机器人操作技能学习方法综述

刘乃军 [1,2] 鲁 涛 [1,2] 蔡莹皓 [1,2] 王 硕 [1,2,3]

摘 要 结合人工智能技术和机器人技术, 研究具备一定自主决策和学习能力的机器人操作技能学习系统, 已逐渐成为机器人研究领域的重要分支. 本文介绍了机器人操作技能学习的主要方法及最新的研究成果. 依据对训练数据的使用方式将机器人操作技能学习方法分为基于强化学习的方法、基于示教学习的方法和基于小数据学习的方法, 并基于此对近些年的研究成果进行了综述和分析, 最后列举了机器人操作技能学习的未来发展方向.

关键词 机器人, 操作技能, 强化学习, 示教学习, 小数据学习

引用格式 刘乃军, 鲁涛, 蔡莹皓, 王硕. 机器人操作技能学习方法综述. 自动化学报, 2019, **45**(3): 458−470

DOI 10.16383/j.aas.c180076

A Review of Robot Manipulation Skills Learning Methods

LIU Nai-Jun[1,2] LU Tao[1,2] CAI Ying-Hao[1,2] WANG Shuo[1,2,3]

Abstract Designing a robot manipulation skill learning system with autonomous reasoning and learning ability has gradually become an important branch of robotics research field in combination with artificial intelligence and robotics technology. In this paper, the main methods and the latest research results of robot manipulation skills learning methods are introduced. We divide the learning methods into three categories, namely reinforcement learning approach, demonstration learning approach, and few data learning approach. Achievements of the robot manipulation skills learning areas based on these methods are discussed thoroughly. Finally, the future research directions are listed.

Key words Robots, manipulation skills, reinforcement learning, imitation learning, few data learning

Citation Liu Nai-Jun, Lu Tao, Cai Ying-Hao, Wang Shuo. A Review of robot manipulation skills learning methods. *Acta Automatica Sinica*, 2019, **45**(3): 458−470

各式机器人正逐渐应用于家庭、工厂、国防以及外太空探索等领域[1−2], 具备诸如衣服整理、机械零件装配、炸弹拆除等操作技能. 随着机器人技术的发展, 人们期望机器人具备更强的自主操作能力, 在更多领域代替人类完成更加复杂的操作任务. 在人工分析机器人行为特性和工作任务要求的基础上, 采用传统复杂编程、遥操作或示教编程等常规方法可使机器人具备一定的操作技能, 较好地胜任于诸多结构化工作环境和单一固定任务的工作场景, 快速准确地完成可重复位置和力控制的任务. 然而伴随机器人应用领域的不断扩大, 机器人往往会面临未知、动态及难预测的复杂环境. 采用传统常规方法设计的机器人操作技能不能动态地适应该类非结构化工作环境或场景多变的工作场合, 且机器人操作技能开发过程中存在周期长、效率低、工作量大及不能满足需求的多样性等诸多难题[3]. 随着人工智能技术研究的快速发展及关键技术的突破, 采用机器学习方法[4−5] 设计具备一定自主决策和学习能力的机器人操作技能学习系统, 使机器人在复杂、动态的环境中学习并获取操作技能, 能弥补传统编程等常规方法的缺陷, 极大提高机器人对环境的适应能力. 机器人操作技能学习作为未来机器人应具备的重要性能之一, 对未来机器人技术的发展具有重要意义, 是未来机器人在各领域得以广泛应用的重要基础. 近年来, 机器人操作技能学习研究正逐渐成为机器人研究领域的前沿和热点[6−8], 新的学习方法被逐渐应用于机器人的操作技能学习中, 诸多著名研究机构和公司, 如 DeepMind[9−10]、加州大学伯克利分校[11−12]、OpenAI[13−14]、Google Brain[15] 等在此领域取得了一定的成果, 但仍面临着巨大挑战. 本文针对近年来机器人操作技能学习领域的主要研究工作进行概述, 并以此为基础列举了机器人操作

收稿日期 2018-01-30 录用日期 2018-08-10
Manuscript received January 30, 2018; accepted August 10, 2018
国家自然科学基金 (U1713222, 61773378, 61703401), 北京市科技计划 (Z171100000817009) 资助
Supported by National Natural Science Foundation of China (U1713222, 61773378, 61703401) and Beijing Municipal Commission of Science and Technology (Z171100000817009)
本文责任编委 魏庆来
Recommended by Associate Editor WEI Qing-Lai
1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190 2. 中国科学院大学 北京 100190 3. 中国科学院脑智卓越中心 北京 100190
1. State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. University of Chinese Academy of Sciences, Beijing 100190 3. Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190

技能学习未来的主要研究方向.

1　研究进展概述

机器人操作技能学习方法涉及众多机器学习算法, 机器人训练数据的产生方式决定了机器人学习所要采用的具体方法[16]. 机器人操作技能学习所需数据大致可由机器人与环境交互产生或由专家提供[5, 17]. 基于此, 本文将机器人操作技能学习方法分为基于强化学习的方法、基于示教学习的方法和基于小数据学习的方法 (如图 1 所示), 并基于该分类对机器人操作技能学习的研究现状进行概述和分析.

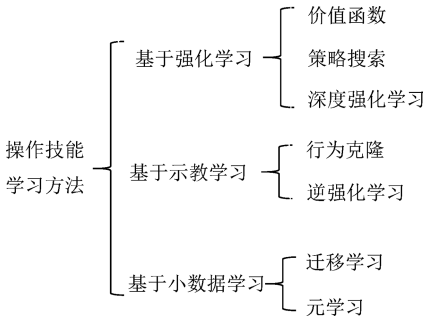


图 1　操作技能学习方法分类

Fig. 1　The categories of robot manipulation skills learning methods

1.1　基于强化学习

在基于强化学习的机器人操作技能学习中, 机器人以试错的机制与环境进行交互, 通过最大化累计奖赏的方式学习到最优操作技能策略[18−19]. 该类方法分为执行策略、收集样本及优化策略三个阶段, 如图 2 所示.

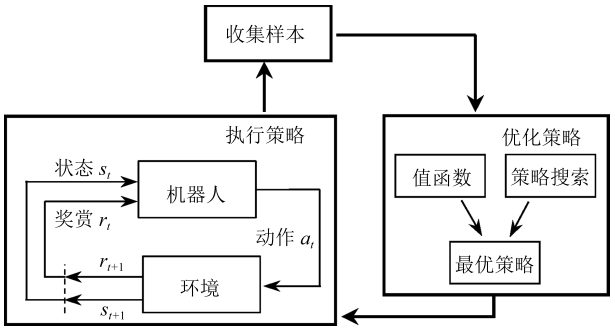


图 2　基于强化学习的操作技能学习示意图

Fig. 2　Illustration of manipulation skills learning method based on reinforcement learning

在执行策略阶段, 机器人在状态 s_t 依据当前策略 π 执行动作 a_t 得到奖赏值 r_{t+1} 并根据状态转移概率 $p(s_{t+1}/s_t, a_t)$ 到达新状态 s_{t+1}, 重复该过程, 直到机器人到达终止状态.

在收集样本阶段, 得到轨迹序列 $\tau: s_0, a_0, s_1, a_1, \cdots, s_H$, 其中 H 为轨迹序列长度. 机器人在环境中执行策略 π 后, 所得累计奖赏值 $R(\tau)$ 为

$$R(\tau) = \sum_{t=0}^{H} \gamma^t r_t, \quad 0 < \gamma \leq 1 \tag{1}$$

其中, γ 为折扣因子. 机器人在状态 s 对应的价值函数 $V^\pi(s)$ 表示其在状态 s 执行策略 π 后得到的累计奖赏值.

$$V^\pi(s) = \mathrm{E}\left[\sum_{k=0}^{H-t} \gamma^k r_{t+k} | s_t = s; \pi\right] \tag{2}$$

在状态 s 实施动作 a 后得到的动作−状态值函数 $Q^\pi(s, a)$ 的定义为

$$Q^\pi(s, a) = \mathrm{E}\left[\sum_{k=0}^{H-t} \gamma^k r_{t+k} | s_t = s, a_t = a; \pi\right] \tag{3}$$

由贝尔曼 (Bellman) 方程[20] 可得动作−状态值函数的迭代关系式为

$$Q^\pi(s_t, a_t) = \mathrm{E}_{s+1}\left[r_{t+1} + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1}))\right] \tag{4}$$

机器人在状态 s_t 所要执行的最优动作 a_t^* 为

$$a_t^* = \arg\max_{a_t} Q^\pi(s_t, a_t) \tag{5}$$

在策略优化阶段, 对机器人操作技能策略进行优化. 依据最优动作的获得是否需要价值函数 $V^\pi(s)$ 或动作−状态值函数 $Q^\pi(s, a)$, 将强化学习方法分为值函数强化学习和策略搜索强化学习. 近年来, 随着深度学习的发展, 诸多学者采用由深度学习和强化学习结合得到的深度强化学习方法来获取机器人的操作技能策略.

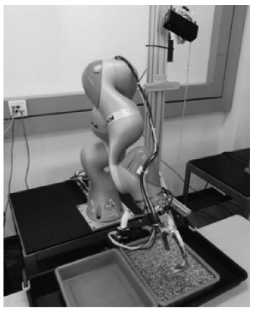
1.1.1　值函数强化学习方法

值函数强化学习方法依据机器人与环境交互是否需要依靠先验知识或交互数据学习得到系统的状态转移模型, 可分为基于学习模型的值函数方法和基于无模型的值函数方法.

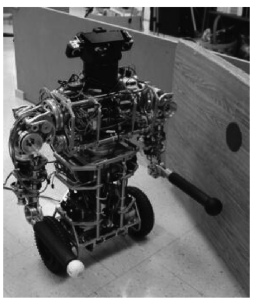
1) 基于学习模型的值函数强化学习. Lioutikov 等[21] 基于局部线性系统估计 (Local linear system estimation) 得到系统的状态转移概率模型, 实现了二连杆机械臂对乒乓球拍的操作 (如图 3 (a) 所示). Schenck 等[22] 基于卷积神经网络结构建立了推断挖取和倾倒动作的预测模型, 实现了 KUKA 机器人挖取豆粒物体的操作技能任务 (如图 3 (b) 所示). Hester 等[23] 基于决策树得到系统的状态转移概率模型, 实现了人形机器人踢足球的操作技能.

2) 基于无模型的值函数强化学习. 机器人各状态的价值函数采用诸如蒙特卡洛[24]、TD(λ)[25]、Q-leaning[26] 及 SARSA[27] 等算法进行估计, 进而得到各状态的最优动作. Konidaris 等[28−29] 基于 CST (Constructing skill tree) 算法将机器人所要执行的任务序列化, 完成了机器人在室内环境中移动到指定位置并执行开门的操作任务 (如图 3 (c) 所示). Asada 等[30] 基于视觉信息构建得到了机器人工作环境中目标物体的几何尺寸及方位信息, 采用 Q-leaning 算法成功实现了机器人将球击打到指定位置的操作任务. Kroemer 等[31] 提出了一种基于强化学习和视觉反馈策略的混合控制器, 以处理抓取任务中的不确定性问题, 成功实现了机器人抓取不同种类物体的任务目标 (如图 3 (d) 所示).

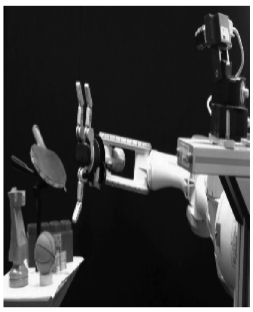
(a) 击打乒乓球[21]
(a) Hitting table tennis[21]



(b) 挖取豆粒[22]
(b) Scooping beans[22]



(c) 导航[28]
(c) Navigation[30]



(d) 抓取[31]
(d) Grasping[31]

图 3 基于值函数强化学习的操作技能
Fig. 3 Manipulation skills based on value function of reinforcement learning

总体而言, 基于无模型的值函数方法不需对系统建模, 计算量小, 但价值函数的获取需要通过机器人与环境的不断交互采样估计得到. 基于学习模型的值函数方法首先需要依据机器人与环境的交互数据学习得到系统模型, 并基于该模型采用仿真形式得到最优策略, 故其在真实环境中所需的样本少, 但计算量大.

1.1.2 策略搜索强化学习方法

与基于通过价值函数推导间接得到最优策略不同, 基于策略搜索的强化学习算法直接基于给定的策略评价函数在策略空间内搜索得到最优控制策略. 将策略表示为参数 θ 的函数 π_θ, 则对策略的优化间接转化为对参数 θ 的优化. 给定的策略评价函数为

$$\eta(\theta) = \mathrm{E}\left[\sum_{t=0}^{H} r(s_t, a_t)|\pi_\theta\right] \tag{6}$$

依据策略搜索是否需要求导, 可将策略搜索分为免求导方法和策略梯度方法. 常见的免求导方法包含 CEM (Cross-entropy method)[32]、CMA (Covariance matrix adaptation)[33] 等. 策略梯度方法通过求解策略评价函数关于参数 θ 的导数, 得到策略参数 θ 的搜索方向 $\nabla_\theta\eta(\theta)$

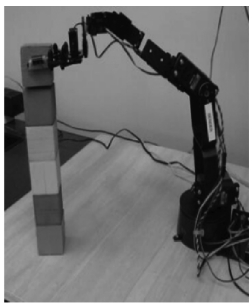
$$\nabla_\theta\eta(\theta) = \nabla_\theta \sum_\tau p(\tau;\theta)R(\tau) = \sum_\tau p(\tau;\theta)\nabla_\theta \log p(\tau;\theta)R(\tau) \tag{7}$$

其中, $p(\tau;\theta)$ 表示执行策略 π_θ 得到轨迹 τ 的概率分布. 进而得到更新后的策略参数 θ_{i+1} 为

$$\theta_{i+1} = \theta_i + \alpha\nabla_\theta\eta(\theta) \tag{8}$$

其中, α 为更新步长. Endo 等[34] 基于策略梯度, 实现了双足机器人行走的操作技能任务. Peters 等[35] 将策略梯度与运动基元相结合, 训练得到了机械臂击打棒球的操作技能策略 (如图 4 (a) 所示). Deisenroth 等[36] 提出了一种基于模型的策略搜索方法, 将深度相机提供的环境图像信息和机器人操作任务的空间约束加入到学习过程, 实现了机器人搭积木的操作任务 (如图 4 (b) 所示), 之后采用高斯过程[37] 建立系统状态转移概率模型, 减小了模型偏差对机器人操作技能学习的不利影响.

(a) 击打棒球[35]
(a) Hitting baseball[35]



(b) 搭积木[36]
(b) Building blocks[36]

图 4 基于策略搜索强化学习的操作技能
Fig. 4 Manipulation skills based on policy search of reinforcement learning

相较而言, 在机器人操作技能学习领域, 策略搜索比基于价值函数的强化学习方法更具优势, 主要

体现在: 1) 采用策略搜索方法可以较为方便地融入专家知识, 可依据获取的专家策略对神经网络参数进行初始化, 以加速策略优化的收敛过程; 2) 策略函数比价值函数具有更少的学习参数, 基于策略搜索的强化学习算法的学习效率更加高效[38].

1.1.3　深度强化学习方法

基于深度神经网络的深度学习作为机器学习领域的新分支, 通过组合低层特征形成更加抽象的高层表示, 得到数据的分布式特征. 近年来, 诸多学者将深度学习和强化学习相结合得到的深度强化学习算法[39] 成功应用于视频游戏[40] 和围棋[41−42] 等领域.

1) 基于价值函数的深度强化学习. DeepMind[40] 提出的 DQN (Deep Q-network) 首次在视频游戏领域超越了人类游戏玩家. DQN 神经网络结构示意图如图 5 所示, 输入是距离当前时刻最近的若干帧图像, 经过若干层卷积网络和全连接网络非线性变换后, 最后输出各动作对应的状态−动作值. 其通过最小化误差函数

$$L_i(\theta_i) = \mathrm{E}_{s,a,r,s'}[y_i - Q(s,a;\theta_i)] \tag{9}$$

对网络参数进行更新, 式中 y_i 为目标状态−动作值.

$$y_i = r + \gamma \max_{a'} Q(s',a';\theta_{i-1}|s,a) \tag{10}$$

其中, θ_i 为第 i 次迭代更新后的网络参数值. 为了防止学习过程中过高估计动作−状态值, van Hasselt 等[43] 提出了双 DQN (Double DQN), 其目标状态−动作值为

$$y_i = r + \gamma Q(s', \arg\max_{a'} Q(s',a';\theta_{i-1}|s,a)) \tag{11}$$

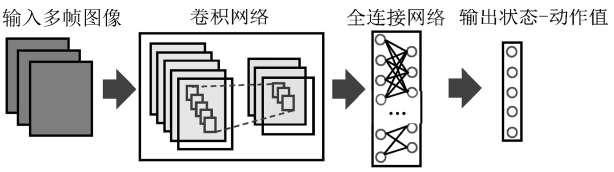


图 5　DQN 网络结构示意图

Fig. 5　Illustration of DQN neural network

之后竞争网路 (Dueling network)[44] 和深度循环网络 (Deep recurrent network)[45] 相继被提出. Zhang 等[46−47] 创建虚拟训练环境将 DQN 算法用于训练三关节机器人抓取任务的控制策略, 然而由于训练环境与真实场景存在一定差异并且其将动作空间进行了离散化, 导致训练后的控制器在真实场景下的抓取效果欠佳. Google Brain 和 DeepMind 联合提出了基于连续动作空间和学习模型的 DQN 改进算法[48], 在虚拟环境中成功实现了机器人抓取、夹手移动等操作任务.

2) 基于策略搜索的深度强化学习. 为解决连续动作空间上的控制问题, Lillicrap 等[9] 通过对确定性策略梯度 (Deterministic policy gradient, DPG)[49] 方法进行改造, 提出了一种基于 Actor-Critic 框架的深度确定性策略梯度 (Deep deterministic policy gradient, DDPG) 算法, 并在模拟仿真环境 Mujoco 中实现了机器人的抓取操作任务目标. 为了保证策略优化过程中性能渐进提高, Schulman 等[50] 提出了 TRPO (Trust region policy optimization) 算法, 其通过优化目标函数

$$\begin{aligned} &\max_{\theta} \mathrm{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta\mathrm{old}}(a_t|s_t)}\hat{A}_t\right] \\ &\text{s.t.} \quad \mathrm{E}_t\left[KL\left[\pi_\theta(\cdot|s_t), \pi_{\theta\mathrm{old}}(\cdot|s_t)\right]\right] \le \delta \end{aligned} \tag{12}$$

对策略参数进行更新, 式中 $\hat{A}_t$ 为优势函数 (Advantage function) 在时刻 t 的估计值, π_θ, $\pi_{\theta\mathrm{old}}$ 分别表示在同一批次训练数据上优化前后的新旧策略, δ 为较小值, 用于限制新旧策略分布的 KL 散度差异. TRPO 算法被成功应用于虚拟场景下的机器人操作技能学习. 随后, DeepMind 和 OpenAI 提出了基于 TRPO 一阶近似形式的改进型算法 PPO (Proximal policy optimization)[10, 13], 在虚拟仿真环境机器人的操作技能学习中取得了优于 TRPO 的效果. 基于异步梯度下降形式 actor-critic 的 A3C (Asynchronous advantage actor-critic)[51] 算法也被用于机器人的操作技能策略学习.

鉴于在策略优化的每个迭代步中, 都需要采集一定量的训练数据来更新策略, 而在真实机器人工作场景中, 训练数据的获取成本高昂, 为此加州大学伯克利分校的 Levine 等[11−12, 52−53] 提出了引导策略搜索 (Guided policy search, GPS) 算法, 通过使用优化轨迹分布来生成具有引导作用的训练样本, 并采用监督学习方法训练神经网络策略. 之后 Levine 等[12] 又将环境的图像信息作为机器人策略状态的一部分, 进行端到端的训练, 获取了机器人抓取、搭衣服等多种操作技能 (如图 6 所示).

与常规强化学习方法相比, 深度强化学习算法将具有强表征能力的深度神经网络用于强化学习中价值函数和策略函数的表达, 避免了人为手工设计特征, 同时也易融入环境中的图像感知信息, 较适合于机器人操作技能学习.

强化学习方法在机器人的操作技能学习领域得到了广泛的应用, 基于机器人操作技能学习的任务特点, 应用于机器人操作技能学习领域的强化学习有别于其他应用领域的不同之处, 主要体现在其状态及动作空间均为高维连续空间、收集训练样本代

价高等方面, 具体如表 1 所示.

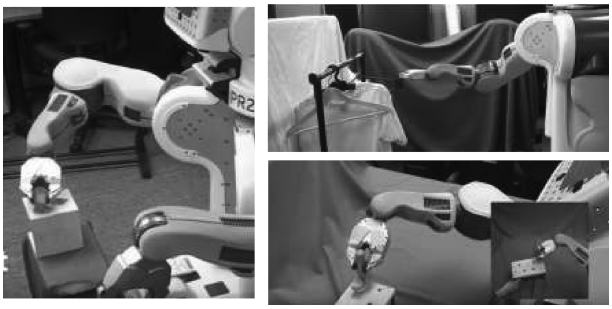


图 6 基于引导策略搜索的机器人操作技能[12]
Fig. 6 Manipulation skills based on guided policy search[12]

表 1 机器人和其他应用中强化学习比较
Table 1 Comparison of reinforcement learning methods applied in robotics and other fields

项目	机器人应用	其他应用
状态、动作空间	均为高维、连续空间	大多为低维、离散空间
训练数据获取	真实环境: 数据获取会损耗硬件, 有潜在危险, 成本高; 虚拟环境: 数据获取方便	不损耗硬件不存在危险性
训练成本	仿真环境低, 真实环境高	低
主流方法	大多基于策略搜索	大多基于价值函数
其他方面	不确定性因素多, 训练过程受诸多条件约束, 学习过程需要人的参与	—

1.2 基于示教学习

在机器人操作技能学习领域, 示教学习通过模仿给定的专家数据学习得到操作技能策略. 示教学习可降低机器人搜索策略空间的复杂度, 在一定程度上提高了机器人操作技能的学习效率. 近年来, 示教学习已成为机器人操作技能学习的热点领域之一[54]. 依据对示教数据的使用方式, 大致可将示教学习分为行为克隆 (Behavior cloning)[55] 和逆强化学习 (Inverse reinforcement learning)[56] 两大类, 如图 7 所示.

行为克隆是基于给定的多个示教轨迹序列 $\tau_1, \tau_2, \cdots, \tau_m$, 其中 τ_i 为 $\{s_1^i, a_1^i, s_2^i, a_2^i, \cdots, s_{n_i}^i\}$, n_i 为轨迹 τ_i 的轨迹长度, 收集得到状态−动作对样本集合 D[57]

$$D = \{(s_1, a_1), (s_2, a_2), \cdots, (s_{\sum_{i=1}^{m} n_i}, a_{\sum_{i=1}^{m} n_i})\} \quad (13)$$

采用常见的监督学习方法, 直接学习到状态到动作的映射关系. 日本东北大学基于隐马尔科夫模型 (Hidden Markov model, HMM)[58] 训练得到了能与人共跳华尔兹舞的机器人策略. Calinon 等[59] 基于高斯混合模型 (Gaussian mixture model, GMM) 学习到机器人移动棋子以及抓取糖块并放到嘴里的操作技能, 之后该课题组又通过可穿戴式运动传感器采集示教数据, 采用高斯混合回归 (Gaussian mixture regression, GMR)[60], 实现了人形机器人完成篮球裁判员诸多判罚动作的操作机能. Rahmatizadeh 等[61] 通过在虚拟仿真环境中采集大量示教数据训练递归神经网络 (Recurrent neural networks, RNN) 策略, 在真实机械臂上实现了抓取不同位置物体的目标. Calinon 等[62] 通过结合隐马尔科夫模型、高斯混合回归与机器人的系统动态特性建立冗余策略模型, 实现了机器人击打乒乓球的操作任务. Levine 等[15] 通过在多台机械臂上收集大量抓取种类各异物体数据 (如图 8 所示), 对深度卷积网络控制策略进行训练, 在无需对相机标定的情况下, 实现了高效准确抓取不同物体的目标. Zhang 等[63] 采用 VR 虚拟设备采集示教数据 (如图 9 所示), 通过监督学习训练神经网络控制策略, 实现了 PR2 机器人抓取、到达指定位置等若干操作技能.

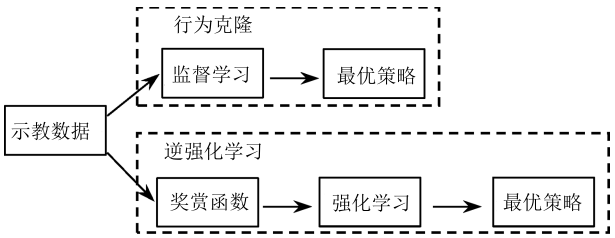


图 7 示教学习分类示意图
Fig. 7 Illustration of classification of imitation learning methods

图 8 多台机器人收集训练数据[15]
Fig. 8 Collecting training data by many robots[15]

图 9　基于 VR 虚拟现实设备的示教学习[63]

Fig. 9　Imitation learning based on VR device[63]

在有限样本条件下, 直接基于监督学习得到的策略适用性不强, 逆向强化学习能够基于给定的有限示教数据反推得到奖赏函数, 从而提高学习策略的泛化性能. 逆强化学习分为两个阶段, 第一阶段基于给定的示教轨迹推导出能使示教轨迹最优的奖赏函数, 第二阶段基于推导出的奖赏函数采用强化学习算法得到机器人执行该示教操作任务的技能策略. Abbeel 等[64] 提出了依据示教数据得到奖赏函数的最大边际原则 (Max margin principle), 依据该原则可使基于奖赏函数学习到的最优策略和其他次优策略之间的差异最大. Ratliff 等[65] 基于最大边际原则提出了最大边际规划框架, 将奖赏函数的学习问题转化为结构化预测问题, 并通过四足机器人对该方法进行了验证. 然而, 基于最大边际原则得到的奖赏函数往往存在二义性问题, 同时基于真实机器人得到的示教数据往往混有噪声, 导致在一些机器人的应用场景中效果不佳. 为此, Ziebart 等[66] 基于最大熵原则构建了序列决策的概率模型奖赏函数, 能保证在示教数据非最优及示教数据混有噪声的情况下, 机器人控制策略也具有较优的性能表现. 上述均为基于线性特征得到奖赏函数的方法, 基于非线性特征的方法如高斯过程[67]、boosting[68] 也被用来求解示教轨迹中潜在的奖赏函数, 其表现效果在一些任务领域优于基于线性特征得到奖赏函数.

为了避免人工设计奖赏函数特征, 同时保证易于处理机器人状态为高维、连续空间, 深度神经网络[69−70] 已逐渐应用于奖赏函数的表达.

此外, Finn 等[71] 提出了引导式奖赏函数的逆强化学习方法, 将奖赏函数作为优化目标生成接近专家示例轨迹数据的奖赏函数. Ho 等[72] 采用生成式对抗网络 (Generative adversarial networks, GAN)[73] 的思想, 将奖赏函数的优化比作判别器, 同时将策略的优化比作生成器, 使奖赏函数优化与策略优化交替迭代以生成能够判别示教轨迹为较优轨迹的奖赏函数. 加州大学伯克利分校提出了 deepmimic 算法[73], 给定示教范例, 采用强化学习中的 PPO 算法[13] 对虚拟仿真环境中的人形机器人等进行训练, 实现了武术、跳舞及多种杂技等高难度操作技能 (如图 10 所示).

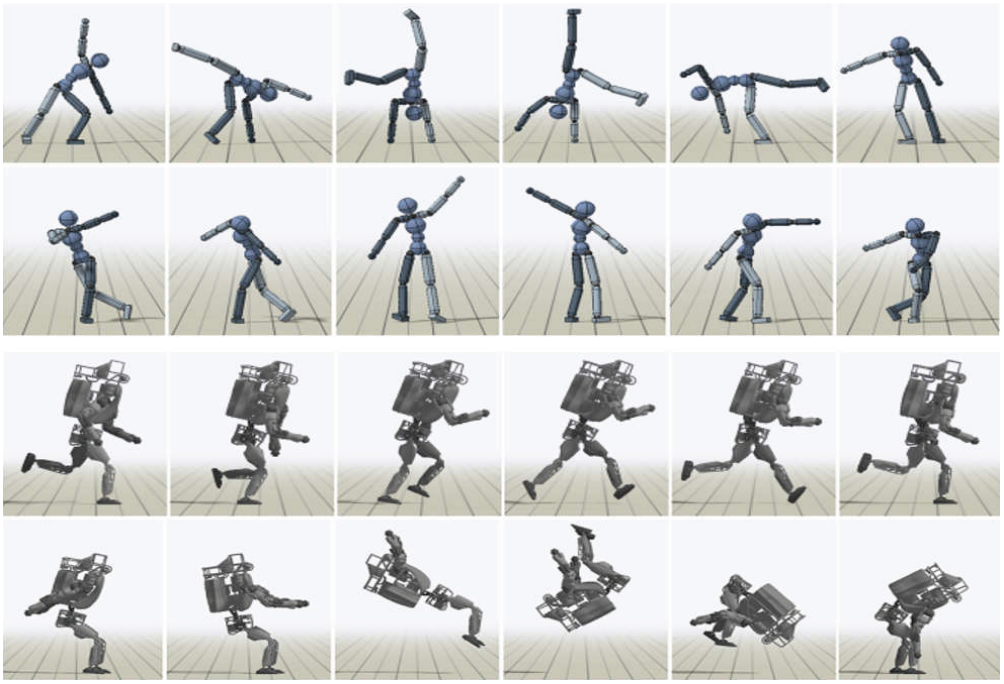


图 10　人形机器人高难度操作技能[73]

Fig. 10　Difficulty manipulation skills learned by human robots[73]

相比于强化学习方法策略起始状态的随机导致的学习效率低, 示教学习方法基于示教数据对策略进行初始化, 可加快机器人操作技能学习速率. 然而示教学习中也存在收集示教数据成本高昂和训练所得策略易陷入局部最优解的问题, 从而可能导致机器人操作技能的学习效果欠佳. 为此有学者将示教学习与强化学习相结合, 以更加高效地获取机器人的操作技能. Zhu 等[74] 提出了无模型的深度强化学习方法, 采用强化学习与示教学习相结合的方式在合成的逼真虚拟仿真环境中对神经网络进行训练, 之后将训练得到的策略直接应用到真实环境中 (如图 11 所示). Hester 等[75] 提出了一种将示教数据添加到 DQN 回放记忆单元 (Replay memory) 中的示教学习方法, 提升了操作技能学习效率.

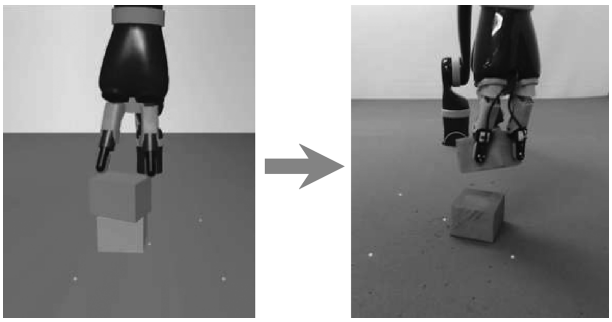


图 11 虚拟环境中训练策略应用于真实环境[74]

Fig. 11 Policies trained in simulated environment applied in real-world environment[74]

1.3 基于小数据学习

无论是基于强化学习还是基于示教学习的机器人操作技能学习方法都需要一定量的训练数据. 使用少量训练数据就可学习到新的操作技能成为了机器人快速应用于各领域的关键.

近年来发展的迁移学习 (Transfer learning) 和元学习 (Meta learning)[76] 具有利用先前数据经验的机制, 在面对新任务少量数据时, 能够实现基于小样本数据的快速任务学习.

迁移学习是从一个或多个源域 (Source domain) 中抽取知识、经验, 然后应用于目标域 (Target domain) 的学习方法[77], 已在诸如计算机视觉[78−79] 及控制[80−81] 等领域取得了一定的进展. 在机器人操作技能学习领域, 迁移学习可将基于一种或多种任务上学习到的能力迁移到另一种新的任务上, 以提高机器人操作技能的学习效率. Ammar 等[82] 提出了一种基于策略梯度的多任务学习方法, 通过从不同的工作任务中迁移知识实现了机器人的高效学习. Gupta 等[83] 通过构建多个机器人之间共有的特征空间, 采用多任务学习的形式在虚拟仿真环境中实现了将 3 连杆机器人抓取、移动指定物体的操作技能通过少量数据迁移给 4 连杆机器人的目标. Tzeng 等[84] 通过在虚拟环境中合成与真实环境中相对应的图像信息对机器人的操作技能进行训练, 之后采用迁移学习的方式将机器人的操作技能应用于真实环境中.

机器人的迁移学习在一定程度上可提高机器人学习操作技能的效率, 然而在面对新任务时, 仍然需要以机器人与环境进行一定的交互为前提, 即仍然不能使机器人通过一次或极少次示教数据成功学习到新的操作技能.

元学习 (Meta learning) 及以此为基础的一次性学习 (One-shot learning) 是一种基于少量训练数据对模型进行学习的机器学习方法. 元学习通过在大量相关任务且每种任务包含少量标记数据的任务集上对策略进行训练, 能够自动学得训练任务集中的共有知识. 诸多学者将该方法应用于图像识别[85−87]、生成式模型[88−89]、强化学习中智能体的快速学习[90−91] 等领域. 还有一些学者尝试将元学习应用在机器人操作技能学习领域. Duan 等[92] 提出了一次性模仿 (One-shot imitation) 学习方法 (如图 12 所示), 基于多种任务采用元学习算法训练得到元学习策略, 学习完成后基于新任务的一次示教就可完成执行新任务的操作技能, 并通过搭积木的操作任务验证了该方法的有效性. Finn 等[93] 提出了 MAML (Model-agnostic meta-learning) 元学习方法, 通过多种任务采用梯度下降方法对同一个深度网络策略模型的参数进行元学习更新, 利用少量训练数据和较少步的梯度下降更新策略参数进行新任务学习 (如图 13 所示), 在虚拟仿真环境中快速学习到了机器人的前进、后退等操作技能. OpenAI[14] 基于策略梯度提出了一种适用于动态环境中的元学习算法, 在虚拟环境中实现了多种构型机器人之间的竞争操作技能学习.

另外, 一些学者提出了面对新任务少数据学习的其他方法. Xu 等[94] 通过采用神经网络推理方

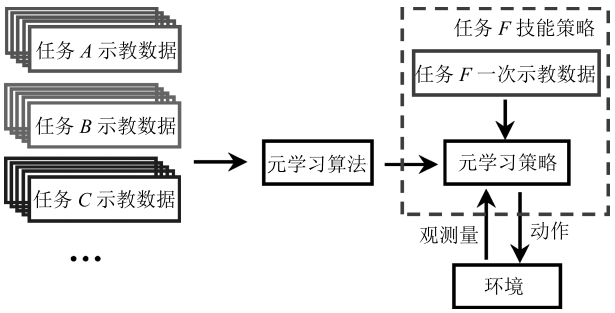


图 12 一次性模仿学习算法示意图[92]

Fig. 12 Illustration of one-shot imitation learning algorithm[92]

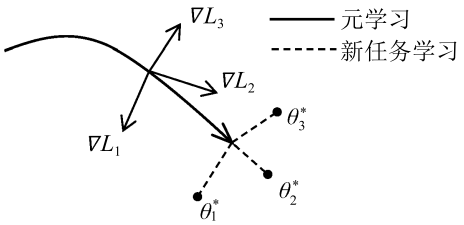


图 13　MAML 元学习方法策略参数梯度更新示意图[93]

Fig. 13　Illustration of gradient update for policy parameters with MAML meta learning algorithm[93]

法[95] 将机器人的操作技能任务进行分解, 在采用大量监督数据对模型训练的基础上, 通过在虚拟环境中进行一次示教, 就可使机器人完成诸如整理餐桌等操作任务. Tobin 等[96] 提出了域随机化 (Domain randomization) 方法, 通过在虚拟环境中改变物体的纹理、光照以及相机的位置等条件对神经网络进行训练, 之后不需额外数据训练即可将在虚拟环境中训练得到的策略直接应用到了真实环境中.

在机器人操作技能学习领域, 迁移学习及元学习都可认为是通过少量数据学习到新操作技能的方法, 但不同之处在于, 迁移学习是将机器人在某一或某几种任务上已经学习好的技能迁移到新任务上, 元学习是通过大量任务对元学习策略进行训练, 基于新任务的少量数据实现机器人操作技能策略的跨任务泛化.

本文将机器人操作技能学习方法分为基于强化学习的方法、基于示教学习的方法和基于小数据学习的方法, 并基于此进行了综述分析, 基于机器人操作技能策略训练数据的使用量、学习效率和学习成本的对比如表 2 所示.

表 2　三类操作技能学习方法特点对比

Table 2　Comparison of three kinds of manipulation skills learning methods

对比项目	基于强化学习	基于示教学习	小数据学习
数据量	不需提供示教数据但需大量机器人与环境的交互数据	需提供较多示教数据	需大量数据面对新任务需少量数据
学习效率	低, 需不断试错	较高	高
学习成本	高	高	低

2　未来发展方向

通过分析已有的机器人操作技能学习研究工作, 机器人操作技能学习问题主要聚焦于两方面: 1) 如何使机器人学习得到的技能策略具有更好的泛化性能; 2) 如何采用较少的训练数据、较低的训练代价学习得到新的操作技能. 如何解决这两方面的问题是机器人操作技能学习的研究重点. 为此, 本文列举了如下的未来研究方向.

2.1　高效学习算法设计

以兼俱感知、决策能力的深度强化学习为核心算法的机器学习方法在机器人操作技能学习领域取得了一定进展, 但由于采用深度学习方法对价值函数或策略函数进行拟合, 通常需要通过多步梯度下降方法进行迭代更新, 采用强化学习得到机器人不同状态所要执行的最优动作也需机器人在环境中经过多步探索得到, 这就导致了该类算法的学习效率较低. 例如人类花费数小时学会的操作技能, 机器人需花费数倍时间才能到达同等水平.

现有的深度强化学习算法, 诸如 DQN, DDPG, A3C, TRPO, PPO 等均为通用的深度强化学习算法, 既能适用于电子游戏, 也能适用于虚拟环境下的机器人控制策略训练. 但在机器人实际操作环境中, 存在数据样本获取困难、数据噪声干扰大等特点, 导致现有操作技能学习方法学习效率低, 学习效果欠佳. 因此, 结合机器人操作技能学习的固有特性及先验知识设计高效学习算法, 实现有限样本下操作技能策略的快速迭代和优化对于机器人操作技能学习具有重要价值.

2.2　技能迁移学习

基于机器人操作技能学习中的迁移学习主要包含两个方面: 1) 基于环境, 将虚拟环境中学习到的操作技能迁移到真实环境中; 2) 基于任务, 将在一种任务上学习到的操作技能迁移到另一种任务上.

在仿真环境中, 机器人操作技能学习的训练成本低廉, 并可避免使用真实机器人训练所带来的诸多不便性和危险性. 但由于仿真环境与机器人真实工作场景不同, 导致仿真环境中学习到的操作技能策略在真实环境中表现效果欠佳, 为此如何将在虚拟环境中学习到的策略较好地应用于真实环境是机器人操作技能学习中研究的关键问题之一.

通过基于一种或多种任务学习的技能策略初始化新任务技能策略, 可加快机器人对新任务操作技能策略的学习效率, 但这仅限于机器人的任务类型和工作环境存在极小差异的情况. 为此如何在具有一定差异的不同任务之间实现操作技能的迁移, 并且避免可能出现的负迁移 (Negative transfer) 现象, 也是机器人操作技能学习中要解决的重要问题.

2.3　层次化任务学习

在机器人的操作技能学习任务中, 复杂操作任务都可以分解成若干简单子任务. 例如机器人倒水操作任务可以分解成机器人从当前位置移动到水杯

位置、机器人末端夹手抓住水杯、移动机器人到指定容器位置、转动末端夹手将水倒入容器中. 机器人开门操作任务可以分解为移动机器人夹手到门把手位置、夹手抓住门把手、转动末端夹手将门打开. 上述任务虽不相同, 但均包含机器人末端执行器到达、末端夹手夹持等子任务, 为此对机器人要执行的任务进行层次化分解可有利于操作技能的学习. 针对复杂操作技能任务, 训练学习将复杂任务分解成多个子任务的高级策略和执行子任务的低级策略, 可使操作技能的学习过程更加高效.

2.4 元学习

元学习作为一种学会学习 (Learning to learn) 的方法, 在机器人操作技能学习领域已取得了一定的进展. 将元学习思想应用于机器人操作技能学习领域可能存在的问题基于两方面: 1) 要确定机器人操作技能学习的训练环境和训练数据集的数据形式; 2) 是设计适宜的元学习网络结构. 目前在计算机视觉领域, 研究者提出了多种类型神经网络结构, 而在基于机器人操作技能学习领域的特定神经网络结构还不多见. 为此借鉴其他研究领域, 设计学习效率高, 性能优异的元学习神经网络结构是机器人操作技能学习的重要研究方向.

元学习作为一种少数据学习方法, 当前还仅限于面对新任务的测试阶段需少量数据, 而在元学习的训练阶段, 仍需提供大量训练数据. 为此基于训练环境、训练数据形式及网络结构等方面, 设计高效的元学习训练算法, 实现真正的少数据学习, 是机器人操作技能学习的未来发展方向之一.

3 结论

相比于传统复杂编程、遥操作及示教编程等常规方法, 机器人操作技能学习方法可使机器人具备一定的决策和学习能力, 动态地适应诸多非结构化工作环境或场景多变的工作场合, 是机器人能够广泛应用于各领域的基础. 机器人操作技能学习作为机器人研究领域的前沿方向吸引了诸多学者的研究兴趣.

目前, 人工智能技术的发展为机器人操作技能的学习提供了新的方法, 开拓了新的思路. 相比于计算机视觉、自然语言处理、语音识别等领域, 机器人的操作技能学习所需代价更高、成本更大. 因此, 基于如何使机器人的操作技能学习更加高效, 如何使学习的操作技能策略泛化性能更强等问题的研究, 也将对机器学习及人工智能技术的发展起到促进作用. 近年来, 人工智能技术中的深度学习技术已开始广泛应用于机器人操作技能学习领域, 除与强化学习结合外, 还应用于示教学习以及元学习中. 但由于机器人应用场景和操作技能学习的特殊性, 决定了应用于机器人领域的深度学习技术与其他应用领域具有不同的特性, 例如在机器人操作技能学习应用领域, 深度学习技术除应用于物体识别外还需进行物体的空间定位. 此外, 深度学习技术目前还缺乏一定的理论支持, 基于深度学习技术获取的机器人操作技能可解释性差, 在操作任务中需要的定位精确性、运动灵巧性和平稳性以及执行任务的实时性暂时还不能从理论上得到保证, 还需进一步开展相关的研究和论证.

References

1 Goldberg K. Editorial: "One Robot is robotics, ten robots is automation". *IEEE Transactions on Automation Science and Engineering*, 2016, **13**(4): 1418−1419

2 Tan Min, Wang Shuo. Research progress on robotics. *Acta Automatica Sinica*, 2013, **39**(7): 963−972
(谭民, 王硕. 机器人技术研究进展. 自动化学报, 2013, **39**(7): 963−972

3 Rozo L, Jaquier N, Calinon S, Caldwell D G. Learning manipulability ellipsoids for task compatibility in robot manipulation. In: Proceedings of the 30th International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017. 3183−3189

4 Siciliano B, Khatib O. *Springer Handbook of Robotics*. Berlin: Springer, 2016. 357−398

5 Connell J H, Mahadevan S. *Robot Learning*. Boston: Springer, 1993. 1−17

6 Dang H, Allen P K. Robot learning of everyday object manipulations via human demonstration. In: Proceedings of the 23rd IEEE International Conference on Intelligent Robots and Systems (IROS). Taipei, China: IEEE, 2010. 1284−1289

7 Gu S X, Holly E, Lillicrap T, Levine S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: Proceedings of the 35th IEEE International Conference on Robotics and Automation (ICRA). Singapore, Singapore: IEEE, 2017. 3389−3396

8 Li D Y, Ma G F, He W, Zhang W, Li C J, Ge S S. Distributed coordinated tracking control of multiple Euler-Lagrange systems by state and output feedback. *IET Control Theory and Applications*, 2017, **11**(14): 2213−2221

9 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Eraz T, Tassa Y, et al. Continuous control with deep reinforcement learning. arXiv: 1509.02971, 2015.

10 Heess N, Dhruva T B, Sriram S, Lemmon J, Merel J, Wayne G, et al. Emergence of locomotion behaviours in rich environments. arXiv: 1707.02286, 2017.

11 Levine S, Abbeel P. Learning neural network policies with guided policy search under unknown dynamics. In: Proceedings of the 28th Advances in Neural Information Processing Systems (NIPS). Montreal, Canada: NIPS Press, 2014. 1071−1079

12 Levine S, Finn C, Darrell T, Abbeel P. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 2016, **17**(1): 1334−1373

13 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv: 1707.06347, 2017.

14 Al-Shedivat M, Bansal T, Burda Y, Sutskever I, Mordatch I, Abbeel P. Continuous adaptation via meta-learning in nonstationary and competitive environments. In: Proceedings of the 6th International conference on Learning Representations (ICLR). Vancouver, Canada: ICLR, 2018.

15 Levine S, Pastor P, Krizhevsky A, Quillen D. Learning hand-eye coordination for robotic grasping with large-scale data collection. In: Proceedings of the 25th International Symposium on Experimental Robotics. Cham: Springer, 2016. 173−184

16 Calinon S. Robot learning with task-parameterized generative models. *Robotics Research*. Cham: Springer, 2018. 111−126

17 Billard A, Grollman D. Robot learning by demonstration. *Scholarpedia*, 2013, **8**(12): 3824

18 Wiering M, van Otterlo M. *Reinforcement Learning: State-of-the-Art*. Berlin: Springer-Verlag, 2015. 79−100

19 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction* (Second edition). Cambridge: MIT Press, 1998.

20 Bellman R. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 1952, **38**(8): 716−719

21 Lioutikov R, Paraschos A, Peters J, Neumann G. Sample-based informationl-theoretic stochastic optimal control. In: Proceedings of the 32nd IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 3896−3902

22 Schenck C, Tompson J, Fox D, Levine S. Learning robotic manipulation of granular media. In: Proceedings of the 1st Conference on Robot Learning (CORL). Mountain View, USA: CORL, 2017.

23 Hester T, Quinlan M, Stone P. Generalized model learning for reinforcement learning on a humanoid robot. In: Proceedings of the 28th IEEE International Conference on Robotics and Automation (ICRA). Alaska, USA: IEEE, 2010. 2369−2374

24 Kocsis L, Szepesvári C. Bandit based Monte-Carlo planning. In: Proceedings of the 2006 European Conference on Machine Learning. Berlin, Germany: Springer, 2006. 282−293

25 Hasselt H, Mahmood A R, Sutton R S. Off-policy TD(λ) with a true online equivalence. In: Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence. Quebec City, Canada: UAI, 2014.

26 Park K H, Kim Y J, Kim J H. Modular Q-learning based multi-agent cooperation for robot soccer. *Robotics and Autonomous Systems*, 2001, **35**(2): 109−122

27 Ramachandran D, Gupta R. Smoothed Sarsa: reinforcement learning for robot delivery tasks. In: Proceedings of the 27th IEEE International Conference on Robotics and Automation (ICRA). Kobe, Japan: IEEE, 2009. 2125−2132

28 Konidaris G, Kuindersma S, Grupen R, Barto A. Autonomous skill acquisition on a mobile manipulator. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI). San Francisco, California, USA: AAAI Press, 2011. 1468−1473

29 Konidaris G, Kuindersma S, Barto A G, Grupen R A. Constructing skill trees for reinforcement learning agents from demonstration trajectories. In: Proceedings of the 24th Advances in Neural Information Processing Systems (NIPS). Vancouver Canada: NIPS Press, 2010. 1162−1170

30 Asada M, Noda S, Tawaratsumida S, Hosoda K. Purposive behavior acquisition for a real robot by vision-based reinforcement learning. *Machine Learning*, 1996, **23**(2−3): 279−303

31 Kroemer O B, Detry R, Piater J, Peters J. Combining active learning and reactive control for robot grasping. *Robotics and Autonomous Systems*, 2010, **58**(9): 1105−1116

32 Gass S I, Fu M C. *Encyclopedia of Operations Research and Management Science*. Boston, MA: Springer, 2013. 326−333

33 Iruthayarajan M W, Baskar S. Covariance matrix adaptation evolution strategy based design of centralized PID controller. *Expert Systems with Applications*, 2010, **37**(8): 5775−5781

34 Endo G, Morimoto J, Matsubara T, Nakanishi J, Cheng G. Learning CPG-based biped locomotion with a policy gradient method: application to a humanoid robot. *The International Journal of Robotics Research*, 2008, **27**(2): 213−228

35 Peters J, Schaal S. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 2008, **21**(4): 682−697

36 Deisenroth M P, Rasmussen C E, Fox D. Learning to control a low-cost manipulator using data-efficient reinforcement learning. *Robotics: Science and Systems VII*. Cambridge: MIT Press, 2011. 57−64

37 Deisenroth M P, Rasmussen C E. PILCO: a model-based and data-efficient approach to policy search. In: Proceedings of the 28th International Conference on Machine Learning (ICML). Washington, USA: Omnipress, 2011. 465−472

38 Deisenroth M P, Neumann G, Peters J. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2013, **2**(1−2): 1−142

39 Zhao Dong-Bin, Shao Kun, Zhu Yuan-Heng, Li Dong, Chen Ya-Ran, Wang Hai-Tao, et al. Review of deep reinforcement learning and discussions on the development of computer Go. *Control Theory and Applications*, 2016, **33**(6): 701−717 (赵冬斌, 邵坤, 朱圆恒, 李栋, 陈亚冉, 王海涛, 等. 深度强化学习综述: 兼论计算机围棋的发展. 控制理论与应用, 2016, **33**(6): 701−717)

40 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemareet M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

41 Silver D, Huang A, Maddison C, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

42 Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature*, 2017, **550**(7587): 354−359

43 van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Arizona, USA: AAAI Press, 2016. 2094−2100

44 Wang Z Y, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling network architectures for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York City, USA: JMLR, 2016. 1995−2003

45 Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Texas, USA: AAAI Press, 2015

46 Zhang F Y, Leitner J, Milford M, Upcroft B, Corke P. Towards vision-based deep reinforcement learning for robotic motion control. arXiv: 1511.03791, 2015.

47 Zhang F Y, Leitner J, Milford M, Corke P. Modular deep Q networks for Sim-to-real transfer of visuo-motor policies. arXiv: 1610.06781, 2016.

48 Gu S X, Lillicrap T, Sutskever I, Levine S. Continuous deep Q-learning with model-based acceleration. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). New York City, USA: JMLR, 2016. 2829−2838

49 Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML). Beijing, China: JMLR, 2014. 387−395

50 Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France: JMLR, 2015. 1889−1897

51 Mnih V, Badia A P, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). New York City, USA: JMLR, 2016. 1928−1937

52 Levine S, Koltun V. Guided policy search. In: Proceedings of the 30th International Conference on Machine Learning (ICML). Atlanta, USA: JMLR, 2013. 1−9

53 Levine S, Koltun V. Learning complex neural network policies with trajectory optimization. In: Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing, China: JMLR, 2014. 829−837

54 Malekzadeh M, Queißer J, Steil J J. Imitation learning for a continuum trunk robot. In: Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). Bruges, Belgium: ESANN, 2017.

55 Ross S, Gordon G J, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: JMLR, 2011. 627−635

56 Ng A Y, Russell S J. Algorithms for inverse reinforcement learning. In: Proceedings of the 17th International Conference on Machine Learning (ICML). Stanford, USA: Morgan Kaufmann Publishers Inc., 2000. 663−670

57 Zhou Zhi-Hua. *Machine Learning*. Beijing: Tsinghua University Press, 2016.
(周志华. 机器学习. 北京: 清华大学出版社, 2016.)

58 Takeda T, Hirata Y, Kosuge K. Dance step estimation method based on HMM for dance partner robot. *IEEE Transactions on Industrial Electronics*, 2007, **54**(2): 699−706

59 Calinon S, Guenter F, Billard A. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 2007, **37**(2): 286−298

60 Calinon S, Billard A. Incremental learning of gestures by imitation in a humanoid robot. In: Proceedings of the 2nd ACM/IEEE International Conference on Human-robot Interaction. Arlington, VA, USA: IEEE, 2007. 255−262

61 Rahmatizadeh R, Abolghasemi P, Behal A, Bölöni L. From virtual demonstration to real-world manipulation using LSTM and MDN. arXiv: 1603.03833, 2016.

62 Calinon S, DHalluin F, Sauser E L, Caldwell D G, Billard A G. Learning and reproduction of gestures by imitation. *IEEE Robotics and Automation Magazine*, 2010, **17**(2): 44−54

63 Zhang T H, McCarthy Z, Jow O, Lee D, Chen X, Goldberg K, et al. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In: Proceedings of the 36th International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE, 2018.

64 Abbeel P, Ng A Y. Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st International Conference on Machine Learning (ICML). Alberta, Canada: ACM, 2004.

65 Ratliff N D, Bagnell J A, Zinkevich M A. Maximum margin planning. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). Pennsylvania, USA: ACM, 2006. 729−736

66 Ziebart B D, Maas A, Bagnell J A, Dey A K. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI). Illinois, USA: AAAI Press, 2008. 1433−1438

67 Levine S, Popovićí Z, Koltun V. Nonlinear inverse reinforcement learning with Gaussian processes. In: Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS). Granada, Spain: Curran Associates, 2011. 19−27

68 Ratliff N D, Bradley D M, Bagnell J A, Chestnutt J E. Boosting structured prediction for imitation learning. In: Proceedings of the 19th Advances in Neural Information Processing Systems (NIPS). British Columbia, Canada: Curran Associates, 2006. 1153−1160

69 Xia C, El Kamel A. Neural inverse reinforcement learning in autonomous navigation. *Robotics and Autonomous Systems*, 2016, **84**: 1−14

70 Wulfmeier M, Ondruska P, Posner I. Maximum entropy deep inverse reinforcement learning. arXiv: 1507.04888, 2015.

71 Finn C, Levine S, Abbeel P. Guided cost learning: deep inverse optimal control via policy optimization. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). New York City, USA: JMLR, 2016. 49−58

72 Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of the 30th Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: Curran Associates, 2016. 4565−4573

73 Peng X B, Abeel P, Levine S, van de Panne M. DeepMimic: example-guided deep reinforcement learning of physics-based character skills. arXiv: 1804.02717, 2018.

74 Zhu Y K, Wang Z Y, Merel J, Rusu A, Erez T, Cabi S, et al. Reinforcement and imitation learning for diverse visuomotor skills. arXiv: 1802.09564, 2018.

75 Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, et al. Deep Q-learning from demonstrations. In: Proceedings of the 32th Association for the Advancement of Artificial Intelligence (AAAI). Louisiana USA: AAAI Press, 2018.

76 Lemke C, Budka M, Gabrys B. Metalearning: a survey of trends and technologies. *Artificial Intelligence Review*, 2015, **44**(1): 117−130

77 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

78 Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T. Deep domain confusion: maximizing for domain invariance. arXiv: 1412.3474, 2014.

79 Shi Z Y, Siva P, Xiang T. Transfer learning by ranking for weakly supervised object annotation. arXiv: 1705.00873, 2017.

80 Gupta A, Devin C, Liu Y X, Abbeel P, Levine S. Learning invariant feature spaces to transfer skills with reinforcement learning. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: ICLR, 2017.

81 Stadie B C, Abbeel P, Sutskever I. Third-person imitation learning. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: ICLR, 2017.

82 Ammar H B, Eaton E, Ruvolo P, Taylor M E. Online multi-task learning for policy gradient methods. In: Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML). Beijing, China: JMLR, 2014. 1206−1214

83 Gupta A, Devin C, Liu Y X, Abbeel P, Levine S. Learning invariant feature spaces to transfer skills with reinforcement learning. arXiv: 1703.02949, 2017.

84 Tzeng E, Devin C, Hoffman J, Finn C, Peng X C, Levine S, et al. Towards adapting deep visuomotor representations from simulated to real environments. arXiv: 1511.07111, 2015.

85 Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Proceedings of the 30th Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: Curran Associates, 2016. 3630−3638

86 Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta-learning with memory-augmented neural networks. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). New York City, USA: JMLR, 2016. 1842−1850

87 Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: ICLR, 2017.

88 Edwards H, Storkey A. Towards a neural statistician. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: ICLR, 2017.

89 Rezende D, Mohamed S, Danihelka I, Gregor K, Wierstra D. One-shot generalization in deep generative models. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). New York City, USA: JMLR, 2016.

90 Duan Y, Schulman J, Chen X, Bartlett P L, Sutskever I, Abbeel P. RL^2: fast reinforcement learning via slow reinforcement learning. arXiv: 1611.02779, 2016.

91 Wang J X, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo J Z, Munoset R, et al. Learning to reinforcement learn. arXiv: 1611.05763, 2016.

92 Duan Y, Andrychowicz M, Stadie B C, Ho J, Schneider J, Sutskever I, et al. One-shot imitation learning. arXiv: 1703.07326, 2017.

93 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. arXiv: 1703.03400, 2017.

94 Xu D F, Nair S, Zhu Y K, Gao J L, Garg A, Li F F, et al. Neural task programming: learning to generalize across hierarchical tasks. arXiv: 1710.01813, 2017.

95 Reed S, de Freitas N. Neural programmer-interpreters. arXiv: 1511.06279, 2015.

96 Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world. In: Proceedings of the 30th International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017. 23−30

刘乃军 中国科学院自动化研究所博士研究生. 2016 年获得山东大学硕士学位. 主要研究方向为智能机器人, 深度强化学习. E-mail: liunaijun2016@ia.ac.cn
(**LIU Nai-Jun** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. He received his master degree from Shandong University in 2016. His research interest covers intelligent robot and deep reinforcement learning.)

鲁 涛 中国科学院自动化研究所复杂系统管理与控制国家重点实验室副研究员. 2007 年获得中国科学院自动化研究所博士学位. 主要研究方向为人机交互、机器人以及人工智能.
E-mail: tao.lu@ia.ac.cn
(**LU Tao** Associate professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2007. His research interest covers human-robot interaction, robotics, and artificial intelligence.)

蔡莹皓 中国科学院自动化研究所副研究员. 2009 年获得中科院自动化所博士学位, 曾任美国南加州大学博士后研究员和芬兰奥卢大学研究科学家. 主要研究方向为机器人视觉.
E-mail: yinghao.cai@ia.ac.cn
(**CAI Ying-Hao** Associate professor at the Institute of Automation, Chinese Academy of Sciences. She received her Ph. D. degree from Institute of Automation, Chinese Academy of Sciences in 2009. She was a postdoctoral research fellow at University of Southern California, USA and senior research scientist in University of Oulu, Finland. Her research interest covers computer vision in robotics.)

王 硕 中国科学院自动化研究所复杂系统管理与控制国家重点实验室和中国科学院脑科学与智能技术卓越创新中心研究员. 主要研究方向为智能机器人, 仿生机器人和多机器人系统. 本文通信作者. E-mail: shuo.wang@ia.ac.cn
(**WANG Shuo** Professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation of the Chinese Academy of Sciences and Center for Excellence in Brain Science and Intelligence Technology of the Chinese Academy of Sciences. His research interest covers intelligent robot, biomimetic robot and multi-robot system. Corresponding author of this paper.)