DOI: 10.11992/tis.201907052

仿生机器人运动步态控制:强化学习方法综述

郭宪,方勇纯
(南开大学 人工智能学院,天津 300350)

摘 要: 仿生机器人是一类典型的多关节非线性欠驱动系统,其步态控制是一个非常具有挑战性的问题。对于该问题,传统的控制和规划方法需要针对具体的运动任务进行专门设计,需要耗费大量时间和精力,而且所设计出来的控制器往往没有通用性。基于数据驱动的强化学习方法能对不同的任务进行自主学习,且对不同的机器人和运动任务具有良好的通用性。因此,近年来这种基于强化学习的方法在仿生机器人运动步态控制方面获得了不少应用。针对这方面的研究,本文从问题形式化、策略表示方法和策略学习方法 3 个方面对现有的研究情况进行了分析和总结,总结了强化学习应用于仿生机器人步态控制中尚待解决的问题,并指出了后续的发展方向。

关键词: 仿生机器人;运动步态;控制方法;强化学习;数据驱动;多关节;非线性;欠驱动
中图分类号: TP18 **文献标志码:** A **文章编号:** 1673−4785(2020)01−0152−08

中文引用格式: 郭宪,方勇纯. 仿生机器人运动步态控制:强化学习方法综述 [J]. 智能系统学报, 2020, 15(1): 152–159.
英文引用格式: GUO Xian, FANG Yongchun. Locomotion gait control for bionic robots: a review of reinforcement learning methods[J]. CAAI transactions on intelligent systems, 2020, 15(1): 152–159.

Locomotion gait control for bionic robots: a review of reinforcement learning methods

GUO Xian, FANG Yongchun
(College of Artificial Intelligence, Nankai University, Tianjin 300350, China)

Abstract: The bionic robot is a typical multi-joint, nonlinear, underactuated system, for which locomotion gait control is of much challenge. For this problem, traditional control and planning methods need to be carefully designed for specific locomotion tasks, which takes a lot of time and efforts, yet lacks generality. On the contrary, data-driven reinforcement learning method can autonomously learn the controller for different locomotion tasks, and it presents the advantage of good generality for different bionic robots and locomotions. Therefore, in recent years, this reinforcement learning-based method has been widely used in the field of bionic robots to construct various locomotion gait controllers. In this paper, the current research status of reinforcement learning-based methods for the locomotion control of bionic robots is comprehensively analyzed, respectively from the following three aspects: formulation of the problem, policy representation, and policy learning. Finally, the problems to be solved in the field are and summarized, and the possible future research directions are provided.

Keywords: bionic robot; locomotion gait; control method; reinforcement learning; data-driven; multi-joint; nonlinear; underactuated

自然界中的动物种类繁多、遍布于世界的各个角落,数亿年的进化使得它们的形态完全适应于所处的环境。为了更好地服务于人类的生产和生活,学者们模仿动物的形态发明了多类仿生机器人,如仿生足类机器人、仿生蛇形机器人等。与轮式机器人相比,这些仿生机器人在各种各样的复杂环境下,如山地、沟壑、海洋、丛林、沼泽

收稿日期: 2019−07−29.
基金项目: 国家自然科学基金项目(61603200);天津市自然科学基金青年项目 (19JCQNJC03200).
通信作者: 方勇纯. E-mail:fangyc@nankai.edu.cn.

等, 展示出更好的运动性能, 因此仿生机器人在民用、军事、星球探测等领域具有广泛的应用空间。与自然界中的动物类似, 仿生机器人通过周期性地改变身体的构型, 如足类机器人改变腿部构型, 蛇形机器人改变身体关节等, 并与环境相互作用从而实现各种各样的运动, 我们将这种运动方式称为运动步态。

由于仿生机器人是多刚体非线性欠驱动系统, 因此其运动步态的控制是一项非常富有挑战性的工作。经过学者们多年不断的努力, 目前最高效的运动步态控制方法是通过将运动任务分解为不同的子模块, 并对子模块分别进行控制。例如, 足类机器人为了进行运动步态控制常常需要进行状态估计、接触点选择、轨迹优化、足端点规划、模型预测控制和操作空间控制等工作[1−2]。这种控制方法能使得目前的仿生机器人获得高机动的运动能力, 如波士顿动力公司的四足机器人可以在野外[3] 和室内[4] 做高机动抗扰动和自平衡运动、MIT 的猎豹可以实现高达 5 m/s 的速度[5], 日本东京工业大学的系列蛇形机器人可以在野外和水下等环境下自由运动[6]、卡内基梅隆大学的模块化蛇形机器人可以爬树、爬管道等[7]。日本通信大学设计的模块化蛇形机器人可以在废墟环境以履带步态方式进行运动[8]。但是, 这些步态控制方法需要大量的专业知识, 且通用性不强, 即使对于同款机器人, 针对不同的运动任务仍需大量工作调整设计控制方法, 而不同款的机器人则需要利用不同的专业知识来重新设计控制器。此外, 利用这种步态控制方法所得到的运动一般并不是最优的, 抗干扰能力较差。

自然界中动物娴熟的运动技能是在出生后不断与环境交互通过试错学习而获得的。近年来广泛应用的强化学习便是这样一种试错的学习方法, 该方法并不直接考虑机器人的运动学和动力学, 而是一种基于数据驱动的控制器设计方法。研究表明: 通过强化学习方法来设计仿生机器人的运动步态控制器可以克服上述分别进行子模块设计所带来的局限。然而, 仿生机器人运动步态的控制涉及连续的高维观测空间和动作空间, 计算量非常大, 因此由于实时性等方面的原因, 传统的强化学习算法往往难以直接应用。随着深度学习技术的出现, 将强化学习与深度学习相结合而形成的深度强化学习技术得到快速发展, 并在视频游戏[9]、围棋[10] 等领域取得突破性进展。近年来, 深度强化学习技术也被广泛应用到仿生机器人的运动步态控制器的设计中。Levine 等[11] 结合轨迹最优和监督学习在仿真环境中训练二维人形机器人实现行走运动步态控制。Schulman 等[12−13] 提出 TRPO 和 PPO 算法并利用 actor-critic 框架实现稳定的训练算法。Peng 等[14] 利用分层深度强化学习算法在仿真环境中训练 3 维仿生机器人的运动。最近, 更高效稳定的深度强化学习算法如 MPO 算法[15]、SAC 算法[16]、TD3 算法[17] 被提出来。除了在仿真环境中进行仿生机器人运动步态的训练, Hwangbo 等[18] 将在仿真环境中的训练结果直接应用到 ANYmal 四足机器人, 实现了稳定高效的四足运动步态控制; Haarnoja 等[19] 利用改进的 SAC 算法直接在 Minitaur 四足机器人上进行训练, 实现了多种运动步态控制。

针对强化学习算法在仿生机器人运动步态控制上的研究, 本文从以下几个方面对当前的研究进行综述: 首先将仿生机器人的运动步态控制问题进行数学形式化, 将该问题纳入到马尔可夫决策过程的理论框架中; 然后对现有的策略表示方法进行调研; 之后介绍不同的策略学习方法, 最后给出总结和展望。

1　问题形式化

强化学习通过试错的方法实现折扣累积回报期望的最大化。智能体与环境交互的过程可以用马尔可夫决策过程来描述, 而马尔可夫决策过程可以利用一个五元组 $(\mathcal{S},\mathcal{A},r,P,\gamma)$ 来描述, 其中 $\mathcal{S}$ 为状态空间, $\mathcal{A}$ 为动作空间, $r(s_t,a_t)$ 为与状态和动作相关的立即回报, $P:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbf{R}$ 为状态转移概率, 即给定状态 s_t 和动作 a_t 转移到新的状态 s_{t+1} 的概率,即 $p(s_{t+1}|s_t,a_t)$, γ 为计算折扣累积回报的折扣因子。令 π 表示一个随机策略, 即 $\pi:\mathcal{S}\times\mathcal{A}\to[0,1]$, 则智能体与环境交互的过程可以表述为智能体在状态 s_t 处, 采取策略 $\pi(s_t)$ 得到动作 a_t, 并与境进行交互, 根据状态转移概率 $p(s_{t+1}|s_t,a_t)$ 得到下一时刻的状态同时获得立即回报 $r(s_t,a_t)$ 的过程。智能体不断地与环境进行交互便产生了一条状态和动作的轨迹, 用 τ 来表示, 即 $\tau=[s_0,a_0,s_1,a_1,\cdots,s_n,a_n]$, 我们用 $\eta(\pi)$ 来表示折扣累积回报的期望, 即

$$\eta(\pi)=\mathrm{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$$

经典强化学习算法在数学上可以进行如下形式化描述[20]:

$$\max_\pi \mathrm{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \tag{1}$$

经典强化学习算法包括基于值函数的方法、基于直接策略搜索的方法以及基于 actor-critic 的方法。仿生机器人运动步态的状态空间和动作空

间都是连续的，因此最常用的为基于直接策略搜索的方法和基于 actor-critic 的方法。其中，前者最常用的为策略梯度的方法及其变种，如 TRPO 的方法和 PPO 的方法，后者最常用的为 DDPG[21] 的方法。

经典强化学习进行数学形式化时并未显式地考虑策略的探索性，因此该类算法容易陷入到局部最优。为了将探索策略考虑在内，学者们提出在优化折扣累积回报的期望的同时优化策略的熵，因此提出了最大熵强化学习的问题形式化[19]：

$$\max_{\pi} \mathrm{E}_{\tau}\left[\sum_{t=0}^{\infty}\gamma^{t}(r(s_t,a_t)-\alpha_t\log\pi_t(a_t|s_t))\right] \tag{2}$$

由式 (2) 描述的最优策略对外界的扰动具有更强的鲁棒性，同时在训练过程中充分考虑了策略的探索性，从而加快了学习速度，实验证明该形式化在仿生机器人运动步态的学习中取得最好的表现。

仿生机器人是典型的机电一体化系统，其本身的机械强度和驱动功率都有限制条件，如最大加速度、最大力矩等。然而，典型的最大期望强化学习或者最大熵强化学习的最优解往往都是高频的 bang-bang 控制信号，实际的仿生机器人在执行这些控制信号时要么非常容易损坏，要么根本达不到要求[22]。因此，对于仿生机器人而言，确保加速度约束、速度约束、力矩约束都是关键的。为了将这些约束考虑进去，仿生机器人的运动步态控制问题可以纳入到约束马尔可夫决策过程的框架中[23]。约束马尔可夫决策过程可以用六元组 $(\mathcal{S},\mathcal{A},P,\gamma,r,c)$ 来表示。其中，前 5 个元素表示了马尔可夫决策过程，第 6 个元素 c 表示约束。约束马尔可夫决策过程形式化为带约束的优化问题，即

$$\max_{\pi} \mathrm{E}_{\tau}\left[\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)\right]$$
$$\text{s.t.}\mathrm{E}_{\tau}\left[\sum_{t=0}^{\infty}\gamma^{t}c(s_t,a_t)\right]\leqslant\bar{C}$$

2　策略表示方法研究

强化学习最终目标是得到最优的控制策略，如前所示用策略 $\pi:\mathcal{S}\times\mathcal{A}\to[0,1]$ 来表示。对于仿生机器人步态控制来说，状态空间包括机器人的每个关节角度 q、相应的角速度 $\dot{q}$、整个身体质心的位姿 p、相应的速度 $\dot{p}$、足端或身体与外部的接触力 F_{ext}，动作空间为每个关节处的驱动力矩 τ。然而，仿生机器人都是多关节系统，少则十几个关节，多则几十个关节，因此仿生机器人的状态空间和动作空间都是高维的连续空间，这给强化学习带来了非常多的挑战。为了对仿生机器人运动步态控制进行有效的学习，学者们提出了很多更有效的策略表示方法。

2.1　基于领域知识的策略参数化方法

2.1.1　基于 CPG 的策略表示方法

CPG 即中枢模式发生器，研究发现 CPG 广泛存在于脊椎动物和非脊椎动物的体内，用于控制动物有节律的呼吸、心跳、肠胃蠕动、行走、奔跑、游动等[24]。仿生机器人学家利用 CPG 模型产生周期信号控制机器人的运动步态，可以大大简化动作空间，因此 CPG 在人形机器人、四足机器人、蛇形机器人等仿生机器人的运动步态控制中得到广泛应用。现有的 CPG 模型包括 Matsuoka 模型[25]、锁相性模型[26] 和范德波尔模型[27]。其中在仿生机器人中被广泛应用的为 Matsuoka 模型。

第 j 个神经元震荡器的动力学模型为[28]

$$\tau\dot{z}_j=-z_j-\sum_{k=1}^{n}w_{jk}q_k-\gamma z'_j+c+a_j$$
$$\tau'\dot{z}'_j=-z_j+q_j$$
$$q_j=\max(0,z_j)$$

式中：τ、τ'、c 为调节常数；z_j 为平均膜势能；z'_j 表示衰减项；γ 为衰减系数；q_j 为神经元的输出；w_{jk} 表示第 k 个神经元对第 j 个神经元的抑制作用；a_j 为反馈信号。当采用 CPG 模型作为运动步态控制器时，常常将 CPG 模型中的反馈信号 a_j 参数化，作为优化控制项。

对于简单的二维人形机器人，Matsubara 将 CPG 模型分别放在两个臀关节和一个状态机控制器来控制膝关节[29]。而对于三维人形机器人，由于关节数目多达 21 个，采用 CPG 模型描述关节角非常复杂。为了对多关节人形机器人进行控制，Endo 等将 CPG 模型用到任务空间，而状态表示也只考虑了表示姿态的角速度，其具体的策略表示方法为

$$a_j(t)=a_j^{\max}\frac{2}{\pi}\arctan\left(\frac{\pi}{2}\nu_j(t)\right)$$

式中：$a_j^{\max}$ 为振幅，$\nu_j\sim P(\nu_j|x;w^{\mu})$，即 ν_j 从随机策略 $P(\nu_j|x;w^{\mu})$ 中采样得到。其中随机策略为高斯策略：

$$p(\nu_j|\boldsymbol{x};\boldsymbol{w}^{\mu})=\frac{1}{\sqrt{2\pi}\sigma_j(w^{\sigma})}\exp\left(\frac{-(\nu_j-\mu_j(\boldsymbol{x};\boldsymbol{w}^{\mu}))^2}{2\sigma_j^2(\boldsymbol{w}^{\sigma})}\right)$$

其中要优化的参数为 $(\boldsymbol{w}^{\mu},\boldsymbol{w}^{\sigma})$，$\mu_j(\boldsymbol{x};\boldsymbol{w}^{\mu})$ 采用的是归一化高斯网络[30]，$\boldsymbol{x}=(\dot{\theta}_{\text{roll}},\ \dot{\theta}_{\text{pitch}})^{\mathrm{T}}$ 为输入观测。

2.1.2　基于仿生运动曲线的策略表示方法

仿生蛇形机器人的运动步态控制常用仿生曲线来表示。最常用的仿生曲线为日本广濑教授提出的 serpenoid 曲线，该仿生曲线可表示为[31]

$$\theta_i=\theta_0+A\sin(w_s s_i-w_T t) \tag{3}$$

式中: θ_0 为角度偏置项; A 为振幅; w_s 为空间频率, 表示波动沿着身体传播的速度; w_T 为时间角频率, 表示单个关节的震荡频率。蛇形曲线来表示控制策略大大简化了控制步态的参数个数。在蛇形曲线式 (3) 中需要控制的参数为 (θ_0, A, w_T)。Fang 等[32] 利用强化学习方法直接对这 3 个参数进行优化, Sartoretti 等[31] 将当前的形状参数作为状态输入并将参数 $\{A, w\}$ 利用神经网络表示成为形状参数的非线性函数。

2.1.3 基于 DMP 的策略表示方法

动态运动基元 (DMP) 方法用一组微分方程表示光滑的运动策略, 通过调整耦合项可以灵活地对运动策略进行调制, 在仿生机器人尤其是足类机器人中得到广泛应用。一组节律运动基元可表示为[33]

$$\tau\dot{y} = z + \frac{\sum_{i=1}^{N} \Psi_i \boldsymbol{w}_i^{\mathrm{T}} \tilde{\boldsymbol{\upsilon}}}{\sum_{i=1}^{N} \Psi_i}$$

$$\tau\dot{z} = \alpha_z(\beta_z(y_m - y) - z)$$

式中: (y, z) 为控制信号; $\Psi_i = \exp(-h_i(\bmod(\phi, 2\pi) - c_i)^2)$; $c_i \in [0, 2\pi]$, $\tilde{\boldsymbol{\upsilon}} = [r\cos\phi \;\; r\sin\phi]^{\mathrm{T}}$, $\tau\dot{\phi}=1$ $\tau\dot{r} = -\mu(r - r_0)$。

强化学习需要优化的参数为 $\boldsymbol{w}_i^{\mathrm{T}}$, 常用的学习方法为策略梯度法[34]。

2.2 基于深度神经网络的策略参数化方法

前面基于领域知识的策略表示方法耦合了大量先验知识, 如步态的周期性、对称性等, 这种表示方法使得强化学习所需要优化的参数量大幅度减少, 因此只需要几百次或上千次的训练就能得到最优解, 而且所得到的解能应用到实际的仿生机器人中。然而, 基于领域知识的策略表示方法表示能力非常有限, 无法表示一般的运动步态。因此, 这些基于领域知识的策略表示方法对不同的运动任务没有通用性, 只能使得仿生机器人实现常见的行走。随着深度学习技术的进步, 深度强化学习在仿生机器人运动步态控制领域得到更深入的研究。通用的深度神经网络用来表示运动步态控制率。与基于领域知识的控制策略不同, 基于深度神经网络的控制策略没有考虑先验知识, 因此更具有通用性, 并且随着网络参数的增加, 其表示能力增强, 而且可以根据不同的运动任务学习不同的运动控制策略。

对于状态的表示, 用 ϕ_i 和 $\dot{\phi}_i$ 来表示仿生机器人的关节角和关节角速度, 用 $\theta_{\mathrm{roll}}, \dot{\theta}_{\mathrm{roll}}, \theta_{\mathrm{pitch}}, \dot{\theta}_{\mathrm{pitch}}$ 来表示仿生机器人整体的滚转角度、角速度, 俯仰角度和角速度。足类仿生机器人与仿生蛇形机器人不同, 足类机器人的状态空间还应该包括足与地面的碰撞作用, 为此在状态空间的表示上应该反映出碰撞作用。Yu 等[35] 利用一个特殊的二值向量 $\boldsymbol{c}$ 表示每条腿是否与地面接触, 状态空间为 $s = [\phi, \dot{\phi}, \boldsymbol{c}, \tilde{\boldsymbol{\upsilon}}]$。Haarnoja 等[19] 利用连续的 5 帧观测总共 112 维的向量作为状态向量来建模实际仿生四足机器人由于信号延迟和地面碰撞作用产生非马尔可夫性。Hwangbo 等[18] 利用关节历史信息建模腿与地面的碰撞作用, 其建立的状态空间为 $s_k = [\phi^g, r_z, v, w, \phi_i, \dot{\phi}_i, \Theta, a_{k-1}, C]$。对于复杂的运动任务[36], 如不同地形下的运动步态控制, 状态往往包括两部分, 即 $s = [s_c, s_T]$, 一部分为仿生机器人本身的状态信息 s_c, 另一部分为面临的地形特征描述 s_T。对于仿生蛇形机器人, 为了得到能量利用率最高的步态, 状态除了自身的特征外, 还需要知道关节力[37], 即 $s_t = [\phi_i, \dot{\phi}_i, \upsilon_1, \tau_i, \upsilon_t]$。

对于动作的表示, 可以直接利用关节的力矩 τ 来表示[11], 也可以利用每个关节的期望关节角度 ϕ_i^d 来表示, 然后利用 PD 控制或者阻抗控制器得到每个关节的关节力矩。Peng 等[38] 研究发现直接优化期望关节角度比直接优化力矩更稳定。

有了状态的表示和动作的表示之后, 策略常常可以利用带有两个隐含层的神经网络来表示从状态到动作的映射。对于带有地形适应的步态控制, 则地形部分的描述往往需要经过若干卷积层后, 再与表示机器人自身特征的状态串接在一起, 输入到一个前向神经网络中。

3 策略学习方法

强化学习方法可以分为基于值函数的方法、基于直接策略搜索的方法和基于 Actor-Critic 框架的方法。基于值函数的方法常用于离散动作空间。仿生机器人的动作空间为高维的连续空间, 因此常用的学习方法为后两种。现有的仿生机器人策略学习方法可以分为两大类, 第 1 类是将仿生机器人视为单智能体, 所有的驱动关节空间为动作空间, 利用单智能体强化学习的方法进行学习; 第 2 类则是将仿生机器人按照不同的足或者身体部位视为不同的智能体, 整个仿生机器人视为多个智能体系统, 其步态运动控制视为多个智能体的协同运动, 利用多智能强化学习的方法学习策略。

3.1 单智能体强化学习方法

3.1.1 基于轨迹最优的方法

基于轨迹最优的 GPS 方法由 Levine 等[11] 提出, 该方法通过交叉优化最优控制策略 $p_i(u|x)$ 和神经网络策略 $\pi_\theta(x)$ 学习得到最优的神经网络策略。具体学习过程如下。

1) 利用轨迹最优算法, 如微分动态规划等来

优化得到最优的策略，其优化目标为

$$L(q) = D_{KL}(q(\tau)\|\rho(\tau)) + \sum_{t=1}^{T}\lambda_t D_{KL}(q(x_t)\pi_\theta(u_t|x_t)\|q(x_t,u_t))$$

2) 得到最优控制策略后，利用该最优策略在数据点 τ 处，利用监督学习的方法对神经网络策略进行训练，其优化目标为

$$L(\theta) = \sum_{t=1}^{T}\lambda_t \sum_{i=1}^{N} D_{KL}(\pi_\theta(u_t|x_{ti})\|q(u_t|x_{ti}))$$

3）更新对偶变量 λ_t 更新规则为

$$\lambda_t \leftarrow \lambda_t + \eta D_{KL}(q(x_t)\pi_\theta(u_t|x_t)\|q(x_t,u_t))$$

与传统的基于随机策略搜索的方法相比，GPS 的方法利用轨迹最优方法得到局部最优解，并利用最优解进行监督训练，避免了随机探索。经过训练得到的通用神经网络策略能泛化到其他情况下。实验证明，该方法能高效地学习游动、跳跃、行走、奔跑等运动。

3.1.2　基于确定性策略梯度的方法

仿生机器人一般拥有十几个甚至是几十个控制输入，使得其动作空间为十几维甚至是几十维的连续空间。基于随机策略梯度的方法在进行梯度估计的时候需要在如此高维的空间进行大量采样，使得学习效率很低。基于确定性策略梯度的方法不需要在动作空间进行采样，只需在状态空间进行采样。尤其是深度确定性策略梯度算法即 DDPG 算法[21] 利用深度神经网络逼近行为值函数，并利用 off-policy 的方法估计行为值函数，大大提升了数据效率，在连续运动控制中得到广泛应用。DDPG 的方法为 Actor-Critic 框架的强化学习算法，在该算法中行为值函数和策略函数都由神经网络来逼近，其更新规则为

$$\delta_t = r_t + \gamma Q^{w^-}(s_{t+1},\mu_{\theta^-}(s_{t+1})) - Q^w(s_t,a_t)$$
$$w_{t+1} = w_t + \alpha_m \delta_t \nabla_w Q^w(s_t,a_t)$$
$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s_t)\nabla_a Q^w(s_t,a_t)|_{a=\mu_\theta(s)}$$

式中：w、w^-、θ、θ^- 分别为行为值函数网络的估计参数、目标网络参数、策略网络的估计参数、目标策略网参数。

DDPG 方法在连续控制问题取得普遍较好的结果。然而，由于 off-policy 的存在，行为值函数的估计普遍存在过优估计的问题。同时，由于行为值函数参数的更新与策略网络参数的更新同时交叉更新，这使得学习过程非常不稳定。为了解决这些问题，各种各样的改进的 DDPG 算法被提出来，其中 Fujimoto 等提出 TD3 的方法[17]，利用 Double Q-learning[39] 的方法来解决过优估计问题，利用策略网络延迟更新的方法解决学习不稳定，在连续控制问题上取得当前最好的结果。

3.1.3　基于最大熵的方法

当仿生机器人的自由度很高时，GPS 的方法难以拟合局部动力学，DDPG 的方法则需要额外的探索策略，学习效率低。基于最大熵的强化学习算法将探索策略耦合到优化之中，因此增强了学习效率。根据策略迭代方法提出来的 soft-actor-critic 方法[16] 充分利用了最大熵原理，在保证收敛性的同时，能快速地收敛到最优解。最大熵强化学习的问题形式化为式 (2) 所示。为了求解式 (1)，需要依次优化如下 3 个损失函数：

$$J_Q(\theta) = \mathrm{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}[(Q_\theta(s_t,a_t) - (r(s_t,a_t) + \gamma V_{\theta_1,\theta_2}(s_{t+1})))^2] \tag{4}$$

$$J_\pi(\phi) = \mathrm{E}_{s_t\sim\mathcal{D},a_t\sim\pi_\phi}\left[\alpha\log\pi_\phi(a_t|s_t) - \min_{i\in\{1,2\}} Q_{\theta i}(s_t,a_t)\right] \tag{5}$$

$$J(\alpha) = \mathrm{E}_{s_t\sim\mathcal{D},a_t\sim\pi_\phi}[-\alpha\log\pi_\phi(a_t|s_t) - \alpha\mathcal{H}] \tag{6}$$

其中式 (4) 为行为值函数的损失函数，由贝尔曼参数的均方和来给出。为了移除过优估计，使用 2 套独立的参数 θ_1 和 θ_2。式 (5) 为最大熵策略的损失函数，式 (6) 为自动调整因此 α 的损失函数，其中 α 为式 (3) 中熵所占的比重，该参数在优化的过程中不断发生变化。SAC 算法被成功应用到四足机器人的运动步态控制中。值得一提的是，该算法可以在真实的四足机器人上直接进行训练，经过 2 h 的训练，四足机器人就能学会行走。

3.1.4　基于最大后验策略优化的方法

最大后验策略优化的方法从概率的角度出发，将强化学习问题建模为推理问题。假设 $p_\pi(O=1)$ 为完成任务的概率，则根据推理问题，该概率为[15]

$$\log p_\pi(O=1) = \log\int p_\pi(\tau)p(O=1|\tau)\mathrm{d}\tau \geqslant \int q(\tau)\left[\log p(O=1|\tau) + \log\frac{p_\pi(\tau)}{q(\tau)}\right]\mathrm{d}\tau$$

令损失函数：

$$J(q,\pi) = \mathrm{E}_q\left[\sum_t r_t/\alpha\right] - KL(q(\tau)\|p_\pi(\tau))$$

式中：$q(\tau)$ 为提议分布。该优化问题可通过 EM 方法进行求解，在 E 步优化得到最优的提议分布 $q(\tau)$，在该步中非参数优化解为

$$q_i(a|s) \propto \pi(a|s,\theta_i)\exp\left(\frac{Q_{\theta_i}(s,a)}{\eta^*}\right)$$

其中最优温度项 η^* 根据式（7）优化得到：

$$g(\eta) = \eta\varepsilon + \eta\int\mu(s)\log\int\pi(a|s,\theta_i)\exp\left(\frac{Q_{\theta_i}(s,a)}{\eta}\right)\mathrm{d}a\mathrm{d}s \tag{7}$$

在 M 步中，利用最优的提议分布更新神经网络策略：

$$\max_\theta J(q_i,\theta) = \max_\theta \mathrm{E}_{\mu_q(s)}[\mathrm{E}_{q(a|s)}[\log\pi(a|s,\theta)]] + \log p(\theta)$$

基于最大后验策略优化的方法与基于轨迹最优的方法类似, 都是先优化得到一个局部最优策略, 然后以该策略为目标进行监督学习。不同的是, 基于轨迹最优的方法需要先拟合一个动力学模型, 然后根据轨迹最优方法得到局部最优策略, 是基于模型的方法; 而基于最大后验策略优化则是完全根据数据进行的无模型优化方法。因此基于最大后验策略优化的方法具有更大的应用范围和通用性。

3.2　多智能体强化学习方法

将仿生机器人的驱动关节分成若干个独立的智能体, 每个智能体共享一套同样的控制策略, 利用异步的分布式方法对多智能体系统进行训练可以加速学习的过程。Sartoretti 等利用 A3C 的方法利用分布式强化学习对仿生蛇形机器人和仿生六足机器人进行训练[31]。更具体来说, 对于仿生蛇形机器人, 整个身体关节可以看成 6 个智能体, 每个智能体的策略利用仿生曲线式 (9) 进行参数化, 并利用共享回报和 A3C 的方法对共享策略进行训练。对于六足机器人, 每条足视为一个智能体, 利用基于 CPG 的方法参数化共享策略。

4　存在的问题及发展趋势

本文对强化学习算法在仿生机器人的步态控制领域的研究和发展进行了综述, 具体包括仿生机器人运动步态控制的问题形式化、现有的策略表示方法研究、现有的策略研究方法研究。总体来说, 目前强化学习算法在仿生机器人领域得到快速发展, 不过目前普遍存在很多问题。

4.1　存在的问题

1) 样本效率低

人类能快速地学会走路、奔跑、跳跃等运动步态, 然而现有的强化学习方法则需要几十万甚至上百万次的尝试。这不仅需要耗费大量的时间和能量, 还会导致仿生机器人严重磨损甚至坏掉。如何提升样本效率, 这是强化学习应用于实际的仿生机器人步态学习中急需解决的重要问题。

2) 无法有效地进行多任务学习

现有的强化学习算法大都只能学习单一的运动步态, 当学习其他类型的运动步态或任务时, 需要重新训练; 如何通过一次训练便可以学会多个运动步态或完成多个运动任务是当前研究中存在的一个重要问题。

3) 从仿真环境到实际平台的迁移性差

现有的强化学习方法大都是先构建机器人的仿真模型, 在仿真环境中训练机器人的运动步态, 然而实际模型与仿真模型往往存在较大差异, 这就导致在仿真环境中训练的策略直接迁移到实际机器人上会产生很大的偏差。因此, 从仿真到实际机器人平台的迁移学习是有待研究的重要问题。

4) 鲁棒性差

对于实际的机器人系统, 由于各种传感器存在着误差, 这就导致机器人实际的观测是带有噪音的, 而在无噪音条件下训练的策略往往失效, 因此如何得到鲁棒的强化学习算法是有待研究的重要问题。

针对上述存在的问题, 目前仿生机器人运动步态学习的发展趋势如下。

4.2　发展趋势

1) 基于模型的强化学习

为了提升强化学习算法的样本效率, 基于模型的强化学习算法近年来成为该领域研究的热点。Ha 等[40] 提出创建世界模型, 在进行策略学习之前先学习一个世界模型, 然后利用世界模型对下一个状态进行预测, 预测的状态作为输入的一部分耦合进策略学习中。当输入为像素时, 状态空间为高维输入, 机器人一般需要大量的交互数据进行步态的学习, Ebert 等[41] 提出创建图像预测模型, 利用该模型创建虚拟环境, 在虚拟环境中进行局部训练, 为了不断降低虚拟环境的误差, 智能体通过与真实世界的交互不断优化虚拟环境, 从而最终减少与真实世界的交互。

2) 元强化学习

为了使得机器人快速地学习多项任务, 元强化学习算法近年来成为研究热点。与经典强化学习算法不同, 元强化学习是在任务空间进行训练, 学习到任务空间的先验知识, 以便在学习新的任务时能利用以前的先验知识进行快速地学习。为了使得智能体具有连续学习的能力, Finn 等[42] 提出在线元强化学习的方法, 从而使得机器人能连续学习多个运动任务。

3) 分层强化学习

在仿生机器人学习越障、踢球等复杂的运动任务时, 任务本身具有很强的不同水平的决策的特性。简单的端到端的学习面临着学习效率低、学习效果差等问题。为此, 分层强化学习算法成为解决该问题的研究热点。Peng 等在解决仿生机器人复杂运动任务时, 将动作空间分成两层: 上层的目标位置和底层的关节动作, 通过两层决策实现学习目的。Mahjourian 等[43] 在解决乒乓球机器人打球的运动任务时采用分层的策略, 其中底层采用基于模型的控制器, 高层采用强化学习的方

法学习无模型的控制器，获得了高效的学习算法。

5　结束语

本文从问题形式化、策略表示方法和策略学习 3 个方面对当前强化学习算法应用到仿生机器人的运动步态控制任务中的研究情况进行了分析和总结，并给出了强化学习算法应用到该领域尚待解决的问题和未来的发展方向。总体而言，不同于仿真环境，仿生机器人的步态运动控制受到实际机器人系统的驱动、机构、通信等多方面的限制，使得强化学习算法在该领域中的应用表现出极大的挑战。一般而言，在形式化方面，需要利用约束马尔可夫决策过程对该问题进行建模；在策略表示方面，更倾向于领域结构化的表示方法；在策略学习方面，高效的直接策略搜索方法表现更佳。然而，目前强化学习算法用于仿生机器人运动步态学习和控制仍然面临着样本效率低、无法有效地进行多任务学习、从仿真环境到实际平台的迁移性差和学习鲁棒性差等问题。新的方法如基于模型的强化学习、元强化学习和分层强化学习等有望解决或缓解这些问题。

参考文献：

[1] GEHRING C, COROS S, HUTTER M, et al. Practice makes perfect: an optimization-based approach to controlling agile motions for a quadruped robot[J]. IEEE robotics & automation magazine, 2016, 23(1): 34–43.

[2] APGAR T, CLARY P, GREEN K, et al. Fast online trajectory optimization for the bipedal robot Cassie[C]//Proceedings of Robotics: Science and Systems 2018. Pittsburgh, USA, 2018.

[3] RAIBERT M, BLANKESPOOR K, NELSON G, et al. BigDog, the rough-terrain quadruped robot[C]//Proceedings of the 17th World Congress of the International Federation of Automatic Control. Seoul, Korea, 2008: 10822−10825.

[4] Spotmini autonomous navigation[EB/OL].[2018-08-11]. https://ucrazy.ru/video/1526182828-spotmini-autonomous-navigation.html.

[5] PARK H W, PARK S, KIM S. Variable-speed quadrupedal bounding using impulse planning: Untethered high-speed 3D running of MIT Cheetah 2[C]//Proceedings of 2015 IEEE International Conference on Robotics and Automation. Seattle, USA, 2015: 5163−5170.

[6] HIROSE S, YAMADA H. Snake-like robots: machine design of biologically inspired robots[J]. IEEE robotics and automation magazine, 2009, 16(1): 88–98.

[7] HATTON R L, CHOSET H. Generating gaits for snake robots by annealed chain fitting and keyframe wave extraction[C]//Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. St. Louis, USA, 2009: 840−845.

[8] TAKEMORI T, TANAKA M, MATSUNO F. Gait design for a snake robot by connecting curve segments and experimental demonstration[J]. IEEE transactions on robotics, 2018, 34(5): 1384–1391.

[9] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529–533.

[10] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354–359.

[11] LEVINE S, KOLTUN V. Learning complex neural network policies with trajectory optimization[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014: 829−837.

[12] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[C]//Proceedings of the 31st International Conference on Machine Learning. Lille, France, 2015: 1889−1897.

[13] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. (2017-08-28). https://arxiv.org/abs/1707.06347.

[14] PENG Xuebin, BERSETH G, YIN Kangkang, et al. DeepLoco: dynamic locomotion skills using hierarchical deep reinforcement learning[J]. ACM transactions on graphics, 2017, 36(4): 1–13.

[15] ABDOLMALEKI A, SPRINGENBERG J T, TASSA Y, et al. Maximum a posteriori policy optimisation[EB/OL]. (2018-06-14). https://arxiv.org/abs/1806.06920.

[16] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications[EB/OL]. (2019-01-29). https://arxiv.org/abs/1812.05905.

[17] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]// Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Sweden, 2018: 1587−1596.

[18] HWANGBO J, LEE J, DOSOVITSKIY A, et al. Learning agile and dynamic motor skills for legged robots[J]. Science robotics, 2019, 4(26): 5872–5880.

[19] HAARNOJA T, HA S, ZHOU A, et al. Learning to walk via deep reinforcement learning[EB/OL]. (2019-06-19). https://arxiv.org/abs/1812.11103

[20] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 1998.

[21] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer science, 2015, 8(6): A187.

[22] BOHEZ S, ABDOLMALEKI A, NEUNERT M, et al. Value constrained model-free continuous control[EB/OL]. (2019-02-12). https://arxiv.org/abs/1902.04623.

[23] ALTMAN E. Constrained Markov decision processes[M]. London: Chapman and Hall, 1999.

[24] DELCOMYN F. Neural basis of rhythmic behavior in animals[J]. Science, 1980, 210(4469): 492–498.

[25] MATSUOKA K. Sustained oscillations generated by mutually inhibiting neurons with adaptation[J]. Biological cybernetics, 1985, 52(6): 367–376.

[26] COHEN A H, HOLMES P J, RAND R H. The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: a mathematical model[J]. Journal of mathematical biology, 1982, 13(3): 345–369.

[27] BAY J S, HEMAMI H. Modeling of a neural pattern generator with coupled nonlinear oscillators[J]. IEEE transactions on biomedical engineering, 1987, BME−34(4): 297–306.

[28] ENDO G, MORIMOTO J, MATSUBARA T, et al. Learning CPG-based biped locomotion with a policy gradient method: application to a humanoid robot[J]. The international journal of robotics research, 2008, 27(2): 213–228.

[29] MATSUBARA T, MORIMOTO J, NAKANISHI J, et al. Learning CPG-based biped locomotion with a policy gradient method[C]//Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots. Tsukuba, Japan, 2005.

[30] DOYA K. Reinforcement learning in continuous time and space[J]. Neural computation, 2000, 12(1): 219–245.

[31] SARTORETTI G, PAIVINE W, SHI Yunfei, et al. Distributed learning of decentralized control policies for articulated mobile robots[J]. IEEE transactions on robotics, 2019, 35(5): 1109–1122.

[32] 方勇纯, 朱威, 郭宪. 基于路径积分强化学习方法的蛇形机器人目标导向运动 [J]. 模式识别与人工智能, 2019, 32(1): 1–9.
FANG Yongchun, ZHU Wei, GUO Xian. Target-directed locomotion of a snake-like robot based on path integral reinforcement learning[J]. Pattern recognition and artificial intelligence, 2019, 32(1): 1–9.

[33] IJSPEERT A J, SCHAAL S. Learning attractor landscapes for learning motor primitives[M]//THRUN S, SAUL L K, SCHOLKOPF B. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 1547−1554.

[34] SCHAAL S, PETERS J, NAKANISHI J, et al. Learning movement primitives[M]//DARIORAJA P, CHATILA R. Robotics Research. The Eleventh International Symposium. Berlin, Germany: Springer, 2005.

[35] YU Wenhao, TURK G, LIU C K. Learning symmetric and low-energy locomotion[J]. ACM transactions on graphics, 2018, 37(4): 144–150.

[36] PENG Xuebin, BERSETH G, VAN DE PANNE M. Terrain-adaptive locomotion skills using deep reinforcement learning[J]. ACM transactions on graphics, 2016, 35(4): 81–88.

[37] BING Zhenshan, LEMKE C, JIANG Zhuangyi, et al. Energy-efficient slithering gait exploration for a snake-like robot based on reinforcement learning[EB/OL]. (2019-04-16). https://arxiv.org/abs/1904.07788v1.

[38] PENG Xuebin, VAN DE PANNE M. Learning locomotion skills using DeepRL: does the choice of action space matter?[C]//Proceeding of ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Los Angeles, USA, 2017: 12−20.

[39] VAN HASSELT H. Double q-learning[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Red Hook, USA, 2010: 2613−2621.

[40] HA D, SCHMIDHUBER J. World Models[EB/OL]. (2018-05-09). https://arxiv.org/abs/1803.10122.

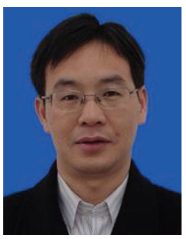
[41] EBERT F, FINN C, DASARI S, et al. Visual foresight: model-based deep reinforcement learning for vision-based robotic control[EB/OL]. (2018-12-03). https://arxiv.org/abs/1812.00568.

[42] FINN C, RAJESWARAN A, KAKADE S, et al. Online meta-learning[EB/OL]. (2019-07-03). https://arxiv.org/abs/1902.08438.

[43] MAHJOURIAN R, MⅡKKULAINEN R, LAZIC N, et al. Hierarchical policy design for sample-efficient learning of robot table tennis through self-play[EB/OL]. (2019-02-17). https://arxiv.org/abs/1811.12927?context=cs.

作者简介:

郭宪, 讲师, 博士, 主要研究方向为仿生机器人设计与智能运动控制。主持国家自然科学基金项目 1 项, 省部级项目 2 项。



方勇纯, 教授, 博士生导师, 南开大学人工智能学院院长, 主要研究方向为机器人视觉控制、欠驱动吊运系统控制、仿生机器人运动控制和微纳米操作。主持国家重点研发计划项目、国家基金重点项目、“十二五”国家技术支撑计划课题、国家基金仪器专项等项目。获吴文俊人工智能自然科学奖一等奖、天津市专利奖金奖、天津市自然科学一等奖、高等教育教学成果一等奖等多项奖励, 发表学术论文 100 余篇。