验权重 $w_1=60$。四足机器人可以稳定前行的一种主要表现是其运动时姿态趋于平衡，所以为规避机器人局限于某种前进策略，需要降低训练时出现不稳定姿态的概率。在机器人 x 轴与 y 轴方向上设置欧拉角经验权重 $w_2=w_3=0.5$。机器人运动前进的方式未知，某些暂时性姿态有可能让未来动作达到更大奖励，而该姿态却可能带来较小的奖励，如跳跃。为解决这种矛盾，将机器人的俯仰姿态 z 轴权重设置 $w_4=0.2$。设置权重这一线性的奖励函数，在实际训练过程中呈现较好效果。

1.3.2　近端策略优化（PPO）

使用策略梯度算法对学习率调整至关重要，同时策略梯度算法对每个数据样本执行梯度跟新，具有高样本复杂度。为解决此问题，在 PPO 算法中让 $r_t(k)$ 表示概率比为：

$$r_t(k) = \frac{\vartheta_k(a_t \mid s_t)}{\vartheta_{k_{old}}(a_t \mid s_t)} \tag{8}$$

式中，k 表示策略 ϑ 参数集，优化目标为：

$$L(k) = E[\min(r_t(k)\hat{A}_t, clip(r_t(k),1-\varepsilon,1+\varepsilon)\hat{A}_t)] \tag{9}$$

其中超参数 $\varepsilon=0.2$，$\hat{A}_t$ 为广义优势，通过 $clip(r_t(k),1-\varepsilon,1+\varepsilon)$ 限制概率比，来限制策略更新幅度，解决更新梯度大找不到梯度下降策略，幅度过小无法收敛的问题。

使用神经网络表示策略 ϑ，通过 PPO 算法来求解 POMDP。即 ϑ^* 求解最优的参数集 k^*，表达为

$$k^* = \mathrm{argmax}_k V^k(o) \tag{10}$$

PPO 算法基于 Actor－Critic 算法实现，该算法相对传统的策略梯度算法收敛更快。Actor 依据概率选择行为，Critic 根据行为得到奖励优势决定 Actor 收敛方向。Actor 和 Critic 均有两个神经网络构成。Actor 网络有两个隐含层，神经元数量依据经验确定，设置第一个隐含层神经元数量为状态空间维度 $n_{O_i}\times20$，第二层为行为空间维度 $n_{A_i}\times10$，由 Critic 作用可知其策略相对较简单，设置一层隐含层，维度为 $n_{O_i}\times20$。Critic 要求相对于 Actor 更快收敛，依据实验分别设定学习率为 0.000 09 和 0.000 18。两组神经网络结构如图 4。

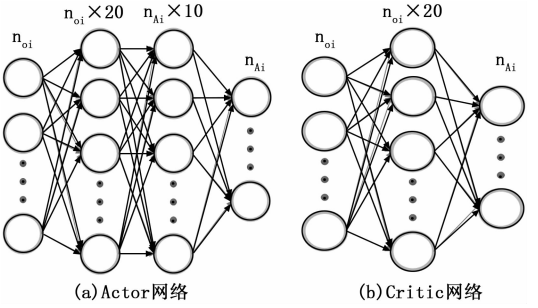

图 4　Actor－Critic 神经网络结构

1.4　算法实现过程

该分层学习模型要求在运动策略成熟后再进行上级策略训练。首先对机器人各种腿部受损状况分别训练 BDC 运动策略，主要针对两个前腿受损问题进行训练及仿真。待 BDC 全部收敛后，引入 SDC 获取全部状态空间 S_{all}，在随机模式下激活 BDC 与环境交互，依据奖励策略（1）对 SDC 策略进行更新。算法实现过程概括为算法 1。

算法 1：分层学习训练算法

随机初始化 SDC 策略参数集 θ

随机初始化 BDC 策略参数集 $k[3]$

初始化运动模式 $i=0, i\in 0,1,2$

循环 $i!=3$

遍历 $step=0,1,2\ldots T_{BDC}$ DO

与 Unity3D 交互得到 N 个批量：$N\times\{O_i,A_i,R_i\}$

依据 $Max(sum(R_i))$

找到最优批次 $\{O_i, A_i, R_i\}^*$
依据目标 $L(k)$
跟新优化策略 $\vartheta_i \leftarrow \{O_i, A_i, R_i\}^*$
结束遍历
i ++
结束循环遍历 $step = 0, 1, 2 \ldots T_{SDC}$ DO
随机运动模式 i，激活 BDC 策略 ϑ_i 与 Unity3D 环境交互
获取估值最大索引 i_o
激活 BDC 策略后获取批量 $\{S_{all}, A_s, R_n\}$
将批量存入记忆中心有 $M \times \{S_{all}, A_s, R_n\}$
随机向记忆中心抽取一个批量 M_n
依据损失函数 $L(\theta)$
跟新优化策略 $\varphi \leftarrow M_n$
结束遍历

分层学习的整个决策过程是由 SDC 激活相关 BDC 开始。SDC 获取四足机器人状态集 S_{all}，策略 φ 表达该状态的估计集，获取估计集中最高估值的索引 i_o，i_o 传递进入策略激活单元，激活固定序列，策略激活单元按序列整理状态空间集 O_i，其对应 BDC 的 Actor 网络输入层，BDC 子策略 ϑ 再输出该状态的控制动作。

2 机器人环境构建

在 Unity3D 中构建反关节四足机器人系统，单个腿部系统为 3 个自由度。策略中心和机器人仿真分为两部分进行，使用 TCP 通讯连接两个系统。重复构建多个四足机器人系统，训练时储存每个机器人连续运动集，在概率 $\rho = 0.95$ 选取最优奖励集进行训练。系统构建如图 5 描述。

机器人 Hip 关节和 Knee 关节角度参考零点垂直于 x 轴。各关节角度区间跨度较大，不利于快速收敛。为使 BDC 输出的角度期望在训练初期更快表现出运动行为，对输出范围进行归一化，行为值 $a \in A_B$ 限制为 $[-0.44, 0.44]$。最终输出期望角度 $\alpha_x = a + b$，其中 b 为各关节运动范围中值，运动范围如表 1 中所示。

机器人的运作目标是四足机器人尽可能极限化向前运动的最大速率。在实验过程中，由于初始训练时，策略随机性导致机器人姿态偏离过多，均发生侧翻，摔倒，严重偏离方向等状况，这些可能性导致样本复杂度上升或增加获取非理想样本的概率，因此构建机器人辅助运动机制，对过度偏离的角度进行周期定角度矫正。并且单一限制某方向欧拉角，不能使机器人具有调整姿态的能力。在策略成熟后，撤销辅助运动机制，再进行一定时间步的训练，机器人具有保持较好的姿态调节能力并保持一定鲁棒性。运动辅助机制主要针对 x 轴和 y 轴进行矫正，y 轴角度稳定范围为 $[-2°, 2°]$，超出范围则进行周期为 6 ms，单次 0.4° 的矫正。同样的，x 轴的稳定范围为 $[-8°, 8°]$，单次 0.6° 矫正。x 轴角度受腿部运动的影响，稳定范围相对偏大，可以使其充分发挥腿部运动机能。次要的，z 轴矫正防止机器人过激运动行为导致倾角过度，z 轴稳定范围为 $[-60°, 60°]$，单次 10° 矫正。对于某些冲击动作造成姿态倾斜较大的，该运动辅助机制在物理环境中具有缓和的效果。

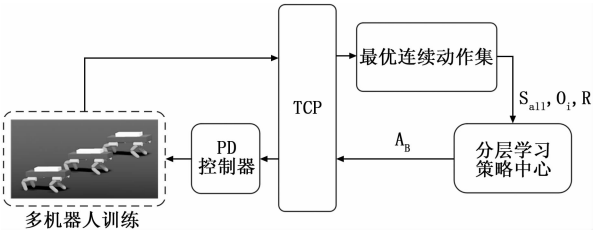

图 5 机器人训练系统构建

3 仿真与测试

在 Intel 单核 2.4 GHz 处理器上训练所有神经网络和运行机器人仿真程序，通过构建 TCP 通信网络连接分层学习策略中心和机器人仿真环境，两者间利用响应机制自动完成交互控制及信息反馈这一过程。

BDC 运动策略的训练目标，主要有正常运动与单腿失效运动。两者均构建基于姿态欧拉角的辅助运动系统。在式（7）奖励策略下，正常运动策略经过 1 200 个训练批量后，综合奖励值稳定在 40。单腿受损后的，同样经过 1 200 个训练批量，综合的奖励值稳定在 39。训练过程平均直接奖励值数据如图 6。分析可知该腿部受损的四足机器人相对未受损状况的，运动速率没有明显差异。

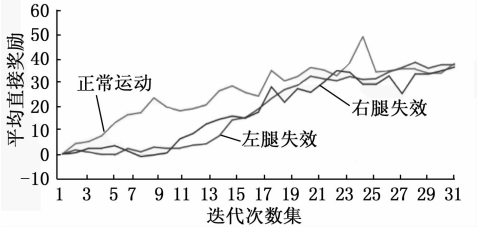

图 6 BDC 策略平均直接奖励曲线

图 7 SDC 策略平均直接奖励曲线

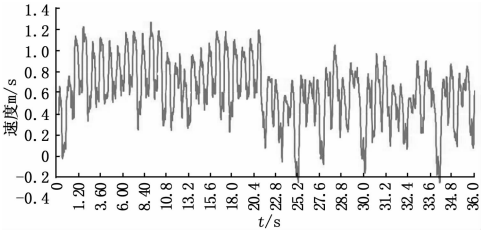

图 8 机器人运动速度曲线

SDC 在各个 BDC 单元训练结束后进行训练，设定机器人运动状态改变周期为 20 s，估计后激活 BDC 表达运动行

为。式（1）中奖励权重依据子运动的稳定奖励值确定。由于 BDC 策略存在神经网络的稀疏性，所以传入 BDC 的状态空间 O_i，不一定就是期望的状态空间，但仍可能控制相关 BDC 实现到运动技能，甚至达到更好的运动策略，所以导致 O_i 可能对应多个 BDC 策略，但不是之前预定的训练目标。将这一问题交给自适应的 Double－DQN 去解决。在单腿受损后或恢复正常状态过渡时，SDC 激活 BDC 同时要考虑过程的时机性，让这一过程快速及稳定完成。SDC 策略在经过 4 小时训练后，每个批量总奖励稳定在 14，训练数据曲线如图 7。

依据实验，机器人从正常运动向腿部受损过渡时，也近乎完美的表现出连贯性及柔和性。从图 8 机器人左腿受损后的运动速率曲线分析可得，在正常运动下，机器人保持跳跃运动行为，峰值运动速率约 1.2 m/s，在接近 18 s 时腿部受损后，仍继续保持速度峰值约 0.9 m/s 的运动行为，在运动过渡期间内，没有出现明显的停滞时间。对机器人运动行为进行连续运动逐帧截取如图 9，其俯仰姿态变化对应了其运动时的跳跃特性。在 7 帧至 8 帧的腿部受损的过渡期间，该模型自适应调整运动行为使机器人保持运动的连续性。

由此可见该模型可以实现机器人连续向前运动，并在机器人腿部受损后可以自适应控制机器人继续保持运作，并保证过程有效的连续及柔和。

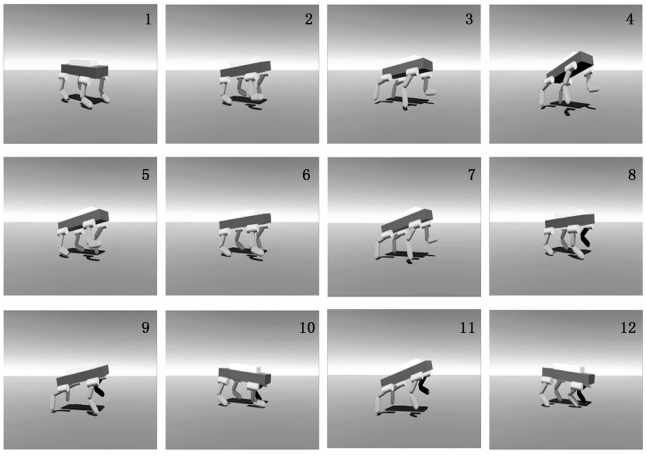

图 9　正常运动至左腿受损的截取帧

4　结束语

为解决机器人腿部受损后无法继续保持前进的问题，构建一种分层学习控制模型。SDC 和 BDC 两部分分别控制运动决策激活和基础运动控制功能，结构化控制方法，降低策略学习的复杂度。依据 Unity3D 中训练及仿真的结果，充分证实该种分层学习组合的可行性。特别地，Double－DQN 离散控制和 PPO 连续控制各发挥关键作用。让激活策略适应相关控制层，由于神经网络稀疏性，尽管可能不符合期望目标，但是由于状态空间信息局部变化，有可能让子策略表达更好的运动行为。另外，在训练中采用运动辅助系统，降低初始训练的混乱度，有效帮助快速收敛，并且该种方法不会降低和影响机器人自我姿态的调整能力。

该种分层学习方法，结构化网络关系，互相联系上下级控制策略关系，但是并没有深层次的互相影响策略执行，仍然是一种浅层次的分层方法。虽然目前的分层学习仍然没有一种通用或者具有深层次理论的方式，依然依据问题所引而确定结构及方式，但其表达的思想是大型及复杂策略解决的理想途径。未来将分类及思考任务间不同策略关系，架构相互优化的算法，深化策略联系，实现高要求的运动控制。考虑将传统控制理论相结合，将会是一个极具价值的方向。

参考文献：

[1] Cully A，Clune J，Tarapore D，et al. Robots that can adapt like animals [J]. Nature，2014，521 (7553)：503－507.

[2] Bongard J，Zykov V，Lipson H. Resilient Machines Through Continuous Self－Modeling [J]. Science，2006，314 (5802)：1118－1121.

[3] Schulman J，Wolski F，Dhariwal P，et al. Proximal Policy Optimization Algorithms [Z]. 2017，1－8.

[4] P. Lillicrap T，J. Hunt J，Pritzel A，et al. Continuous control with deep reinforcement learning [Z]. 2015

[5] Duan Y，Chen X，Houthooft R，et al. Benchmarking Deep Reinforcement Learning for Continuous Control [Z]. 2016.

[6] Ha S，Kim J，Yamane K. Automated Deep Reinforcement Learning Environment for Hardware of a Modular Legged Robot [A]. 15th International Conference on Ubiquitous Robots (UR) [C]. 2018.

[7] Peng X B，Berseth G，Yin K，et al. DeepLoco：dynamic locomotion skills using hierarchical deep reinforcement learning [J]. ACM Transactions on Graphics，2017，36 (4)：1－13.

[8] Levine S，Pastor P，Krizhevsky A，et al. Learning hand－eye coordination for robotic grasping with deep learning and large－scale data collection [A]. Springer Proceedings in Advanced Robotics 2016 International Symposium on Experimental Robotics [C]. 2017：173－184.

[9] Kulkarni T D，Narasimhan K R，Saeedi A，et al. Hierarchical Deep Reinforcement Learning：Integrating Temporal Abstraction and Intrinsic Motivation [Z]. 2016.

[10] Van Hasselt H，Guez A，Silver D. Deep Reinforcement Learning with Double Q－learning [J]. Computer Science，2015.

[11] Chernova S，Veloso M. An evolutionary approach to gait learning for four－legged robots [A]. IEEE/RSJ International Conference on Intelligent Robots & Systems [C]. IEEE，2004.

[12] Hwangbo J，Lee J，Dosovitskiy A，et al. Learning agile and dynamic motor skills for legged robots [J]. Science Robotics，4.

[13] Yu W，Tan J，Liu C K，et al. Preparing for the Unknown：Learning a Universal Policy with Online System Identification [Z]. 2017.

[14] Endo G，Morimoto J，Matsubara T，et al. Learning CPG Sen-

sory Feedback with Policy Gradient for Biped Locomotion for a Full－Body Humanoid [A]. Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference [C]. July 9 － 13, 2005, Pittsburgh, Pennsylvania, USA. AAAI Press, 2005.

[15] Wampler K, Popovi Z, Popovi J. Generalizing locomotion style to new animals with inverse optimal regression [J]. ACM Transactions on Graphics, 2014, 33 (4): 1－11.

[16] Mordatch I, Lowrey K, Todorov E. Ensemble－CIO: Full－body dynamic motion planning that transfers to physical humanoids [A]. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) [C]. Hamburg, Germany, 2015: 5307－5314.

[17] Bansal T, Pachocki J, Sidor S, et al. Emergent Complexity via Multi－Agent Competition [Z]. 2017.

[18] Kohl N, Stone P. Machine Learning for Fast Quadrupedal Locomotion [A]. National Conference on Artifical Intelligence [C]. AAAI Press, 2004.

[19] KIMURA H. Reinforcement Learning of Walking Behavior for a Four－Legged Robot [C]. 40th IEEE Conference on Decision and Control. IEEE, 2001.

[20] 刘 帅，邬树楠，刘宇飞，等 空间机器人抓捕非合作目标的自主强化学习控制 [J]. 中国科学：物理学、力学、天文学，2019 (2)：109－118.

[21] 多南讯，吕 强，林辉灿，等. 迈进高维连续空间：深度强化学习在机器人领域中的应用 [J]. 机器人，2019 (2)：276－288.

[22] 赵玉婷，韩宝玲，罗庆生. 基于 deep Q－network 双足机器人非平整地面行走稳定性控制方法 [J]. 计算机应用，2018，38 (9)：2459－2463.

(上接第 94 页)

3.2 软件功能实现

地面服务器行车日志下载与数据分析软件为数据分析提供 AP 场强分布电子地图、轮径值、雷达因子、应答器电子地图和软件运行参数等信息，同时可以对分析策略进行实时编辑，以进行不同模式的数据分析。同时软件还可以实时查看列车状态信息。软件界面如图 7 所示，软件分析结果如图 8 所示。

图 7 地面服务器数据分析软件界面

单端日志分析结果汇总

日志起始时间	日志结束时间	日志行数
2017.12.28 2:15:10	2017.12.28 22:21:50	132067
报警次数汇总		故障率
SRP接收时间戳-单数据	0	0.00000000%
SRP接收时间戳-双数据	0	0.00000000%
SRP反馈时间戳-单数据	0	0.00000000%
SRP反馈时间戳-双数据	0	0.00000000%
SRP错误代码-单数据	0	0.00000000%
SRP错误代码-双数据	30	0.02271574%
激活的SRP	0	0.00000000%
激活的BTM	0	0.00000000%
ATO_A端当本地状态	0	0.00000000%
ATO_A错误计数	9	0.00681472%
ATO-A发送的时间戳	0	0.00000000%
单数BTM位置的合法性	0	0.00000000%
双数BTM位置的合法性	5	0.00378596%
钳停流	0	0.00000000%
单端-雷达导致SRP切换	0	0.00000000%
双端-雷达导致SRP切换	0	0.00000000%
单端-通信类故障	0	0.00000000%
双端-通信类故障	30	0.02271574%
单端-OPG-A-CH1故障	0	0.00000000%
单端-OPG-A-CH2故障	0	0.00000000%
单端-OPG-B-CH1故障	0	0.00000000%
单端-OPG-B-CH2故障	0	0.00000000%
双端-OPG-A-CH1故障	0	0.00000000%
双端-OPG-A-CH2故障	0	0.00000000%
双端-OPG-B-CH1故障	0	0.00000000%
双端-OPG-B-CH2故障	0	0.00000000%
单端-雷达-A故障	0	0.00000000%
单端-雷达-B故障	0	0.00000000%
双端-雷达-A故障	0	0.00000000%
双端-雷达-B故障	0	0.00000000%
单端-OPG速度突变	0	0.00000000%
双端-OPG速度突变	0	0.00000000%
单端-雷达速度突变	0	0.00000000%
双端-雷达速度突变	0	0.00000000%

双端日志分析结果汇总

日志起始时间	日志结束时间	日志行数
2017.12.28 1:08:50	2017.12.28 22:30:31	109046
报警次数汇总		故障率
SRP接收时间戳-单数据	0	0.00000000%
SRP接收时间戳-双数据	0	0.00000000%
SRP反馈时间戳-单数据	0	0.00000000%
SRP反馈时间戳-双数据	0	0.00000000%
SRP错误代码-单数据	21	0.01925793%
SRP错误代码-双数据	1	0.00091704%
激活的SRP	0	0.00000000%
激活的BTM	0	0.00000000%
ATO_A端当本地状态	0	0.00000000%
ATO_A错误计数	0	0.00000000%
ATO-A发送的时间戳	0	0.00000000%
单数BTM位置的合法性	4	0.00366818%
双数BTM位置的合法性	2	0.00183409%
钳停流	1	0.00091704%
单端-雷达导致SRP切换	0	0.00000000%
双端-雷达导致SRP切换	0	0.00000000%
单端-通信类故障	21	0.01925793%
双端-通信类故障	1	0.00091704%
单端-OPG-A-CH1故障	0	0.00000000%
单端-OPG-A-CH2故障	0	0.00000000%
单端-OPG-B-CH1故障	0	0.00000000%
单端-OPG-B-CH2故障	0	0.00000000%
双端-OPG-A-CH1故障	0	0.00000000%
双端-OPG-A-CH2故障	0	0.00000000%
双端-OPG-B-CH1故障	0	0.00000000%
双端-OPG-B-CH2故障	0	0.00000000%
单端-雷达-A故障	0	0.00000000%
单端-雷达-B故障	0	0.00000000%
双端-雷达-A故障	0	0.00000000%
双端-雷达-B故障	0	0.00000000%
单端-OPG速度突变	0	0.00000000%
双端-OPG速度突变	0	0.00000000%
单端-雷达速度突变	0	0.00000000%
双端-雷达速度突变	0	0.00000000%

图 8 地面服务器分析结果

4 结束语

本文提出了一种非侵入式车载信号设备在线检测运维系统方案，方案利用非接触式的磁通门传感器对车载 ATC 设备的实际输入输出和传感器电气工作状态等信息进行监测，在不影响系统可靠性的前提下，测量精度满足要求；利用既有无线通信网络实现车地行车数据实时传输，为轨旁分析提供数据，也为列车行车日志下载提供了经济便捷的解决方案；行车日志下载与数据分析软件对整合监测信息的行车日志进行分析，从中筛选异常信息进而监测各子系统服役状态，并在设备性能下降时进行预警。本文所提出的在线监测办法、车地通信实现办法及轨旁软件分析决策办法都能对提升既有的信号系统运维效能提供参考和借鉴。

参考文献：

[1] 张金玉，张 伟. 装备智能故障诊断与预测 [M]. 北京：国防工业出版社，2013.

[2] 雷 东. 车载信号系统与车辆接口故障优化处理方法 [J]. 城市轨道交通研究，2015，(z2)：53－56.

[3] 陈 磊，郭秀清，霍 勇. 轨道交通车载 ATC 数据监控系统的设计与实现 [J]. 机电一体化，2011，(11)：89－93.

[4] 徐玉宝. 地铁车载 ATC 系统的论述及分析 [J]. 数字化用户，2016，(49)：64.

[5] 石杰豪. 地铁车载信号与车辆接口功能及电路分析 [J]. 技术与市场，2017，(8)：63－64.

[6] 程昌焰，曾 成. 广州地铁三号线 B1 型列车冲标问题分析及改善对策 [J]. 价值工程，2017，(5)：89－90.

[7] 和劭延，吴春会，田建君. 电流传感器技术综述 [J]. 电气传动，2018，(1)：65－75.

[8] Cao Yi, Daping Cao. Theory of fluxgate sensor: Stability condition and critical resistance [J]. Sensors&Actuators: A. Physical, 2015, 233.

[9] 刘腾飞. 环型磁通门传感器的研究与设计 [D]. 武汉：华中科技大学，2010.

[10] 张京晶. 城市轨道交通车地无线通信频率规划的探讨 [A]. 中国科学技术协会 [C]. 广东省人民政府，2015：1－6.

[11] 丛亚闻. 基于移动闭塞的移动授权生成机制研究 [D]. 成都：西南交通大学，2011.