基于触觉传感器和强化学习内在奖励的机械臂抓取方法

宋相兵[1]，季玉龙[2]，俎文强[3]，何　扬[1]，杨红雨[1,3]
(1. 四川大学视觉合成图形图像技术国防重点学科实验室，成都 610065；
2. 四川大学空天科学与工程学院，成都 610065；
3. 四川大学计算机学院，成都 610065)

摘　要：触觉在机器人抓取过程中扮演着重要的角色，但在大多数强化学习任务中，触觉仅被用于拓展状态空间，其提供的位置和压力等信息很少被完全利用. 针对该问题，同时受内在奖励机制启发，首先设计了一种“倒 T”形传感器阵列布局；然后基于这种传感器阵列提出了新的内在激励方法，该方法根据机械臂末端与物体接触位置的不同，给予不同的重视程度，鼓励智能体以更有效的姿态来夹取物体；最后将该方法在仿真环境中进行测试，结果表明该方法在夹取椭球和圆球物体任务中收敛速度比最新的基准方法平均提高了约 20%.

关键词：深度强化学习；机械臂；抓取；触觉；内在奖励
中图分类号：TP242.6　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2022.032003

The method for manipulator grasping based on tactile sensor and reinforcement learning intrinsic reward

SONG Xiang-Bing[1], *JI Yu-Long*[2], *ZU Wen-Qiang*[3], *HE Yang*[1], *YANG Hong-Yu*[1,3]
(1. National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China;
2. School of Aeronautics and Astronautics, Sichuan University, Chengdu 610065, China;
3. College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Although play an important role in the process of the robot grasping, haptics is only used to extend the state space, and the information provided by it, such as position and pressure, is rarely fully utilized in most reinforcement tasks. Inspired by the intrinsic reward mechanism, an intrinsic incentive method based on the “inverted T” array sensor is proposed. According to the position where the end effector of the robot touches the object, the method gives degrees of importance, and encourages the agent to achieve the goal with a more effective posture. Finally, the method was tested in the simulation environment, and the results showed that the speed of convergence of the method in the task gripping ellipsoid objects was about 20% faster than the latest benchmark method.

Keywords: Deep reinforcement learning; Robot arm; Grasping; Haptics; Intrinsic reward

收稿日期：2021-11-03
基金项目：国家自然科学基金重点项目(U20A20161)
作者简介：宋相兵(1995－)，男，硕士研究生，四川乐山人，主要研究方向为强化学习、智能机器人.
通讯作者：杨红雨. E-mail：yanghongyu@scu.edu.cn

1 引　言

人们对周遭环境和物体进行探索时,非常依赖触觉,并依靠触觉感知来引导下一步操作. 比如,人类可以在没有视觉的情况下灵活操纵某一物体,但是在失去触觉后,人们操纵物体的灵活性将大大降低[1,2]. 同样机器人在密集接触的任务中也非常依赖触觉提供的局部信息(如压力、震动、接触印记[3]等),因为这些信息在机器人进行灵巧操作[4]或抓握动作[5]时起着至关重要的作用.

与视觉传感器相比,触觉传感器提供的接触信息能让智能体更充分地感知环境和获取交互情况,比如,感知到物体发生滑动,或者已经牢牢抓住物体[6]. 通常在视觉被遮挡或者识别不准确的情况下,智能体将无法准确感知环境,因而无法完成任务,而 Wu 等[7]发现智能体仅依靠触觉信息仍可以通过尝试与探索完成抓取任务. 因此触觉传感器在智能体中的应用逐渐成为研究热点. 目前基于视觉的触觉传感器因其能获得高分辨率的反馈信息在机器人领域很受欢迎,并已经产生了很多应用和设计. 比如 Dong 等研发的 GelSight[3]、GelSlim 2.0[8]、GelSlim MPalm[9]、GelSlim 3.0[10],Lambeta 等[11]设计的 DIGIT,Padmanabha 等[12]利用多个微型摄像头制作的 OmniTact 传感器,以及 She 等[13]在柔性手指中嵌入的触觉传感器等. 另外 TacTip[14]虽然也是通过摄像头获取触觉信号,但是其触觉印记的分辨率远低于上述几种传感器. 除此之外,阵列式传感器在机器人领域也很受青睐. 它们大都利用新材料将外部刺激或压力转化为电信号,并且感知单元的布局相对规则. 这类传感器根据感知原理主要分为:压阻式传感器[15]、电容式传感器[16]和气压传感器[4]等. 不论信息的获取方式是怎样,每一个新的传感器都有其固有的属性,如脆性、体积、分辨率、延迟和生产成本等[17].

虽然基于视觉的触觉传感器能获取高分辨率的触觉信息,但由于其数据大都是图像格式,若想获取有效信息还需要经过复杂处理,而基于阵列的触觉传感器不仅获取触觉信息效率较高,而且还易于进行模拟仿真,这使得我们可以方便地根据触觉信号给予智能体相应的奖励. 因此考虑到这些特性和其不同的应用场景,本文将仿真阵列传感器进行实验,具体传感器仿真细节在 3.1 节中详细阐述.

在机器人抓取方面,触觉信息大多数情况下都作为深度神经网络的输入用于机器人训练. 比如 Hogan 等[9]、Calandra 等[5]和 Hellman 等[18]先利用机器学习训练了一个预测模型,然后让机器人根据其结果来判断状态后再执行相应任务. 这类方法计算复杂,会拖慢训练速率. 而如今深度强化学习已在多个领域表现出巨大的潜力,如电子游戏、仿真模拟、机器人控制[19]以及图像处理[20]等. 且 Dong 等[21]试验得出强化学习在机器人抓取等复杂任务中的表现要优于监督学习. 另外 Chebotar 等[22]将深度学习与强化学习结合,利用触觉的时空特征训练了一个抓握稳定判别器,让智能体学会以更优姿态重新抓取物体.

而另一些研究人员则仅将触觉信号当成是强化学习环境的状态观测值,让智能体通过试错与奖励反馈学会抓握物体,比如文献[6]让机器人学会自适应控制力的大小避免物体掉落、文献[23]研究表明加入触觉反馈可以显著提高多指机器人抓握物体的鲁棒性,同样的方法也被用于其他非抓取的任务,如文献[24]让智能体根据触觉反馈学会控制方块;文献[4]将气压传感作为触觉让机械手指学会旋转和移动物体. 这种方法虽然能提高智能体学习效率,但其本质是增加了环境的状态维度,提高了采样机器人的采样效率,而没有真正利用有价值的触觉信息.

与以上工作不同的是,另一些学者则根据触觉信息制定奖惩函数来诱导智能体尝试和探索,比如 Huang 等[25]根据触觉反馈制定奖惩机制,让机械手学会温柔的触摸和操纵物体;文献[26]将触觉信号作为内在奖励,鼓励智能体进行更有效的探索. 本文的工作也受此启发,首先基于阵列式传感器设计了一种“倒 T”字形的触觉传感器的排列方式,然后通过分析机械末端与物体接触时的受力情况,提出了新的奖励函数来鼓励智能体能以更合适、更稳定的姿态去抓取物体,最后将该方法适配到新的机器人仿真环境中,验证该方法的有效性.

2 问题描述和模型建立

2.1 问题描述

机器人抓取是指将物体从起始位置拾取到另一目标位置的一套连续动作. 在强化学习的训练过程中,智能体需要不断试错,不停地用末端接触物体,尝试不同位置将其夹起. 当物体表面比较粗糙,末端和物体之间摩擦力比较大时,机器人能非常轻松完成抓取任务,而当物体比较光滑且表面是弧面时,机器人夹取物体的成功率则会大大降低.

机器人用刚性末端夹取光滑物体类似于我们生活中用筷子夹取某些豆类食物或者夹弹珠，其成功率主要取决于接触力的大小和接触点位置. 如果接触力太小，接触点的摩擦力小于重力，物体自然会因此掉落；若接触点处于错误位置，增加接触力只会增大物体滑落的可能性. 以夹取球型物体为例，如图 1a 所示，当夹具与物体接触点位于物体形心上方时，接触力方向偏离形心，如果此时夹具与物体间的摩擦系数太小，将不能保证有效夹取. 图 1b 和 1c 所示的方式则可以进行有效抓取. 但是图 1b 中接触点和接触力虽然处于被抓取物体的最佳位置，但该接触点位于夹具边缘，容错率很低，若机器稍有震动将导致物体滑落. 因此该夹取方式仍不是最优解. 而图 1c 所示的夹取情形中，物体和夹具上的接触点都到达最佳位置，即使在夹取过程中物体发生轻微偏移，也可以将物体成功抓取. 图 1d 展示了俯视时的最佳抓取位置.

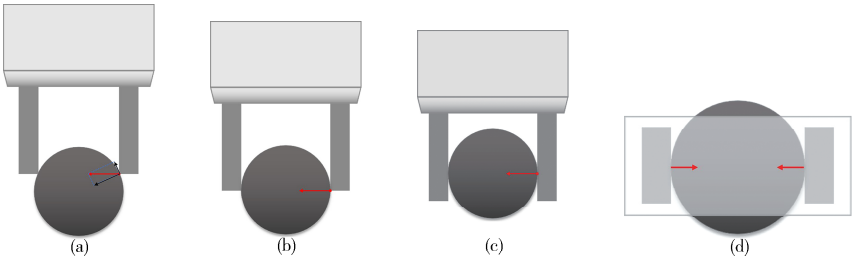

图 1　夹具与物体接触位置：(a) 最差接触点；(b) 次优接触点；(c) 最优接触点；(d) 俯视情况下的最优接触

Fig. 1　The contact position of thegripper and the object: (a) The worst contact point; (b) the suboptimal contact point; (c) the optimal contact point; (d) the optimal contact in the overhead case

本文认为图 1c 和 1d 是机械臂夹取过程中非常理想的中间状态，因此我们希望塑造一个合理的奖励函数来引导智能体到达这种中间状态. 这种中间状态可利用触觉传感器来间接表示，但是在强化学习中，奖励塑造是一个很棘手的问题. 因为一项复杂任务往往具有多个中间状态，如果我们针对每一个状态都进行奖励，那样奖励函数将会非常复杂，而且这样往往也会使智能体找到得分漏洞，然后陷入刷分的循环，从而导致训练失败[27]. 而如果仅将达到最终目标作为奖励体条件，会使得智能体进行过多随机且无用的探索，从而拖慢训练速度甚至出现算法一直无法收敛的情况.

所以我们将用触觉信号作为内在奖励引导智能体去探索，使其更快速地到达中间状态(即末端以更优位置去接触物体)，但同时又允许智能体能够进行其他不同的探索，也就是利用现有的机械臂仿真环境进行实验，通过增加触觉传感区域来扩展强化学习环境的状态空间，并根据触觉累计值修改奖励函数，从而引导智能体进行更有效的探索.

2.2　强化学习模型

本文通过深度强化学习和内在触觉激励来优化机械臂抓取物体的任务，该任务是将目标物体夹取到一个目标位置，从而获得任务奖励. 其中机械臂抓取物体的过程可归纳为一个马尔可夫决策过程(Markov Decision Process, MDP). 该过程可以用元组(S,A,P,R,γ)来表示. 其中 S 表示状态空间；A 表示动作；而 P 表示在状态 S 下执行动作 A 后，状态变成 S'的概率，可写作$(P:S\times A\rightarrow S')$；$R$ 表示智能体在状态 S 下执行动作 A 所获得的奖励函数，即 $R:S\times A\rightarrow[0,1]$；$\gamma$ 代表折扣因子，它表示未来奖励对现在的重要性，其值越大代表智能体越看重未来奖励，其值越小则表示智能体更重视短期回报.

在整个交互过程中，机械臂作为智能体，在 t 时刻，观察当前状态，然后根据策略 π 选择动作 a_t，$a_t\in A$ (其动作由一个四维向量表示，前三维数据表示一个与机械臂夹具中心绑定的 Mocap 动作捕捉点的世界坐标，后一维数据表示夹具开合的状态)，然后观察得到新的状态 s_{t+1}，最后根据状态计算上一步动作的奖励，再进行接下来的决策. 智能体的最终目标就是找到最优策略 π^*，得到最大化累计奖励 R_t，即：

$$R_t=\sum_{i=t}^{\infty}\gamma^{i-t}r_i \tag{1}$$

其中 r_i 代表即时奖励.

3　方　法

3.1　触觉传感器的设计与任务设置

与文献[26]类似，我们在机械臂末端夹具上添加了传感区域，本文和他们不同的是，他们将整个

夹具用一个传感区域覆盖,对所有接触位置一视同仁,而我们认为夹具的区域是有优劣之分的,正如 2.1 节所讲到的,物体与夹具的接触位置不同,会影响夹取的稳定性,因此我们根据前面定义的接触位置的好坏给予了不同传感器不同的重视程度(重视程度与位置关系如图 2 所示),另外为了既鼓励夹具与物体接触,又能引导智能体以最优位置夹取物体,本文在夹具上设计了如图 2 所示的倒"T 型"传感区域阵列. 该阵列由 3 个传感区域组成,最下面的一整块传感区域是为了能更全面地捕捉刚接触的信号,上面的两块方形传感区域是为了引导智能体用这些区域(尤其是区域 1)去夹取物体. 每当末端夹具与物体在某一传感区域接触时,物理引擎就会计算出相应的值. 通过奖励函数设置,我们将鼓励智能体用末端去接触目标物体,并鼓励它尽量用重视程度最高的位置去接触物体.

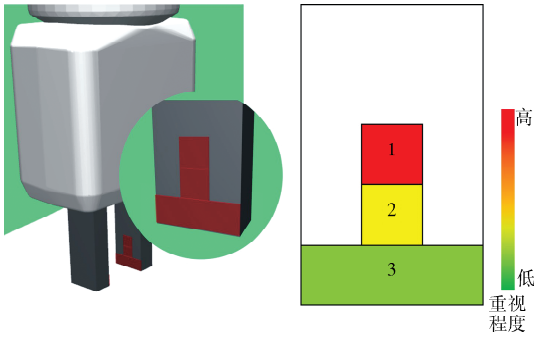

图 2　传感器布局与相应区域重视程度

Fig. 2　Sensor layout and the importance of the corresponding area

接着为了证明本文方法的有效性,我们在提高原始任务难度的情况下(即增加了目标位置在空中生成的概率和高度,以及将目标物体由方块替换成圆球和椭球),使用不添加传感器、只加传感器、最新内在奖励[26]以及本文内在奖励 4 种方法分别进行球形抓取和椭球抓取的任务训练,通过对比它们各自夹取的成功率来佐证本文观点.

3.2　奖励函数设定

强化学习算法旨在让智能体学习一种能在环境中获得最大长期回报的策略,而在大多数任务中这些奖励都是稀疏的,即智能体只有完成了目标才能获得相应奖励. 这就会让智能体产生太多无意义的尝试,从而降低学习效率. 因此,为了提高智能体探索的效率,本文引入了内部激励机制[28,29]来鼓励智能体进行更有效的探索. 我们的奖励函数由外部奖励 $r_{\text{ext}}(s,g)$ 和与目标无关的内部奖励 $r_{\text{int}}(s)$ 两部分组成,其表示如下.

$$r(s,\ g)=\omega_{\text{ext}}*r_{\text{ext}}(s,\ g)+\omega_{\text{int}}*r_{\text{int}}(\text{s}) \tag{2}$$

其中 ω_{ext} 和 ω_{int} 分别表示外部奖励和内部奖励的权重. 外部奖励是完成目标后获得的稀疏奖励,只要物体位置在目标范围之内即返回 1,否则返回 0,具体表示如下.

$$r_{\text{ext}}(s,g)=\begin{cases}1,\ \text{if}\ \|g-x_{\text{obj}}\|<\varepsilon_{\text{pos}}\\0,\ \text{else}\end{cases} \tag{3}$$

其中 g 表示目标位置;x_{obj} 表示物体位置;ε_{pos} 表示距离阈值. 而本文内部奖励又分为两部分,表示如下.

$$r_{\text{int}}(s_t)=\omega_{c\_f}r_{c\_f}+\omega_{c\_p}r_{c\_p} \tag{4}$$

其中 c_f、c_p 分别表示接触力和接触位置. 接触力奖励仅根据触觉信号来设定,只要一幕训练过程中所有触觉信号累计值 $\sum v$ 超过阈值 $\varepsilon_{\text{touch}}$ 就返回 1,否则返回. 其表示如下.

$$r_{c\_f}(s_t)=\begin{cases}1,\ \text{if}\quad \sum_{i=0}^{t}v_i>\varepsilon_{\text{touch}}\\0,\ \text{else}\end{cases} \tag{5}$$

而对于接触位置奖励,我们按照图 2 传感区域的重要程度,根据接触信号累计值的区域的不同给予不同奖励,其表示如下.

$$r_{c\_p}(s_t)=\begin{cases}1, & \text{if}\quad \sum_{i=0}^{t}v_{1i}>\varepsilon'_{\text{touch}}\\0.8, & \text{if}\quad \sum_{i=0}^{t}v_{2i}>\varepsilon'_{\text{touch}}\\0.4, & \text{if}\quad \sum_{i=0}^{t}v_{3i}>\varepsilon'_{\text{touch}}\\0, & \text{else}\end{cases} \tag{6}$$

其中,v_{1i}、v_{2i}、v_{3i} 分别表示区域 1、2、3 的触觉信号值.

为了能获得最佳的训练效果,本文需确定 ω_{ext}、ω_{int}、ω_{c_f}、ω_{c_p}、$\varepsilon_{\text{touch}}$、$\varepsilon'_{\text{touch}}$ 的最优值. 因此本文采用控制变量法,在第 4 章介绍的实验环境下进行了 4 组对比实验(因为 $\omega_{\text{ext}}+\omega_{\text{int}}=1$,$\omega_{c\_f}+\omega_{c\_p}=1$,所以这 4 个参数只需通过两组实验即可确定). 每组实验只改变其中一组参数值,固定其他所有条件,然后进行 150 回合的训练,最终以测试成功率(该成功率计算方式将在 4.2 节中阐述)作为训练效果的判断指标.

4 组实验结果如图 3 所示,4 幅图中的黄色曲线分别表示所选参数在各自最优值下的训练效果. 最终本文确定以表 1 中的值作为奖励函数的参数.

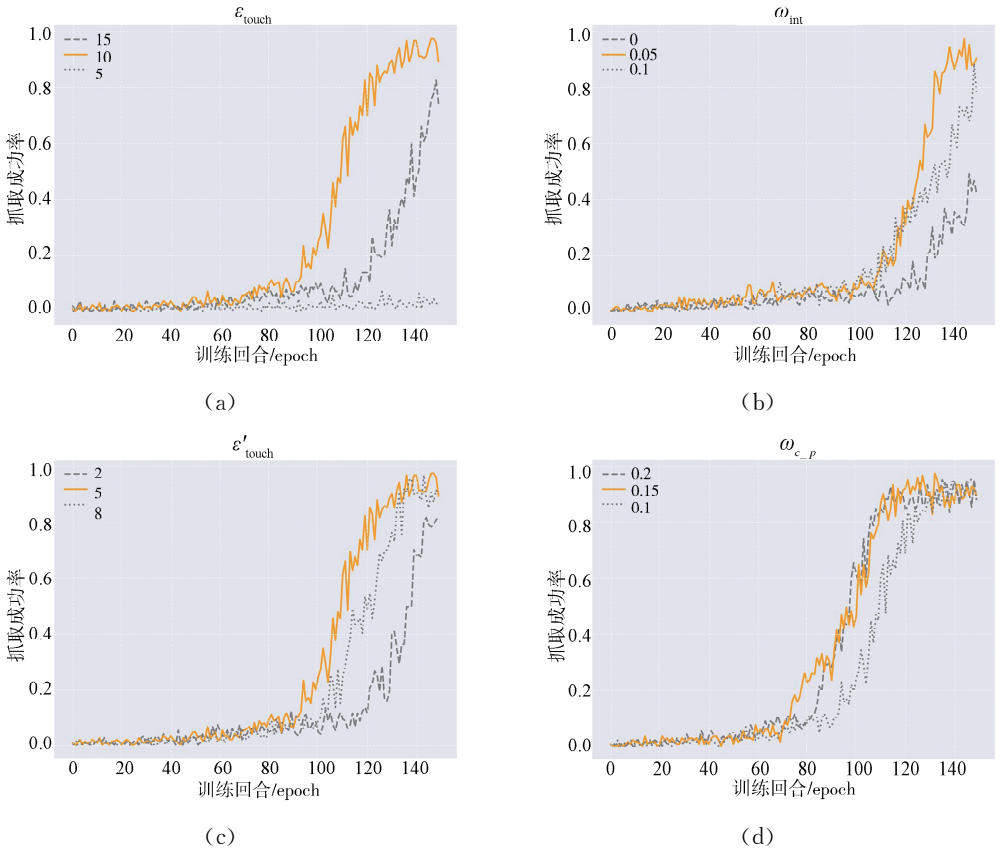

图 3　对比实验结果：(a) ε_{touch} 的选取范围为 5～15；(b) ω_{int} 的选取范围为 0～0.1；(c) ε'_{touch} 的选取范围为 2～8；(d) ω_{c_p} 的选取范围为 0.1～0.2

Fig. 3　The results of comparative experiments: (a) the value range of ε_{touch} is 5～15; (b) the value range of ω_{int} is 0～0.1; (c) the value range of ε'_{touch} is 2～8; (d) the value range of ω_{c_p} is 0.1～0.2

表 1　奖励函数中的参数值

Tab. 1　The value of the parameter in the reward function

参数名	参数值
ω_{ext}	0.95
ω_{int}	0.05
ω_{c_f}	0.85
ω_{c_p}	0.15
ε_{touch}	10
ε'_{touch}	5

3.3　训练算法

由于机器人抓取是一个连续动作控制问题，所以本文采用集成了时间差分学习和策略梯度的深度确定性策略梯度算法(Deep Deterministic Policy Gradient，DDPG)[30]来训练智能体，并且将之与事后经验回放技术(Hindsight Experience Replay，HER)[31]结合来提高数据利用效率.

其中，DDPG 是一种基于演员-评论家(Actor-Critic)框架的离线策略(Off-policy)算法，即使用两套不同的网络进行动作的选择和评价. 其中 Actor 是直接学习策略的网络，它接收从环境中获取的当前状态 S，然后根据策略 π，输出相应的动作 A. 而 Critic 网络则根据当前状态和动作计算 Q 值来评价动作的好坏，即学习动作价值函数 Q^{π}. 在训练期间，Actor 网络通过行为策略去探索环境，该行为策略是目标策略加上了一些噪声后的策略，即 $\pi_b=\pi(s)+N(0,1)$. 在式(7)中，Critic 通过最小化 $r_t+\gamma Q(s_{t+1},\pi(s_{t+1}))$ 和 $Q(s_t,a_t)$ 的 loss L_c 来更新网络.

$$L_c=r_t+\gamma Q(s_{t+1},\pi(s_{t+1}))-Q(s_t,a_t) \tag{7}$$

其中，$r_t+\gamma Q(s_{t+1},\pi(s_{t+1}))$ 表示真实的动作状态值；$Q(s_t,a_t)$ 表示估计的动作状态值. 而 Actor 使用策略梯度通过损失函数的梯度下降来训练网络，该损失函数表示为

$$L_a=-E_s[Q(s,\pi(s))] \tag{8}$$

其中，s 是从经验回放池采样而来. 而 L_a 关于 Actor 网络的参数的梯度可以通过结合了 Critic 和 Actor 网络的反向传播计算得到.

HER[31]是 Andrychwicz 等提出的一种数据增强技术. 在机器人任务中，如果目标比较复杂而且

奖励很稀疏,那么智能体在学到一些经验前会进行很多失败且无效的尝试. 因此 HER 就鼓励智能体从失败的经验中学习一些东西. 在探索过程中,智能体根据真实目标对一些轨迹进行采样. HER 的主要思想就是将选定的一次状态转移中的原始目标替换为已经达到的目标,即用一个虚拟目标替换真实目标. 这样智能体就可以获得足够数量的奖励信号来开始学习.

HER 包含 4 种采样策略,每种策略具体内容如下:(1) future 模式:当回放某一状态转移时,从同一幕的该状态之后,随机选择 k 个状态进行回放,即,如果现在的样本为(s_t,a_t,s_{t+1}),就从 $t+1$ 开始到最后的状态中选择 k 个已经达到的目标(achieved goal)作为新目标;(2) final 模式:把每一幕的最后一个已达到目标作为新目标;(3) episode 模式:与 future 模式有些类似,但是该模式直接从同一幕中随机选择 k 个已达到目标作为新目标,没有限制是否要往后采样;(4) random 模式:随机选择整个训练过程中的 k 个状态进行回放;

本文采用的是 future 模式,HER 计算流程图如图 4 所示,其主要步骤如下.

(1) 随机选取一幕训练的完整样本;

(2) 用智能体当前的已完成状态替换掉最终目标(desired goal);

(3) 更新“info”信息;

(4) 重新计算奖励.

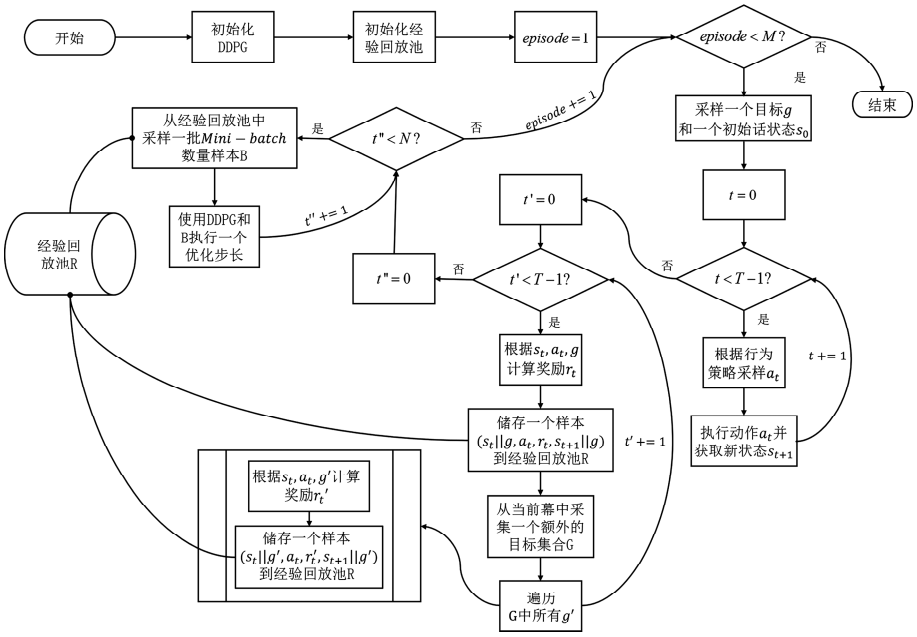

图 4　HER 计算流程图

Fig. 4　The calculation flow chart of HER

4　仿真实验环境及结果

4.1　实验环境

本实验环境是在 OpenAI Gym[32] 中的 Fetch 机器人抓取与放置(Fetch-PickAndPlace)仿真实验环境的基础上进行的改进,其整个实验场景如图 5 所示. 该实验环境是以 MuJoCo[33] 作为物理引擎. 在机器人运动仿真过程中,MuJoCo 具备关节防卡死、多约束、多驱动以及细节化仿真等特点,十分适用于机器人姿态控制及机械臂运动仿真. MuJoCo 可以仿真许多类型的传感器,比如触摸传感器、惯性测量单元、力传感器、力矩传感器、关节速度传感器等. 这些传感器在仿真环境里不参与模型的碰撞计算,它们只为用户的计算提供相关信息数据信息. 例如触摸传感器是在执行器上定义一块特定形状的传感区域,只要执行器在该区域与其他物体产生接触行为,相应的接触力就会被计算出来. 该传感器的读数是非负标量,它是通过将被包含在接触区域的所有法向力(标量)相加来计算的.

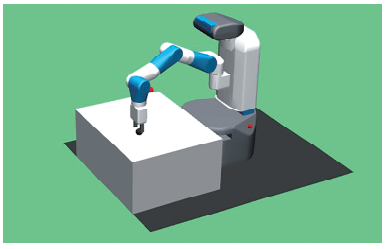

图 5　实验场景

Fig. 5　The experimental scene

该实验任务就是将桌上的目标物体抓取并移动到空中红色的目标位置，只要物体与目标的距离不超过阈值即可判定完成任务.

而本文对环境的主要修改如图 2 所示. 这 3 块传感区域的宽度都相等，而第 3 块区域的长度是前两块的 3 倍，且与末端夹具等宽. 同时，为了避免智能体在探索过程中产生误触(如夹具触碰桌面而产生读数)，本文将 3 块传感区域沿末端长边依次上移了 2.5 mm. 然后将这 3 块区域作为 site 被绑定在末端夹具 body 实体下，其在 xml 文件中的表现形式如下：

<body (body 参数)……>
　　(其他属性)……
　　<site (site 参数)……/>
　　(其他 site)……
</body>

最终将环境的观察值增加了 2 * 3 维，即从 25 维增加到 31 维. 最终改进后的环境状态信息如表 2 所示.

表 2　仿真环境状态值信息

Tab. 2　The information of state value in simulation environment

状态值名称	含义	维度
object_rot	目标物体的朝向	3
object_pos	目标物体的坐标位置	3
object_rel_pos	物体与末端夹具的相对位置	3
object_velp	目标物体的移动速度	3
object_velr	目标物体的旋转速度	3
grip_pos	末端夹具的坐标位置	3
grip_velp	末端夹具的移动速度	3
gripper_vel	夹具之间的相对移动速度	2
gripper_state	夹具的开合状态	2
touch_values(本文中添加)	触觉传感器读数	6
合计		31

4.2　实验设置

本实验中的所有超参数以及训练过程都在文献[31,34]中有详细描述，并且这些超参数和强化学习算法都被集成于 OpenAI 的 Baselines[35] 中. Baselines 是基于 TensorFlow 而开发的一套强化学习算法的实现框架. 本文利用其中的 DDPG 和 HER 来训练智能体.

本文设置了 4 组实验环境，每组环境完成 2 项不同物体(球体和椭球体)的夹取任务. 其中，1 组为本文的实验环境，其观察值如 4.1 节所述是 31 维. 另外 3 组为对比实验环境，而对比实验环境中 2 组的观察值也是 31 维，另一空白组的观察值只有初始的 25 维，每组详细设置如表 3 所示.

表 3　实验环境信息

Tab. 3　The information of experimental environment

环境类型	观察值维度	奖励类型	简称
本文传感器＋内在奖励	31	本文内在奖励	本文环境
文献[26]传感器＋内在奖励	31	文献[26]内在奖励	情况 1
只加传感器	31	原始奖励	情况 2
不加传感器	25	原始奖励	情况 3

然后基于以上 4 种环境在 Windows 平台 Intel 16 核电脑上用 15 核通过 MPI 实现 150 回合的并行训练，最终以每回合训练结束后测试抓取的成功率作为主要判断指标，以第一次达到基准成功率的回合数作为辅助判断指标. 抓取成功率就是在每回合训练结束后，再在每个核心上进行 10 次确定性的测试试验，接着综合计算所有核心上试验夹取成功次数而得到的成功率.

4.3　结果及分析

本实验通过夹取球体(图 6)和椭球体(图 7)任务对比了 3.1 节中提到的 4 种情况(即未加传感器的原始情况、只加传感器不修改奖励函数、最新提出的传感器结合内在奖励机制以及本文的传感器结合内在奖励机制，后文中这些情况简称如表 3 最后一列所示)，所得结果如图 8、图 9、表 4 以及表 5 所示. 其中图 8、图 9 横坐标表示训练回合，纵坐标表示每回合得出的抓取成功率. 表 4 前 4 列给出了 4 种情况下 130～150 回合的平均成功率，后两列计算了本文平均成功率与情况 1 和 2 平均成功率的比率(由于在 150 回合内情况 3 的成功率未见显著提升，因此本文未计算该情况下的比率). 表 5 以表 4 中情况 1 的平均成功率(由于情况 2 在 150 回合内未见收敛的趋势，故不予考虑)作为基准成功率，列出了情况 1 与本文环境第一次到达该成功率的回合数以及回合比率. 比如，情况 1 的圆球抓取任务中第一次达到 0.861 成功率的回合是第 133 回合，而本文环境第一次达到该成功率则是在 112 回合，收敛速度是前者的 1.188 倍.

图 6　实验中的球体

Fig. 6　The sphere of the experiment

图 7　实验中的椭球体

Fig. 7　The ellipsoid of the experiment

图 8 和图 9 中蓝色实线是本文所使用的方法得出的抓取成功率曲线，橘黄色虚线是 Vulin 等[26]提出的传感器结合内在奖励方法(即情况 1)得出的结果，红色和绿色虚线分别表示只加传感器不修改奖励(即情况 2，仅拓展智能体的观测空间)和没加传感器(即情况 3)情况下的训练情况. 可以看出，在球体和椭球体抓取任务中，加入传感器增加了观测空间后，智能体的学习效率相较于不加传感器的情况有较大改善. 而橘黄色虚线则表明情况 1 的确能大大提高智能体的学习效率. 不过，结合表 4 和表 5 可知本文所提方法同样远远优于未经优化的传统方法(情况 2 和情况 3)，而且在同样条件下，比最新的情况 1 提前约 20 回合收敛，并且在后者收敛之前就达到了最优的正确率. 这是由于情况 1 没对接触位置作区分，仅仅是鼓励智能体去接触目标物体，而忽略了接触位置的差异，从而导致智能体在训练过程中仍会多次尝试以 2.1 节中描述的次优或最差姿态去接触物体，最终导致夹取失败而拖慢训练速度.

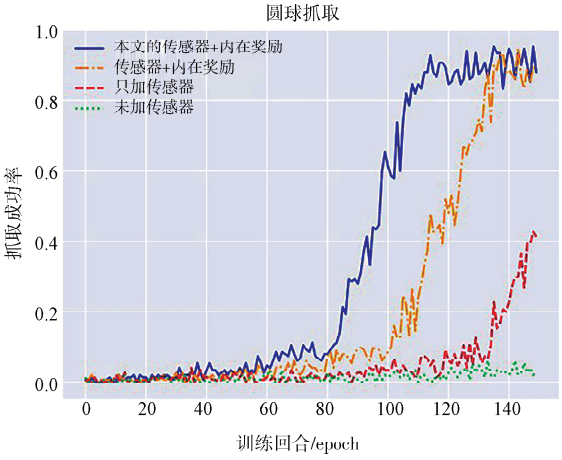

图 8　圆球抓取任务结果

Fig. 8　The result of grasping sphere

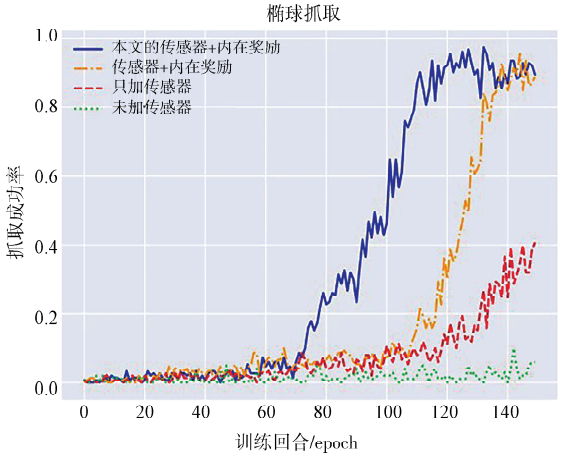

图 9　椭球抓取任务结果

Fig. 9　The result of grasping ellipsoid

表 4　130～150 回合平均成功率

Tab. 4　Average success rate in the range of 130～150 epochs

任务	情况 3	情况 2	情况 1	本文环境	相对于情况 2	相对于情况 1
圆球抓取	0.032	0.235	0.861	0.905	3.851	1.051
椭球抓取	0.031	0.296	0.845	0.902	3.047	1.067
平均					3.449	1.059

表 5　第一次到达基准平均成功率的回合数

Tab. 5　First epoch when a curve reaches the baseline success rate

任务	情况 1	本文环境	相对于情况 1
圆球抓取	133	112	1.188
椭球抓取	137	110	1.245
平均	/	/	1.217

因此，本文所提方法极大地提高了智能体抓取球形物体时的学习效率. 在本实验中，我们利用了传感器提供的位置和压力信息，有效地捕捉到了物体的接触信息，然后通过内在奖励机制鼓励智能体进行有效的探索. 在这里内在奖励就像一个指导员，引导智能体达到一个容易达到且有意义的状态. 有了内在奖励引导，智能体能够返回一个比较合适的状态，并且不会因为随机探索而丢失该轨迹，因此内在奖励机制可以提高智能体有意义的探索.

5　结　论

针对智能体在对球形物体进行抓取时容易产生滑动的问题，本文提出了一种新的呈“倒 T”形排布传感器阵列和相应的内在奖励函数；并且结合 DDPG+HER 强化学习算法在 MuJoCo 仿真环境中对比了 4 种方法；最后验证了以触觉传感器累计值和接触位置作为内在奖励信号能够引导智能体更快地完成球体的夹取任务，同时又不会限制智能

体探索其他状态.

当然，我们可以设计一个完美的末端夹具来执行相关任务.但在现实世界，物体的形状有成千上万种，要设计这样一种末端机械装置十分困难，因此目前的大多数末端执行器能达到的都只是在特定情况下的最优控制[36].人类的手之所以能抓握大部分物体，除了因为其灵活的关节外，还因为其具有敏锐的触觉.所以在现有末端装置条件下加入触觉传感是提高物体抓取的有效途径.

未来，我们将继续开展相关工作，如将该方法应用于更多类型的物体，包括但不限于圆柱体、正方体以及不规则物体等；并且将该方法应用于不同类型的末端执行器上来验证该方法的鲁棒性.此外，本文是仅在仿真环境下进行的相关实验，下一步我们会将算法迁移到真实机器人上，在复杂真实的环境中验证其效果.

参考文献：

[1] Jones L A, Lederman S J. Human hand function [M]. USA: Oxford University Press, 2006.

[2] Kosoy E, Collins J, Chan D M, *et al*. Exploring exploration: Comparing children with RL agents in unified environments [EB/OL]. [2021-05-22]. https://doi.org/10.48550/arXiv.2005.02880.

[3] Dong S, Yuan W, Adelson E H. Improved gelsight tactile sensor for measuring geometry and slip [C]// Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017.

[4] Van Hoof H, Hermans T, Neumann G, *et al*. Learning robot in-hand manipulation with tactile features [C]//Proceedings of the 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). Seoul, South Korea: IEEE, 2015.

[5] Calandra R, Owens A, Jayaraman D, *et al*. More than a feeling: Learning to grasp and regrasp using vision and touch [J]. IEEE Robot Autom Lett, 2018, 3: 3300.

[6] Koh K H, Farhan M, Liu Y F, *et al*. Learning to grasp unknown objects using force feedback [C]// Proceedings of the 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). Honolulu, US: IEEE, 2017.

[7] Wu B, Akinola I, Varley J, *et al*. MAT: multi-fingered adaptive tactile grasping via deep reinforcement learning [C]//Proceedings of the Conference on Robot Learning(CoRL). [S.l.]: PMLR, 2020.

[8] Ma D, Donlon E, Dong S, *et al*. Dense tactile force estimation using GelSlim and inverse FEM [C]// Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019.

[9] Hogan F R, Ballester J, Dong S, *et al*. Tactile dexterity: manipulation primitives with tactile feedback [C]//Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA). [S.l.]: IEEE, 2020.

[10] Taylor I, Dong S, Rodriguez A. GelSlim3.0: high-resolution measurement of shape, force and slip in a compact tactile-sensing finger [EB/OL]. (2021-03-23)[2021-05-22]. https://doi.org/10.48550/arXiv.2103.12269.

[11] Lambeta M, Chou P W, Tian S, *et al*. Digit: a novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation [J]. IEEE Robot Autom Lett, 2020, 5: 3838.

[12] Padmanabha A, Ebert F, Tian S, *et al*. Omnitact: A multi-directional high-resolution touch sensor [C]//Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2020.

[13] She Y, Liu S Q, Yu P, *et al*. Exoskeleton-covered soft finger with vision-based proprioception and tactile sensing [C]//Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2020.

[14] Ward-Cherrier B, Pestell N, Cramphorn L, *et al*. The TacTip family: soft optical tactile sensors with 3d-printed biomimetic morphologies [J]. Soft Robot, 2018, 5: 216.

[15] Park J, You I, Kim T Y, *et al*. Ag nanowire-based transparent stretchable tactile sensor recognizing strain directions and pressure [J]. Nanotechnology, 2019, 30: 315502.

[16] Kang S, Lee J, Lee S, *et al*. Highly sensitive pressure sensor based on bioinspired porous structure for real - time tactile sensing [J]. Adv Electronic Mater, 2016, 2: 1600356.

[17] Zou L, Ge C, Wang Z J, *et al*. Novel tactile sensor technology and smart tactile sensing systems: a review [J]. Sensors: Basel, 2017, 17: 2653.

[18] Hellman R B, Tekin C, van der Schaar M, *et al*. Functional contour-following via haptic perception

and reinforcement learning[J]. IEEE T Haptics, 2017, 11: 61.

[19] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41: 1.

[20] 谢树钦, 陈梓天, 徐超, 等. 针对不可微多阶段算法的环境升级式强化学习方法[J]. 重庆邮电大学学报:自然科学版, 2020, 32: 857.

[21] Dong S, Jha D K, Romeres D, *et al*. Tactile-rl for insertion: Generalization to objects of unknown geometry [C]//Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi′an: IEEE, 2021.

[22] Chebotar Y, Hausman K, Su Z, *et al*. Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning[C]//Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, South Korean:IEEE, 2016.

[23] Merzić H, Bogdanović M, Kappler D, *et al*. Leveraging contact forces for learning to grasp[C]// Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, 2019.

[24] Melnik A, Lach L, Plappert M, *et al*. Tactile sensing and deep reinforcement learning for in-hand manipulation tasks [C]//Proceedings of the IROS Workshop on Autonomous Object Manipulation. Macau, China: IEEE, 2019.

[25] Huang S H, Zambelli M, Kay J, *et al*. Learning gentle object manipulation with curiosity-driven deep reinforcement learning [EB/OL]. [2021-05-22]. https://doi.org/10.48550/arXiv.1903.08542.

[26] Vulin N, Christen S, Stevšić S, *et al*. Improved learning of robot manipulation tasks via tactile intrinsic motivation[J]. IEEE Robot Autom Lett, 2021, 6: 2194.

[27] Ng A Y, Harada D, Russell S. Policy invariance under reward transformations: Theory and application to reward shaping [C]//Proceedings of the International Conference On Machine Learning (ICML). Bled, Slovenia: Morgan Kaufmann, 1999.

[28] Pathak D, Agrawal P, Efros A A, *et al*. Curiosity-driven exploration by self-supervised prediction [C]//Proceedings of the International Conference On Machine Learning(ICML). Sydney, Australia: PMLR, 2017.

[29] Singh S P, Barto A G, Chentanez N. Intrinsically motivated reinforcement learning [C]// Proceedings of the Advances in Neural Information Processing Systems on DBLP. Vancouver, British Columbia, Canada: DBLP, 2013.

[30] Lillicrap T P, Hunt J J, Pritzel A, *et al*. Continuous control with deep reinforcement learning[EB/OL]. (2015-09-09)[2021-05-22]. https://arxiv.org/abs/1509.02971.

[31] Andrychowicz M, Wolski F, Ray A, *et al*. Hindsight experience replay [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS). California, USA: Curran Associates Inc,2017.

[32] Brockman G, Cheung V, Pettersson L, *et al*. Openai gym [EB/OL]. (2016-06-05)[2021-05-22]. https://doi.org/10.48550/arXiv.1606.01540.

[33] Todorov E, Erez T, Tassa Y. Mujoco: a physics engine for model-based control [C]// Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Vilamoura-Algarve, Portugal: IEEE, 2012.

[34] Plappert M, Andrychowicz M, Ray A, *et al*. Multi-goal reinforcement learning: Challenging robotics environments and request for research[EB/OL]. (2018-02-28)[2021-05-22]. https://doi.org/10.48550/arXiv.1802.09464, 2018.

[35] Dhariwal P, Hesse C, Klimov O, *et al*. Openai baselines [DB/OL]. (2017-07-28)[2021-05-22]. https://github.com/openai/baselines.

[36] Marwan Q M, Chua S C, Kwek L C. Comprehensive review on reaching and grasping of objects in robotics [J]. Robotica, 2021, 39: 1849.

引用本文格式:

中　文: 宋相兵, 季玉龙, 俎文强, 等. 基于触觉传感器和强化学习内在奖励的机械臂抓取方法[J]. 四川大学学报: 自然科学版, 2022, 59: 032003.

英　文: Song X B, Ji Y L, Zu W Q, *et al*. The method for manipulator grasping based on tactile sensor and reinforcement learning intrinsic reward [J]. J Sichuan Univ: Nat Sci Ed, 2022, 59: 032003.