基于强化学习的波动鳍推进水下作业机器人悬停控制

马睿宸[1,2], 白雪剑[2,3†], 王 宇[2], 王 睿[2], 王 硕[1,2]

(1. 中国科学院大学, 北京 100049; 2. 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;
3. 中国科学院大学 人工智能学院, 北京 100049)

摘要: 本文针对波动鳍推进水下作业机器人的悬停控制问题开展研究. 首先, 给出了波动鳍推进水下作业机器人的运动学模型、动力学模型和波动鳍的参数–力映射模型, 建立了基于马尔可夫决策过程的悬停控制训练框架. 其次, 基于模型结构和训练策略, 使用强化学习的方法进行网络训练, 得到最佳的悬停控制器. 最终, 在室内水池中完成了波动鳍推进水下作业机器人的悬停控制实验, 实验结果验证了所提方法的有效性.

关键词: 水下作业机器人; 悬停控制; 波动鳍; 神经网络; 强化学习

引用格式: 马睿宸, 白雪剑, 王宇, 等. 基于强化学习的波动鳍推进水下作业机器人悬停控制. 控制理论与应用, 2022, 39(11): 2092 – 2099

DOI: 10.7641/CTA.2022.11054

Hovering control of an underwater vehicle-manipulator system propelled by undulatory fins via reinforcement learning

MA Rui-chen[1,2], BAI Xue-jian[2,3†], WANG Yu[2], WANG Rui[2], WANG Shuo[1,2]

(1. University of Chinese Academy of Sciences, Beijing 100049, China;
2. The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
3. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper addresses the hovering control of an underwater vehicle-manipulator system (UVMS) propelled by undulatory fins. First, the kinematic and dynamical models of the UVMS and a mapping model between the control parameters of undulatory fins and the driving force of the UVMS are introduced, and a hovering control training framework based on Markov decision process (MDP) is designed. Then, based on the framework and training strategies, the hovering controller is fully trained via reinforcement learning method. Finally, the well-trained controller is applied in the real environment, and the experimental results demonstrate that the proposed method can accomplish the UVMS's hovering control effectively.

Key words: underwater vehicle-manipulator system; hovering control; undulatory fin; neural network; reinforcement learning

Citation: MA Ruichen, BAI Xuejian, WANG Yu, et al. Hovering control of an underwater vehicle-manipulator system propelled by undulatory fins via reinforcement learning. *Control Theory & Applications*, 2022, 39(11): 2092 – 2099

1 引言

海洋资源的探索与开发是人类可持续发展的重要方向之一, 其中水下机器人在海洋资源勘探、环境监测等方面起到了重要的作用. 近年来, 随着海洋探索的深入, 水下打捞、水下救援、水下生物样本采集等许多领域需要水下机器人具备一定的作业能力. 因此, 水下机器人–作业臂系统(underwater vehicle-manipulator system, UVMS)应运而生.

有关UVMS控制方法的研究已取得了较大的进展: 文献[1]中使用任务优先级方法解决了UVMS的运动学冗余问题, 实现了UVMS在悬浮状态下对水下连接器的自主插/拔作业. 文献[2]提出了一种改进的任务优先级框架, 并将其用于UVMS的悬浮抓取控制中, 最终完成了对飞机黑匣子的自主打捞任务. 文献[3]针

收稿日期: 2021−11−01; 录用日期: 2022−04−15.
†通信作者. E-mail: baixuejian2018@ia.ac.cn.
本文责任编委: 闫敬.
国家自然科学基金项目(62122087, 62073316, U1806204, 62033013, U1713222), 中国科学院对外合作重点项目(173211KYSB20200020)资助.
Supported by the Natioanal Natural Science Foundation of China (62122087, 62073316, U1806204, 62033013, U1713222) and the Key Projects of Foreign Cooperation of CAS (173211KYSB20200020).

对水下作业臂与艇体间的耦合作用, 提出了一种准滑膜控制器, 仿真结果表明所提方法可以提高UVMS在水下悬停时抓取控制的稳定性. 近年来, 随着机器学习技术的不断发展, 科研人员将示范学习、强化学习等方法应用于UVMS的控制中. 文献[4]提出了一种集成了扰流动态观测的强化学习方法, 并将其应用于UVMS的悬停控制中. 文献[5]提出了一种基于参数示教学习的UVMS自主控制策略, 通过演示、建模学习、任务复现等过程, 所提算法在Girona 500水下机器人平台上实现了自主的阀门开/闭作业. 上述UVMS均采用水下螺旋桨推进器, 螺旋桨推进器可以产生可观的稳定推力, 但其推力方向较为单一.

不同于螺旋桨推进器, 南美海洋中的黑魔鬼刀鱼通过在带状鳍面上生成的行波推进方式产生推力[6], 这种推进方式被归纳为中间鳍/对鳍 (median/paired fin, MPF)推进模式. MPF推进模式可以在低速波动状态下产生平滑稳定的推进力[7–8]. 此外, 当鳍面上产生相向传播的两个正弦波时, 通过控制正弦波的相位以及波动鳍面的偏角, MPF推进模式可以同时产生可控的纵向、横向以及竖向的推进力[9–10], 其推力特性适合于UVMS的悬停控制. 通过模仿黑魔鬼刀鱼的MPF推进模式, 本课题组研制了图1所示的波动鳍推进水下作业机器人[11]. 该UVMS采用模块化的设计思想, 包含两个波动鳍推进器、一个视觉分析舱、一个主控舱和一套作业臂系统, 每部分配备了独立的电池和控制系统.

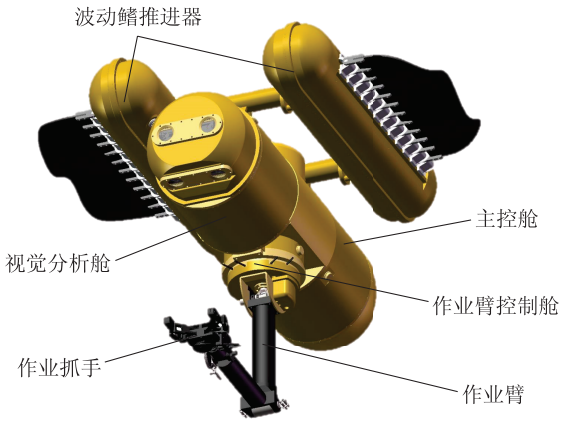
近年来, 强化学习技术被广泛用于解决各类控制问题[12]. 在使用强化学习技术解决优化控制问题时, 需要将控制问题构造为马尔可夫决策过程(Markov decision process, MDP), 一个标准的MDP如图2所示, 智能体按照一定策略对环境执行动作, 并根据感知到的环境状态获得相应的奖励, 为获得长期的最大累积奖励, 智能体会不断更新策略, 最终会获得特定性能指标下的最优控制策略. 本文采用强化学习方法, 针对波动鳍推进水下作业机器人的悬停控制问题开展研究, 建立了UVMS模型, 提出了基于MDP的悬停控制训练框架, 并基于模型结构和训练策略进行网络训练, 得到了最佳的悬停控制器, 最终通过水池实验验证了所提方法的有效性.

2　UVMS 的建模

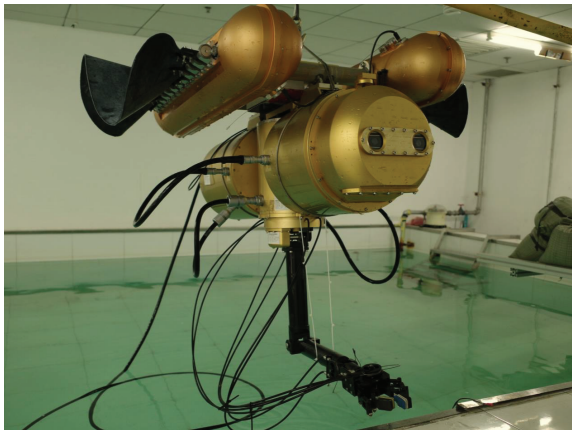
本节给出了UVMS的坐标系统, 并对相关模型进行构建. 首先, 建立了UVMS的运动学与动力学模型. 其次, 描述了波动鳍的波形与力学特性. 最后, 建立了波动鳍参数与UVMS所受推进力之间的映射模型.

2.1　UVMS 的运动学与动力学模型

控制对象UVMS可看做一个刚体, 其重心在竖直方向上低于浮心, 不易产生俯仰和横滚运动, 因此可忽略这两个自由度上的运动. 如图3所示, 在世界坐标系 $E-xyz$ 中, 机器人的空间位姿可描述为 $\boldsymbol{\chi}=[x\ y\ z\ \psi]^{\mathrm{T}}$. 随体坐标系$O-uvw$固定在机器人的几何中心, 在该坐标系中, 机器人的速度可表示为$\boldsymbol{\nu}=[u\ v\ w\ r]^{\mathrm{T}}$, 所受外力可表示为$\boldsymbol{\tau}=[X\ Y\ Z\ N]^{\mathrm{T}}$.



(a) UVMS结构图



(b) UVMS原型机照片

图 1 UVMS的结构图与照片

Fig. 1 Schematic and photo of the UVMS prototype

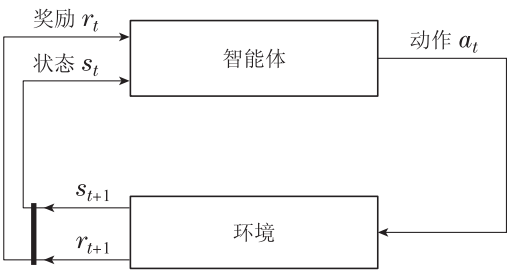


图 2　MDP的结构

Fig. 2　Structure of a MDP

所述UVMS的运动学模型可表示为

$$\dot{\boldsymbol{\chi}}=\boldsymbol{J}(\psi)\boldsymbol{\nu}, \tag{1}$$

所述UVMS的动力学模型可表示为

$$\boldsymbol{M}\dot{\boldsymbol{v}}+\boldsymbol{C}(\boldsymbol{v})\boldsymbol{v}=\boldsymbol{\tau}_{\mathrm{p}}-\boldsymbol{\tau}_{\mathrm{h}}(\boldsymbol{v},\dot{\boldsymbol{v}})-\boldsymbol{\tau}_{\mathrm{d}}, \tag{2}$$

其中: $\boldsymbol{J}(\psi) \in \mathrm{SO}(4)$是坐标变换矩阵, $\boldsymbol{M} \in \mathbb{R}^{4\times4}$是惯性矩阵, $\boldsymbol{C}(\boldsymbol{\nu}) \in \mathbb{R}^{4\times4}$是向心力和科氏力矩阵, $\boldsymbol{\tau}_{\mathrm{p}}$是UVMS所受推进力, $\boldsymbol{\tau}_{\mathrm{h}}(\boldsymbol{v},\dot{\boldsymbol{v}}) \in \mathbb{R}^4$是水作用力, $\boldsymbol{\tau}_{\mathrm{d}} \in \mathbb{R}^4$是其他外力.

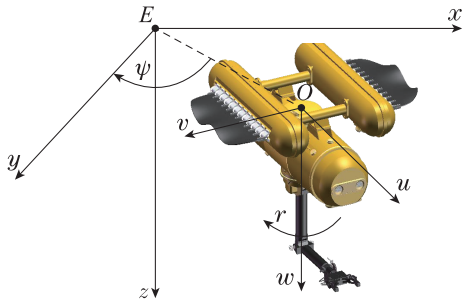


图 3 UVMS 的坐标系统

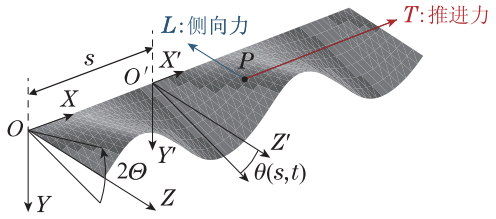
Fig. 3 Coordinate systems for the UVMS

2.2 波动鳍的波形与力学特性

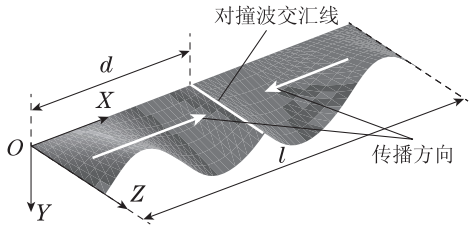
如图4所示, 波动鳍通过正弦波的行波运动来改变周围的水流状态从而产生推力.本文中的波动鳍推进器使用两种波形, 它们可由下式表示:

$$\begin{cases} \theta_1(s,t) = \Theta\sin[2\pi(\dfrac{s}{\lambda} - ft)], & s \leqslant d, \\ \theta_2(s,t) = \Theta\sin[2\pi(\dfrac{2d-s}{\lambda} - ft)], & s > d, \end{cases} \tag{3}$$

其中: $\theta_i(s,t)\ (i=1,2)$是波动鳍鳍面在t时刻s位置绕X轴旋转的角度值, Θ是最大旋转角度, λ是波长, f是频率, l是鳍面的总长度, d是两个波交汇处的位置. $\theta_1(s,t)$和$\theta_2(s,t)$分别描述了图4(b)中相向传播的两个波形. 在实际控制中, 波动鳍的控制量仅采用频率f和交汇位置d, 其他设为定值. 为方便表示, 令$\eta = d/l$, 并将η代替d作为控制量. 当$\eta = 0$或$\eta = 1$时, 可视为采用图4(a)中的单独正弦波.



(a) 正弦波



(b) 对撞正弦波

图 4 波动鳍采用的波形

Fig. 4 Wave patterns for the undulatory fin

波动鳍在水中产生的力主要作用在两个方向. 如图4(a)所示, 假设力的作用点为P, 则这两个力为: 沿X轴方向的推进力T和垂直于X轴方向的侧向力L. 通过波动鳍的力测量平台[13], 将多个等间距分布的f和η对T和L的影响进行测量后, 使用二元样条插值法, 可把结果拟合成这两种力的表达: $T(f,\eta)$和$L(f,\eta)$. 这两个力同时由f和η决定, 相互之间存在耦合关系.

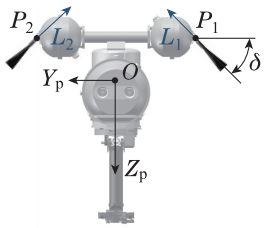
2.3 波动鳍参数–力映射模型

本小节分析了波动鳍推进器的波形参数与UVMS所受推进力$\boldsymbol{\tau}_{\mathrm{p}}$之间的映射关系. 首先, 在图5中对$\boldsymbol{\tau}_{\mathrm{p}}$和波动鳍推力之间的关系进行分析.

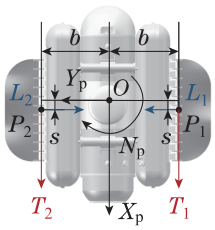
由图5知, $\boldsymbol{\tau}_{\mathrm{p}} = [\,X_{\mathrm{p}}\ Y_{\mathrm{p}}\ Z_{\mathrm{p}}\ N_{\mathrm{p}}\,]^{\mathrm{T}}$可表示为

$$\boldsymbol{\tau}_{\mathrm{p}}(f_1,\eta_1,f_2,\eta_2) = \begin{bmatrix} T_1(f_1,\eta_1) + T_2(f_2,\eta_2) \\ \cos\delta[L_2(f_2,\eta_2) - L_1(f_1,\eta_1)] \\ \sin\delta[L_1(f_1,\eta_1) + L_2(f_2,\eta_2)] \\ b[T_1(f_1,\eta_1) - T_2(f_2,\eta_2)] + s[L_2(f_2,\eta_2) - L_1(f_1,\eta_1)] \end{bmatrix}, \tag{4}$$

其中: 角标1代表左鳍, 2代表右鳍, δ为波动鳍整体绕连接处的偏转角度.



(a) 主视图



(b) 俯视图

图 5 波动鳍所产生的合力

Fig. 5 Resultant force of undulatory fins

3 训练框架

本节基于UVMS的模型, 构建了悬停控制任务的训练框架, 如图6所示, 图中带角标的参数表示下一时刻的该参数(如位姿信息$\boldsymbol{\chi}'$表示下一时刻的$\boldsymbol{\chi}$), dt表示训练框架中的时间间隔. 接下来本文将从悬停控制问题定量描述和MDP的构建两个方面来详细介绍.

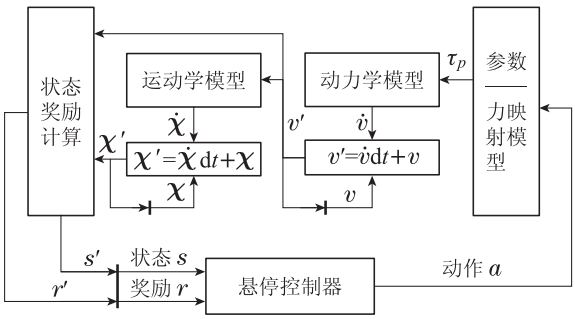


图 6 悬停控制任务的训练框架

Fig. 6 The training framework for the hovering control task

3.1 悬停控制问题定量描述

如图7所示, $\boldsymbol{\chi}_{\rm h}=[\,x_{\rm h}\ y_{\rm h}\ z_{\rm h}\ \psi_{\rm h}\,]^{\rm T}$表示悬停位姿. 悬停控制旨在让UVMS从当前位姿出发, 移动并保持到悬停位姿. 通过定量分析, 悬停控制需要满足以下3项指标:

1) 水平距离

$$\sqrt{(x-x_{\rm h})^2+(y-y_{\rm h})^2}<e_{xy}, \tag{5}$$

2) 竖直距离

$$|z-z_{\rm h}|<e_z, \tag{6}$$

3) 偏航角度差

$$|\psi-\psi_{\rm h}|<e_\psi, \tag{7}$$

其中: e_{xy}, e_z, e_ψ为3项指标的最大误差, 具体值为: $e_{xy}=0.05$ m, $e_z=0.05$ m, $e_\phi=0.05$ rad.

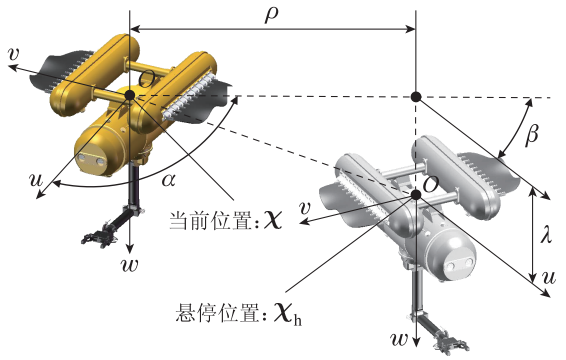


图 7 悬停控制的问题描述

Fig. 7 Description of the hovering control

3.2 MDP的构建

一个MDP包含4个部分: 动作空间$\mathcal{A}$, 状态空间$\mathcal{S}$, 奖励$r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$以及状态转移矩阵$p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{a}_{t-1})$, 本文悬停控制任务的MDP可构建为

1) 动作$\boldsymbol{a}\in\mathcal{A}$.

$$\boldsymbol{a}=[\,f_1\ \eta_1\ f_2\ \eta_2\,]^{\rm T}. \tag{8}$$

动作是MDP中悬停控制器所能控制的参数, 这里沿用了波动鳍波形的控制参数.

2) 状态$\boldsymbol{s}\in\mathcal{S}$.

$$\boldsymbol{s}=[\,\rho\ \lambda\ \alpha\ \beta\ \gamma\ \boldsymbol{\nu}\,]^{\rm T}. \tag{9}$$

状态是MDP中悬停控制器所能感知到的关于环境的描述, 如图7所示, $\boldsymbol{s}$中的部分参数已标出, 前5项参数的具体计算公式为

$$\begin{cases}\rho=\sqrt{(x-x_{\rm h})^2+(y-y_{\rm h})^2},\\ \lambda=z-z_{\rm h},\\ \alpha=\arctan2(y-y_{\rm h},x-x_{\rm h}),\\ \beta=\alpha-\psi+\psi_{\rm h},\\ \gamma=\psi-\psi_{\rm h},\end{cases} \tag{10}$$

其中: ρ, α, β是UVMS在水平方向上趋近悬停位姿需要的状态信息, λ是UVMS在竖直方向上趋近悬停位姿需要的状态信息, γ是UVMS 在偏航角上趋近悬停位姿需要的状态信息, $\boldsymbol{\nu}$是速度, 如式(10)所示, 状态$\boldsymbol{s}$中各项的设计大多数采用相对值, 这样能用较少的状态代表更多的情况, 有利于缩小状态空间, 加快求解.

3) 奖励$r(\boldsymbol{s})$.

$$r(\boldsymbol{s})=-k_1\rho^2-k_2\lambda^2-k_3\gamma^2-k_4\boldsymbol{\nu}^{\rm T}\boldsymbol{M}\boldsymbol{\nu}. \tag{11}$$

奖励函数是MDP中悬停控制器为完成任务所需的引导, 通常情况下越趋近于完成任务, 其值应当越大. 对于悬停任务, 式(11)中$k_i(i=1\cdots,4)$是每项的权重, $k_1\rho^2$ 驱使UVMS与悬停位姿间水平距离的缩小, $k_2\lambda^2$驱使竖直间距的缩小, $k_3\gamma^2$驱使偏航角差值的缩小, $k_4\boldsymbol{\nu}^{\rm T}\boldsymbol{M}\boldsymbol{\nu}$驱使动能缩小, 便于最终的悬停.

4) 状态转移矩阵.

状态转移矩阵用于描述悬停控制器在与环境交互时, 由动作$\boldsymbol{a}$ 引起的状态$\boldsymbol{s}$的变化. 如图6所示, 基于UVMS的运动学、动力学模型和波动鳍的参数–力映射模型, 由$\boldsymbol{a}$到$\boldsymbol{s}$的变化得以描述(文献[14]中更具体地介绍了该过程).

4 悬停控制器的训练

本节介绍了悬停控制器的训练过程. 首先, 给出了基于柔性行动器–评判器(soft actor-critic, SAC)的强化学习算法的求解函数. 其次, 使用神经网络来拟合强化学习训练中涉及的非线性关系. 再次, 针对UVMS特点, 提出了几种训练策略. 最终, 给出了整个求解过程的具体算法.

4.1 基于SAC的强化学习方法

本文中所使用的强化学习方法是基于SAC[15]设计的. 其根据当前状态会给出最佳动作的概率分布而不是一个特定的动作, 可将这种策略表示为$\pi(\cdot|\boldsymbol{s})$. 式(11)中已设计了奖励来引导UVMS完成任务, 但仅靠奖励函数无法对未来的奖励进行判断. 为了描述长时间内能获得的累积奖励并用最大熵的方式强化探索范围, SAC中设立了如下的函数:

$$V^{\pi}(\boldsymbol{s})=\mathop{\rm E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t(r(\boldsymbol{s}_t)+\delta H(\pi(\cdot|\boldsymbol{s}_t)))|\boldsymbol{s}_t], \tag{12}$$

$$Q^{\pi}(\boldsymbol{s},\boldsymbol{a})=\mathop{\rm E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{s}_t)+\delta\sum_{t=1}^{\infty}\gamma^t H(\pi(\cdot|\boldsymbol{s}_t))|\boldsymbol{s}_t,\boldsymbol{a}_t], \tag{13}$$

其中: $V^{\pi}(\boldsymbol{s})$被称为价值函数, 用于描述一个状态的未来累积价值. $Q^{\pi}(\boldsymbol{s},\boldsymbol{a})$ 被称为动作价值函数, 也被称作Q–函数, 用于描述一个状态动作对组合的未来累

积价值. 两式中t是时间步, $\tau=(\boldsymbol{s}_0,\boldsymbol{a}_0,\boldsymbol{s}_1,\boldsymbol{a}_1,\cdots)$是状态与动作的序列, γ是奖励函数的折扣, δ是熵的权重, $H(\pi(\cdot|\boldsymbol{s}_t))=\mathrm{E}[-\log\pi(\boldsymbol{a}_t|\boldsymbol{s}_t)]$是熵. 价值函数和$Q$–函数之间的关系可表示为

$$V^{\pi}(\boldsymbol{s}_t)=\mathop{\mathrm{E}}_{\tau\sim\pi}[Q^{\pi}(\boldsymbol{s}_t,\boldsymbol{a}_t)]+\delta H(\pi(\cdot|\boldsymbol{s}_t)). \quad (14)$$

贝尔曼方程是动态规划中实现最优化的必要工具. 在SAC中, 贝尔曼方程为

$$Q^{\pi}(\boldsymbol{s}_t,\boldsymbol{a}_t)=\mathop{\mathrm{E}}_{\boldsymbol{s}_{t+1}\sim P}\left[r\left(\boldsymbol{s}_t\right)+\gamma V^{\pi}\left(\boldsymbol{s}_{t+1}\right)\right]. \quad (15)$$

为了更新Q–函数, 本文用贝尔曼均方差误差(mean-squared Bellman error, MSBE)来描述Q–函数满足贝尔曼方程的程度, 定义如下:

$$L\left(\mathcal{D}|\phi_i,\right)=\mathop{\mathrm{E}}_{b\sim\mathcal{D}}[(Q_i^{\pi}(\boldsymbol{s}_t,\boldsymbol{a}_t|\phi_i)-y)^2], \quad (16)$$

其中: $\mathcal{D}$是回放缓存, 用于存储训练过程中已产生的信息, ϕ是Q–函数的拟合系数, $i=1,2$表示有两个Q–函数, 这是采用了双Q–函数学习[16]的策略, $b=(\boldsymbol{s}_t,\boldsymbol{a}_t,r(\boldsymbol{s}_t),\boldsymbol{s}_{t+1},d_t)$是一段从$\mathcal{D}$中生成的序列, 其中$d$是任务是否完成的指示信号, y是MSBE中所趋近的目标

$$y=r(\boldsymbol{s}_t)+\gamma(1-d_t)\big(\min_{i=1,2}Q_i'^{\pi}\left(\boldsymbol{s}_{t+1},\tilde{\boldsymbol{a}}_{t+1}|\phi_i\right)-\delta\log\pi\left(\tilde{\boldsymbol{a}}_{t+1}|\boldsymbol{s}_{t+1}\right)\big), \quad (17)$$

其中: $\tilde{\boldsymbol{a}}_{t+1}\sim\pi\left(\cdot|\boldsymbol{s}_{t+1},\theta\right)$是现有策略最新执行的动作, $Q_i'^{\pi}\ (i=1,2)$ 是目标函数, θ是策略拟合系数.

控制问题求解的目标是得到最优策略π. 基于式(14), 本文采用了一种重新赋参[17]的策略, 将策略优化的标准描述为

$$\max_{\theta}\mathop{\mathrm{E}}_{\boldsymbol{s}_t\sim\mathcal{D}}[\min_{i=1,2}Q_i^{\pi}\big(\boldsymbol{s}_t,\tilde{\boldsymbol{a}}(\boldsymbol{s}_t,\xi|\theta)|\phi_i\big)-\delta\log\pi\big(\tilde{\boldsymbol{a}}(\boldsymbol{s}_t,\xi|\theta)|\boldsymbol{s}_t\big)], \quad (18)$$

其中: $\tilde{\boldsymbol{a}}(\boldsymbol{s}_t,\xi|\theta)=\tanh\big(\mu(\boldsymbol{s}|\theta)+\sigma(\boldsymbol{s}|\theta)\odot\xi\big)$, ξ是高斯分布, $\mu(\boldsymbol{s}|\theta)$是期望, $\sigma(\boldsymbol{s}|\theta)$是标准差.

4.2 神经网络参数拟合

控制策略和Q–函数都是非线性的, 需要采用神经网络进行拟合. 控制策略的神经网络如图8(a)所示, $\boldsymbol{\theta}$表示神经元的权重. 隐藏层采用ReLU作为激活函数, 共有两层, 其神经元数分别为400和300. 输出层采用Tanh作为激活函数, 神经元数为4, 与式(8)中$\boldsymbol{a}$参数的维度相匹配. Q–函数的神经网络如图8(b)所示, $\boldsymbol{\phi}$表示神经元的权重. 隐藏层采用ReLU作为激活函数, 共有两层, 神经元数分别为600和400. 在输出层中, 神经元数为10, 采用Linear激活函数.

4.3 训练策略

强化学习所训练出的悬停控制器最终要在实验环境中的UVMS上进行使用, 但仿真环境与实验环境间必然存在误差. 为减小这些误差所带来的影响, 本文针对UVMS的特点, 提出了如下处理方案:

1) 引入随机负载.

在实验环境中UVMS的负载量可能会因为UVMS本身的不理想配平或作业臂抓/放重物而发生改变. 为此本文设计了随机负载来描述UVMS负载的不确定性

$$l=l_0(1+2l_1(\mathrm{rand}(1)-0.5)), \quad (19)$$

其中: l_0为基础负载, l_1为随机范围, rand是随机函数, 可生成$[0,1]$内的一个随机数. 用l即可生成$[l_0(1-b),l_0(1+b)]$范围内的一个随机数. l作用在动力学模型(2)中$\boldsymbol{\tau}_{\mathrm{d}}$的$z$方向上.

2) 动作中引入干扰.

在强化学习训练过程中, 对动作(8)进行处理可模拟环境中的干扰, 为此本文设计了如下的干扰因子

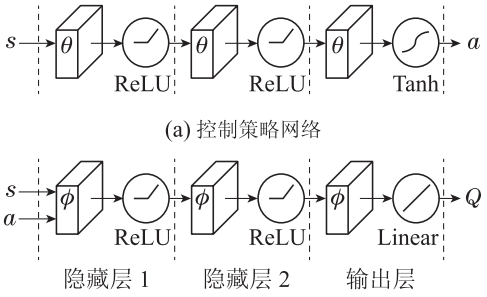
$$p=(1-p_0\mathrm{rand}(1)), \quad (20)$$

其中: p_0为扰动范围, 用p即可生成$[1-p_0,1]$之间的随机数, 与动作$\boldsymbol{a}$中各项直接相乘来进行干扰.

3) 动力学模型预处理.

动力学模型(2)中$\boldsymbol{M}$, $\boldsymbol{C}$和$\boldsymbol{\tau}_{\mathrm{h}}\left(\boldsymbol{v},\dot{\boldsymbol{v}}\right)$反映了UVMS的力学特征. 它们中的各个参数虽然已计算得出[14], 但必然和真实环境存在误差. 因此在强化学习的不同训练周期, 可对这些参数进行随机化处理, 从而提升控制器对不同模型的适应性, 以便适应实验环境中的UVMS模型. 为此本文在$\boldsymbol{M}$, $\boldsymbol{C}$和$\boldsymbol{\tau}_{\mathrm{h}}\left(\boldsymbol{v},\dot{\boldsymbol{v}}\right)$中的每一项中加入高斯噪声来进行处理, 即

$$n=\mathcal{N}(1,{n_0}^2). \quad (21)$$



(a) 控制策略网络

(b) Q–函数网络

图 8 神经网络结构图

Fig. 8 Structure of neural networks

4.4 强化学习的训练过程与结果

本文所设计的强化学习方法需要不断训练来优化悬停控制器, 算法1详细介绍了该训练过程, 其中d是任务的完成信号, 当连续30个训练步数内3项误差指标同时满足式(5)–(7)时$d=1$, 否则$d=0$.

强化学习算法的训练周期$\boldsymbol{M}=1.5\times10^4$, 每个周期最大步数$T=500$. 图9展示了训练过程中累积奖励的变化过程, 其中, 每100个周期累积奖励的平均值用黑线表示, 最大值用灰色区域的上边缘表示, 最小

值用下边缘表示. 由图9可见累积奖励的波动逐渐变小, 说明控制策略逐渐稳定. 最后很长一段训练周期内, 累积奖励无较大波动, 表明控制策略的训练已经收敛. 图10展示了训练过程中每100 个周期的任务的成功率, 在训练后期, 成功率可维持在1附近, 这进一步反映了本文所提出训练方法的有效性.

算法 1: 悬停任务控制策略的训练过程

Input: 训练周期M, 最大步数T, 缓存容量R, 批处理大小N, 奖励折扣γ, 熵权重δ, Polyak系数ϵ

Output: 悬停控制器$\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})$

初始化策略π与其$\boldsymbol{\theta}$;
初始化Q–函数Q_1, Q_2与其权重$\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$;
初始化目标Q_1', Q_2'与其权重$\boldsymbol{\phi}_1'$, $\boldsymbol{\phi}_2'$;
初始化回放缓存$\mathcal{D}$与其容量R;
for 周期= 1 **to** M **do**
  用式(21)预处理动力学模型的参数;
  初始化$\boldsymbol{\chi}$, $\boldsymbol{\nu}$和$\boldsymbol{\tau}$均为$[0\ 0\ 0\ 0\ 0]^{\mathrm{T}}$;
  随机生成目标位姿$\boldsymbol{\chi}_{\mathrm{h}}$;
  依据式(19)初始化随机负载;
  依据式(10)计算初始状态$\boldsymbol{s}$.
  for 步数= 1 **to** T **do**
    依据$\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})$选择动作;
    依据式(20)初始化干扰因子, 并处理$\boldsymbol{a}$;
    执行$\boldsymbol{a}$;
    观测状态$\boldsymbol{s}'$, 计算奖励r.
    观测信号d;
    把$(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}',d)$ 存入回放缓存$\mathcal{D}$;
    if d = 1 **or** 步数 = T **then**
      从$\mathcal{D}$中生成样本:
      $B=\{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}',d)_i\}\ (i=1,\cdots,N)$;
      根据式(17)计算目标y;
      用梯度下降更新Q_i:
      $\nabla_{\phi_i}\dfrac{1}{|B|}\sum\limits_{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}',d)\in B}\left(Q_i(\boldsymbol{s},\boldsymbol{a}|\boldsymbol{\phi}_i)-y\right)^2\ (i=1,2)$;
      用梯度上升更新π:
      $\nabla_\theta\dfrac{1}{|B|}\sum\limits_{\boldsymbol{s}\in B}\big(\min\limits_{i=1,2}Q_i\big(\boldsymbol{s}_t,\tilde{\boldsymbol{a}}(\boldsymbol{s}_t,\xi|\boldsymbol{\theta})|\boldsymbol{\phi}_i\big)-\delta\log\pi\big(\tilde{\boldsymbol{a}}(\boldsymbol{s}_t,\xi|\boldsymbol{\theta})|\boldsymbol{s}_t\big)\big)$;
      更新目标网络:
      $\boldsymbol{\phi}_i'\leftarrow\epsilon\boldsymbol{\phi}_i'+(1-\epsilon)\boldsymbol{\phi}_i\ \ (i=1,2)$;
    end
  end
end

5　实验与分析

为了验证控制方法的有效性, 本节使用训练的悬停控制器, 在室内水池开展了UVMS的悬停位姿切换实验. 首先, 介绍了实验环境. 然后, 介绍了实验过程与结果分析.

5.1　实验环境

图11给出了悬停实验的控制系统框图. 实验在一个5 m×4 m×1.5 m(长×宽×高)的室内水池中进行, 其水深为1.2 m. 为获取UVMS在水平面上的位姿信息, 本文在UVMS顶部的前后舱体上分别粘贴了红色、蓝色标记. 通过固定在天花板上的工业摄像头(MER–160–227U3C)识别这两种标记并进行图像处理, 即可得到两个标记的位置信息, 进而推算得到UVMS的x, y, ψ坐标信息. 为获取机器人的深度信息, 本文用UVMS主控舱中的深度传感器(MS5837–30BA) 对UVMS的z坐标进行计算. 基于上述位姿信息, 可通过微分跟踪器计算出速度信息. 结合位置与速度信息, 可计算得到式(9)中的状态$\boldsymbol{s}$. 训练出的控制器以图8(a)中神经网络的形式存在, 可基于当前状态$\boldsymbol{s}$得到UVMS所需执行的动作$\boldsymbol{a}$, 进而实现悬停控制.

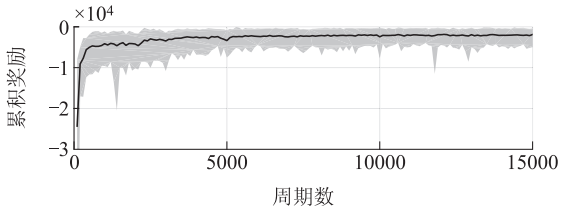


图 9　训练过程中的累积奖励

Fig. 9　Cumulated rewards during training

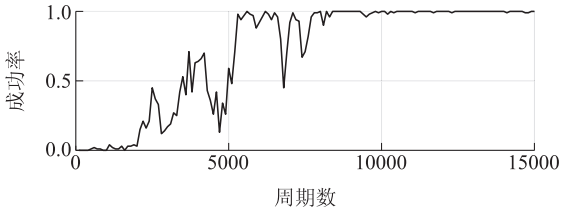


图 10　训练过程中的成功率

Fig. 10　Success rates during training

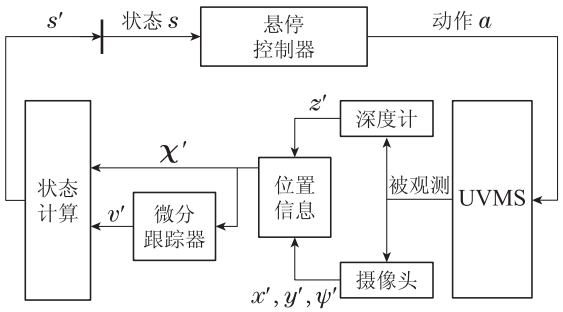


图 11　悬停实验的控制系统框图

Fig. 11　The control system for the hovering control experiment

5.2　悬停位姿切换实验

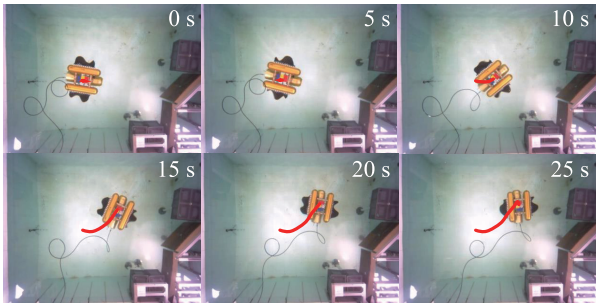
悬停位姿切换实验的目标是使UVMS从一个悬停位姿运动到另一个悬停位姿并保持在一定的误差范围内. UVMS初始悬停位姿为

$$\boldsymbol{\chi}_{\mathrm{s}}=[-0.5\ \ 0\ \ 0.4\ \ -\pi/12]^{\mathrm{T}},$$

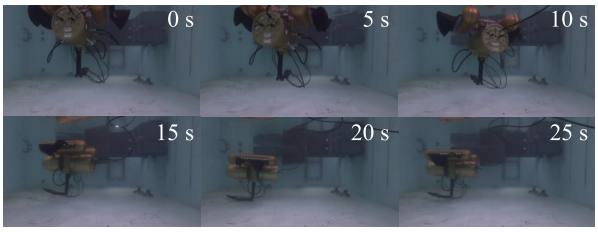
目标悬停位姿为

$$\boldsymbol{\chi}_{\mathrm{h}}=[0.7\ \ -0.5\ \ 0.75\ \ -\pi/2]^{\mathrm{T}}.$$

图12和图13分别给出了UVMS悬停位姿切换实验的视频截图序列和空间航迹示意图. 由图12可以看出, UVMS在25 s的时间内从初始悬停位姿运动到了目标悬停位姿.



(a) 俯视视角



(b) 水下视角

图 12 实验视频截图序列

Fig. 12 Snapshot sequences of experiment

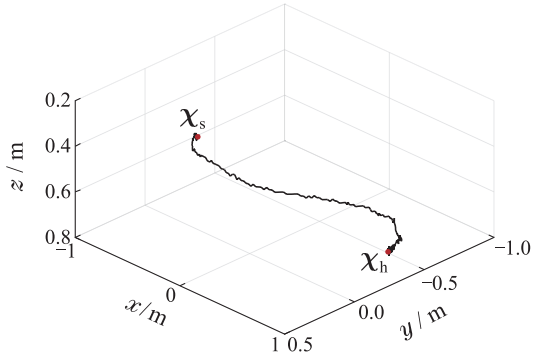


图 13 实验中UVMS的空间航迹

Fig. 13 Spatial path of the UVMS in the experiment

实验过程中UVMS坐标值的变化见图14, 控制量的变化见图15, 实时奖励见图16. 从图14中可看出, 控制器可使UVMS各坐标收敛到目标悬停位姿坐标的有界区域内, 并较长时间的保持. 保持期间, UVMS和目标悬停位姿$\boldsymbol{\chi}_{\rm h}$之间的平均水平距离为0.0240 m, 平均竖直距离为0.0301 m, 平均偏航角度差为0.0428 rad, 满足式(5)–(7)所给出的3项指标要求. 图15中记录的控制量是归一化后的, 用上标$*$加以表示. 归一化前, f的取值范围为$[0, 1.7]$ Hz, η的取值范围为$[0, 1]$. 由图11可知, 在悬停控制器和实验环境的交互过程中, 并不需要对奖励r进行计算, 但r也能侧面反应实验进程, 如图16所示, 可看出悬停位姿的切换过程: 在时间为5 s时, 奖励r瞬间突变, 说明此时的期望悬停位姿由当前所保持的$\boldsymbol{\chi}_{\rm s}$ 切换至了目标位姿$\boldsymbol{\chi}_{\rm h}$. 从式(11)可知, r值越接近0, 悬停误差越小, 而该控制器在初始悬停位姿$\boldsymbol{\chi}_{\rm s}$ 和目标悬停位姿$\boldsymbol{\chi}_{\rm h}$处都能保持较接近0的奖励值, 这也进一步说明了悬停控制器的有效性. 由上述实验结果可知, 本文所设计的控制方法在应用于UVMS的悬停位姿切换实验时控制效果较好, 具有一定的可行性.

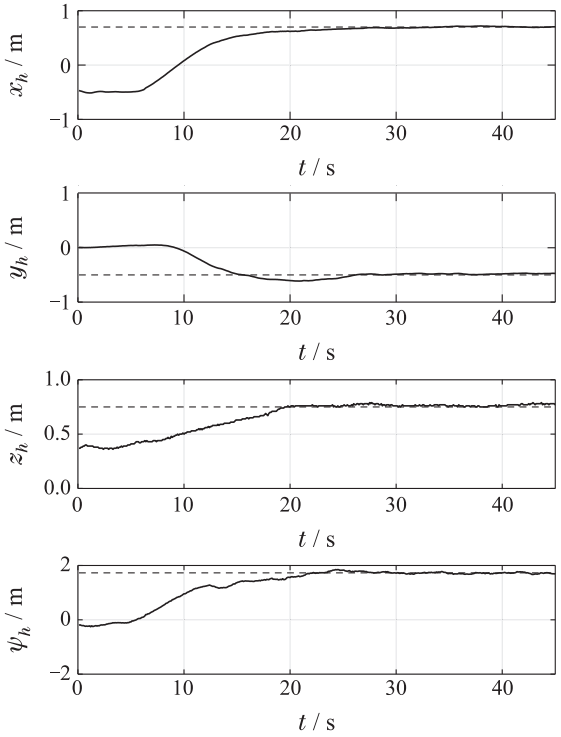


图 14 实验中UVMS的坐标值

Fig. 14 Coordinate value of the UVMS in the experiment

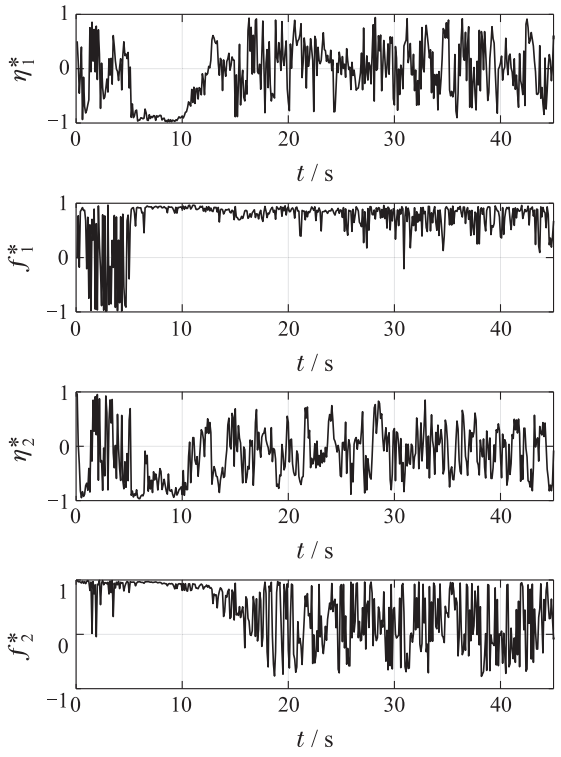


图 15 实验中UVMS的控制量

Fig. 15 Controlled variables of the UVMS in the experiment

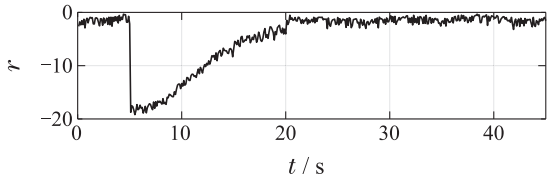


图 16 实验中 UVMS 的实时奖励

Fig. 16 Rewards of the UVMS in the experiment

6 结论

本文针对波动鳍推进水下作业机器人的悬停控制问题开展研究. 首先, 建立了UVMS的模型, 根据MDP结构构建了UVMS的悬停控制训练框架. 其次, 给出了结合UVMS特点的训练策略, 使用强化学习的方法, 通过神经网络训练获得了UVMS的悬停控制器. 最后, 在室内水池开展了UVMS的悬停控制实验, 实验结果验证了所提方法的有效性和可行性.

参考文献:

[1] YOUAKIM D, RIDAO P, PALOMERAS N, et al. MoveIt!: Autonomous underwater free-floating manipulation. *IEEE Robotics & Automation Magazine*, 2017, 24(3): 41 – 51.

[2] SIMETTI E, CASALINOG, TORELLI S, et al. Floating underwater manipulation: Developed control methodology and experimental validation within the TRIDENT project. *Journal of Field Robotics*, 2014, 31(3): 364 – 385.

[3] ZHU Qi. *Research on attitude control method of intervention UVMS.* ZheJiang: Zhejiang University, 2018.
(朱琦. 作业型水下机器人姿态控制方法研究. 浙江: 浙江大学, 2018.)

[4] WANG T, LU W, YAN Z, et al. Dob-net: Actively rejecting unknown excessive time-varying disturbances. *International Conference on Robotics and Automation*, Singapore: IEEE, 2020: 1881 – 1887.

[5] CARRERA A, PALOMERAS N, RIBAS D, et al. An intervention-AUV learns how to perform an underwater valve turning. *Oceans 2014-taipei*. Taipei, China: IEEE, 2014: 1 – 7.

[6] NEVELN I D, BAI Y, SNYDER J B, et al. Biomimetic and bio-inspired robotics in electric fish research. *Journal of Experimental Biology*, 2013, 216(13): 2501 – 2514.

[7] NEVELN I D, BALE R, BHALLA A P S, et al. Undulating fins produce off-axis thrust and flow structures. *Journal of Experimental Biology*, 2014, 217(2): 201 – 213.

[8] CURET O M, PATANKAR N A, LAUDER G V, et al. Aquatic manoeuvering with counter-propagating waves: A novel locomotive strategy. *Journal of The Royal Society Interface*, 2011, 60(8): 1041 – 1050.

[9] RUIZ-TORRES R, CURET O M, LAUDER G V, et al. Kinematics of the ribbon fin in hovering and swimming of the electric ghost knifefish. *Journal of Experimental Biology*, 2013, 216(5): 823 – 834.

[10] SEFATI S, NEVELN I, MACIVER M A, et al. Counter-propagating waves enhance maneuverability and stability: A bio-inspired strategy for robotic ribbon-fin propulsion. *International Conference on Biomedical Robotics and Biomechatronics*. Rome, Italy: IEEE/RAS-EMBS, 2012: 1620 – 1625.

[11] WANG Y, WANG R, WANG S, et al. Underwater bioinspired propulsion: From inspection to manipulation. *IEEE Transactions on Industrial Electronics*, 2019, 67(9): 7629 – 7638.

[12] XU S, LIU J, YANG C, et al. A learning-based stable servo control strategy using broad learning system aplied for microrobotic control. *IEEE Transactions on Cybernetics*, 2021, DOI: 10.1109/TCYB.2021.3121080.

[13] MA R, WANG Y, WANG R, et al. Development of a propeller with undulating fins and its characteristics. *International Conference on Real-time Computing and Robotics*. Irkutsk, Russia: IEEE, 2019: 737 – 742.

[14] MA R, WANG Y, GAO Z, et al. Position control of an underwater biomimetic vehicle-manipulator system via reinforcement learning. *Data Driven Control and Learning Systems Conference*. Liuzhou, China: IEEE, 2020: 573 – 578.

[15] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning*. Hanoi, Vietnam: PMLR, 2018: 1861 – 1870.

[16] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods. *International Conference on Machine Learning*. Hanoi, Vietnam: PMLR, 2018: 1587 – 1596.

[17] KINGMA D P, WELLING M. Auto-encoding variational bayes, *arXiv preprint*, 2013, arXiv:1312.6114.

作者简介:

马睿宸　博士研究生, 目前研究方向为智能控制、机器人学、水下机器人、仿生机器人, E-mail: maruichen2016@ia.ac.cn;

白雪剑　博士研究生, 目前研究方向为智能控制、机器人学、水下机器人、仿生机器人, E-mail: baixuejian2018@ia.ac.cn;

王　宇　博士, 副研究员, 目前研究方向为智能控制、机器人学、水下机器人、仿生机器人, E-mail: yu.wang@ia.ac.cn;

王　睿　博士, 副研究员, 目前研究方向为智能控制、机器人学、水下机器人、仿生机器人, E-mail: rwang5212@ia.ac.cn;

王　硕　博士, 研究员, 博士生导师, 目前研究方向为水下机器人、机器学习、多机器人系统, E-mail: shuo.wang@ia.ac.cn.