doi:10.19306/j.cnki.2095-8110.2022.06.004

基于深度强化学习的无人机栖落机动控制策略设计

黄　赞,何　真,仇靖雯

(南京航空航天大学自动化学院,南京 211106)

摘　要:无人机栖落机动飞行是一种无需跑道的降落方法,能够提升无人机在复杂环境下执行任务的适应能力。针对具有高非线性、多约束特性的无人机栖落机动过程,提出了一种基于模仿深度强化学习的控制策略设计方法。首先,建立了固定翼无人机栖落机动的纵向非线性动力学模型,并设计了无人机栖落机动的强化学习环境。其次,针对栖落机动状态动作空间大的特点,为了提高探索效率,通过模仿专家经验的方法对系统进行预训练。然后,以模仿学习得到的权重为基础,采用近端策略优化方法学习构建无人机栖落机动的神经网络控制器。最后,通过仿真验证了上述控制策略设计方法的有效性。

关键词:栖落机动;深度强化学习;固定翼无人机;神经网络

中图分类号:V249　　**文献标志码**:A　　**文章编号**:2095-8110(2022)06-0025-08

Design of UAV Perching Maneuver Control Strategy Based on Deep Reinforcement Learning

HUANG Zan, HE Zhen, QIU Jing-wen

(College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: UAV perching maneuvering is a landing method that does not require a runway, which can improve the adaptability of UAV to perform tasks in complex environments. Aiming at the UAV perching maneuver process with high nonlinearity and multi-constraint characteristics, a control strategy design method based on imitating deep reinforcement learning is proposed. Firstly, a longitudinal nonlinear dynamic model of fixed-wing UAV perching maneuver is established, and a reinforcement learning environment for UAV perching maneuver is designed. Secondly, in view of the large action and state space of the perching maneuver, to improve the exploration efficiency, the system is pre-trained by imitating the experience of experts. Then, based on the weights obtained by imitation learning, the proximal policy optimization method is used to learn to build a neural network controller for UAV perching maneuver. Finally, simulations verify the effectiveness of the control strategy design method.

Key words: Perching maneuver; Deep reinforcement learning; Fixed-wing UAV; Neural network

收稿日期:2022-08-09;**修订日期**:2022-09-22

基金项目:国家自然科学基金(61873126)

作者简介:黄赞(1998-),男,硕士研究生,主要从事无人机栖落机动飞行控制方法方面的研究。

通信作者:何真(1981-),女,博士,副教授,主要从事飞行控制、非线性控制和智能控制方面的研究。

0 引言

降落是固定翼无人机飞行性能的关键阶段，固定翼飞行器的降落通常需要一定的滑跑距离，才能保证安全着陆。鸟类的降落方式有着飞行器无法达到的优势。文献[1]观察到鸽子可以从空中直接降落到栖木上，降落过程中翅膀和身体存在 40°～90°之间的较大俯仰角。研究者借鉴鸟类的降落方式，提出了无人机栖落机动的概念。无人机在栖落机动过程中，通过大迎角过失速的机动，产生很大的空气阻力，实现快速减速，并最终以较低的速度准确降落在预定落点[2]。栖落机动飞行的过程不再需要滑行跑道，能够拓宽固定翼无人机的应用场合。

文献[3-4]研究了无人机栖落机动的空气动力学模型，文献[5]研究了无人机栖落机动过程的轨迹优化，并设计了变体无人机用于改善栖落机动的性能。文献[6]研究了基于扰动观测器的无人机栖落机动控制方法，并建立了由高度非线性的纵向动力学转化而成的分段线性模型。目前，栖落机动控制方面的研究思路大多是给定参考轨迹，基于近似线性化的模型设计轨迹跟踪控制器[7-8]。这类方法有一个共同的特点，即所设计的控制器依赖于参考轨迹，只在参考轨迹附近有效。而栖落机动中的无人机是一个高度非线性快时变的系统，同时栖落机动需要满足许多约束(尤其是落点处的位置和速度约束)，这使得不同条件下的参考轨迹偏差很大，且不容易计算获得。针对这个问题，本文研究了一种无需给定参考轨迹的栖落机动控制策略设计方法。

强化学习通过与环境之间的交互，更新控制策略。一些专家学者提出了基于强化学习的飞行控制方法，可以根据当前的飞行状态直接得到控制量，具有良好的适应性。文献[9]针对无人机自主飞行到热上升气流的问题，设计了基于强化学习的飞行器导航控制器。文献[10]利用强化学习训练无人机在野外自主导航到热上升气流，学习并验证了局部垂直风加速度和滚转力矩对于导航的重要性。文献[11]将深度强化学习应用到无人机自动特技飞行领域，能够显著地缩短学习时间。但在现实环境中，希望无人机从学习开始就具有良好的在线性能，如果先利用良好的历史数据进行预训练，将会增加深度强化学习在真实环境中的应用场景。文献[12]利用专家数据进行预训练，加快深度 Q 学习的学习过程。文献[13]基于演示数据中的深层确定性策略梯度算法，解决了机器人将柔性对象插入刚性对象的搬运问题。在学习的前期阶段，加入少量的专家演示数据可以帮助智能体的探索和学习。

针对具有高非线性、多约束特性的无人机栖落机动过程，本文基于深度强化学习，提出了一种无需给定参考轨迹的栖落机动控制策略设计方法。本文第一章介绍了固定翼无人机栖落机动的纵向动力学建模；第二章设计了固定翼无人机栖落机动的强化学习模型；第三章设计了能从专家经验数据中学习的基于深度强化学习中近端策略优化(Proximal Policy Optimization，PPO)的栖落机动控制策略设计方法；第四章进行了仿真实验，验证了本文设计方法的有效性；第五章总结了本文的工作。

1 无人机栖落机动动力学建模

本文研究的对象为固定翼无人机，为了简化研究模型，仅针对固定翼无人机栖落机动的纵向动力学建模，假设横向运动对纵向运动方程无影响，方程如下[5]

$$\begin{cases}\dot{V}=(T\cos\alpha-D-mg\sin\mu)/m\\ \dot{\mu}=(T\sin\alpha+L-mg\cos\mu)/(mV)\\ \dot{\alpha}=q-(T\sin\alpha+L-mg\cos\mu)/(mV)\\ \dot{q}=M/I_y\\ \dot{x}=V\cos\mu\\ \dot{h}=V\sin\mu\end{cases}\tag{1}$$

式中，V 为固定翼无人机的飞行速度；μ 为航迹倾斜角；α 为迎角；q 为俯仰角速度；x 为无人机的水平位置；h 为垂直高度；m 为无人机的质量；T 为发动机推力；M 为空气动力矩；I_y 为俯仰转动惯量；L 和 D 分别为无人机所受的升力和阻力。

固定翼无人机纵向运动的空气动力方程如下

$$\begin{cases}L=\dfrac{1}{2}\rho V^2SC_L\\ D=\dfrac{1}{2}\rho V^2SC_D\\ M=\dfrac{1}{2}\rho V^2SC_M\end{cases}\tag{2}$$

式中，ρ 为空气密度；S 为固定翼无人机的空气动力面积；C_L、C_D 和 C_M 分别为升力、阻力和力矩系数，其中 C_L 和 C_D 可由平板模型方法[14-15] 得到与 α 之间的表达式

$$\begin{cases} C_L = 0.8\sin(2\alpha) \\ C_D = 1.4(\sin\alpha)^2 + 0.1 \end{cases} \tag{3}$$

假设固定翼无人机装有全动水平尾翼，能够帮助无人机在低速飞行的状态下获得较大的控制力矩，则空气动力矩系数的表达式为

$$C_M = -\frac{S_e l_e}{S}(0.8\cos\alpha\sin(2\alpha + 2\delta_e) + 1.4\sin\alpha\sin^2(\alpha + \delta_e) + 0.1\sin\alpha) \tag{4}$$

式中，S_e 为升降舵的表面积；l_e 为升降舵空气动力重心到无人机质心的距离；δ_e 为升降舵偏转角。

2　栖落机动的强化学习模型

2.1　基本模型与价值函数

本文采用标准马尔可夫决策过程(Markov Decision Process，MDP)形式对无人机的栖落机动控制过程进行阐述。MDP 由状态 S、动作 A、收益 R、概率分布 P 以及折扣因子 γ 构成。

在每一个时间步 t，无人机与环境进行交互，返回一个观测值 $s_t \in S$，观测值包括无人机的速度 V、航迹倾斜角 μ、迎角 α、俯仰角速率 q、水平位移 x 以及纵向位移 h，在这个观测值的基础上选择并执行动作 $a_t \in A$，动作包括推力 T 以及升降舵偏转角 δ_e。下一时刻，作为执行动作的结果，获得一个数值化的收益 $r_{t+1} \in R$(奖励函数的设计将在 2.2 节中给出)，根据当前无人机的状态和所执行的动作结合概率分布 $P(s_{t+1} \mid s_t, a_t)$，确定下一时刻无人机的状态 s_{t+1}。每一幕轨迹在无人机抵达目标点并成功栖落或达到时间上限值时终止。

在基于策略的强化学习方法中，无人机在每个状态下所采取的动作都遵循策略 π，优化的目标是在策略 π 下的收益累加和

$$J_G(\pi) = E_{\tau\sim p(\tau\mid\pi)}\left[\sum_{k=t+1}^{T}\gamma^{k-t-1} r_k\right] \tag{5}$$

式中，$\tau = \{(s_0, a_0, r_1), (s_1, a_1, r_2)\cdots\}$ 是无人机在策略 π 下的运动轨迹。

采取带参数 θ 的神经网络近似描述策略 π，记为 $\pi(a \mid s, \theta)$。

本文采用策略梯度法学习策略参数 θ，其目标是最大化目标函数 $J(\theta)$

$$J(\theta) = E_t[\log\pi(a_t \mid s_t, \theta) A_t] \tag{6}$$

式中，A_t 为优势函数，是动作价值函数 $Q(s, a)$ 和价值函数 $V(s)$ 的差值

$$A_t = Q(s_t, a_t) - V(s_t) \tag{7}$$

策略参数 θ 的更新近似于 $J(\theta)$ 的梯度上升

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla J(\theta) \tag{8}$$

式中，$\nabla J(\theta)$ 是对目标函数 $J(\theta)$ 梯度的估计；α_θ 是策略参数的学习率。

2.2　奖励函数塑造

在强化学习中，根据控制任务设计适当的奖励函数非常重要。奖励函数选取是否恰当，对学习过程的收敛性以及可行性有着重大影响，它的选取与任务的目标、控制对象所受的约束条件，以及所希望达到的性能指标密切相关。根据无人机无跑道降落的需求，要求栖落机动的无人机在规定的时间内到达预设栖落点完成栖落机动，并希望终点时刻的速度、俯仰角以及与目标值的误差越小越好。因此，无人机栖落机动强化学习算法的奖励函数如下所示：

1)定义 r_d^t 为无人机在 t 时刻与栖落点的归一化距离

$$r_d^t = |x_t - x_f| / x_{max} \tag{9}$$

式中，x_f 为终点位置；x_{max} 为允许的最大水平位移。无人机栖落机动的最终目的是降落到预设地点，算法鼓励无人机越接近预设栖落点越好。

2)定义 r_s^t 为无人机在 t 时刻的归一化速度

$$r_s^t = |v_t - v_f| / v_{max} \tag{10}$$

式中，v_f 为终点速度；v_{max} 为过程中允许的最大速度。无人机栖落机动在终点时刻时，速度越小越好，算法鼓励无人机接近预设降落速度。

3)将上述与无人机位置和速度相关的奖励综合起来，定义综合奖励项 r_p^t

$$r_p^t = 1 - (0.3\sqrt{r_d^t} + \sqrt{r_s^t}) \tag{11}$$

即无人机离终点位置的距离和终点时刻的速度变小，算法鼓励无人机朝着栖落点接近。

4) r_b^t 是无人机栖落机动过程中的约束量，惩罚无人机在栖落机动过程中超出约束范围

$$\begin{aligned} r_b^t = t > t_f \| v_t > v_{max} \| \mu_t > \mu_{max} \| \mu_t < -\mu_{max} \| \\ \alpha_t > \alpha_{max} \| \alpha_t < -\alpha_{max} \| q_t > q_{max} \| \\ q_t < -q_{max} \| x_t > x_{max} \| h_t > h_{max} \end{aligned} \tag{12}$$

式中，t_f 为最大完成时间；μ_{max} 为最大航迹倾斜角；α_{max} 为最大迎角；q_{max} 为最大俯仰角速率；h_{max} 为最大飞行高度；符号‖ 表示逻辑运算符或。

5) r_c^t 是无人机栖落终点的约束量，奖励无人机完成栖落机动，即无人机的终点位置和速度在允许的误差范围内

$$r_c^t = |v_t - v_f| \leqslant \sigma_v \&\& |x_t - x_f| \leqslant \sigma_x \&\& |h_t - h_f| \leqslant \sigma_h \quad (13)$$

式中，σ_v、σ_x 和 σ_h 分别是终点速度、终点位置和终点高度的允许误差；h_f 为终点高度；符号 && 表示逻辑运算符与。

6）r_a^t 是 t 时刻与迎角相关的奖励项，针对无人机栖落飞行是大迎角机动的特性，在奖励函数中加入迎角鼓励项，有利于加快学习的前期探索

$$r_a^t = |\alpha_t|^2 \quad (14)$$

根据无人机栖落机动的目标以及性能要求，对上述奖励项设置了不同的比例因子。特别地，为了让无人机在栖落机动的过程中始终满足过程约束，以及为了保证在落点处满足终点约束，对奖励项 r_b^t 和 r_c^t 设置了较大的比例系数。具体地，奖励函数设计为

$$r_t = 10(r_p^{t+1} - r_p^t) + 500r_c^t - 100r_b^t + 0.3r_a^t \quad (15)$$

3 栖落机动的控制策略优化算法

3.1 模仿强化学习框架

本文利用模仿强化学习（Imitation Reinforcement Learning，IRL）对无人机的栖落机动进行轨迹优化。该学习通过从人类专家的演示数据中学习，并促进深度强化学习。对无人机栖落机动的轨迹优化设计分为两个阶段，一个是模仿学习阶段，另一个是使用 PPO 算法的强化学习阶段。整体架构如图 1 所示。

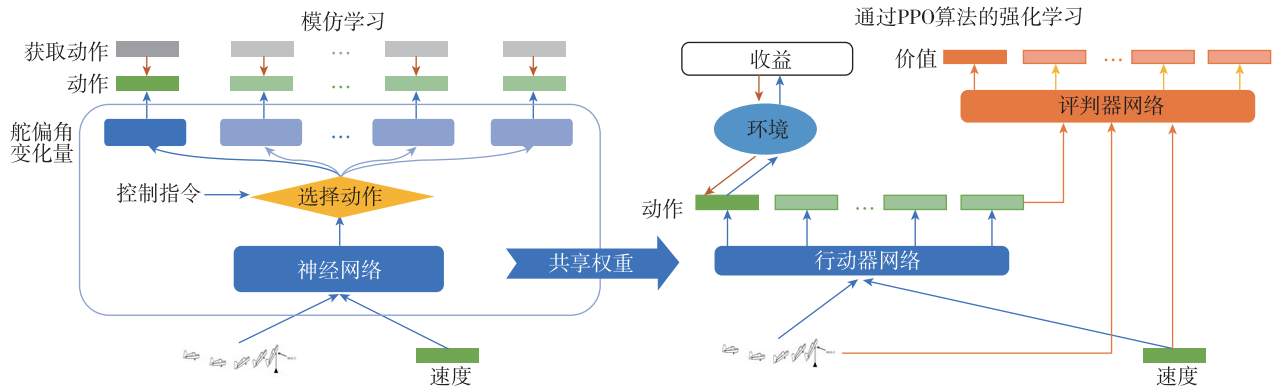


图 1 从模仿中进行强化学习的框架图

Fig. 1 Architecture of reinforcement learning from demonstration

IRL 基于 PPO 算法学习实现，但传统的 PPO 算法会在大的动作空间中进行过多的失败探索，往往会陷入局部最优。IRL 则通过为行动器网络的动作空间搜索提供更好的探索方向，以解决陷入局部最优的问题[16]。

3.2 模仿学习

在第一阶段，本文使用模仿学习的方法对网络进行预训练，并在训练结束后，将学习到的神经网络权重共享到下一阶段的行动器网络中。该阶段首先使用广义伪谱 GPOPS（General Pseudospectral Optimization Software）工具包生成学习所需的轨迹，得到 N 个参考轨迹序列 τ。利用生成的轨迹通过模仿学习来训练策略网络，以模仿专家数据。策略 π 的网络结构如图 2 所示。策略网络中的输入为无人机栖落机动的状态量，即飞行时的速度、航迹倾斜角、迎角、俯仰角速率、水平位移以及纵向位移，输出为升降舵偏转角的变化量。

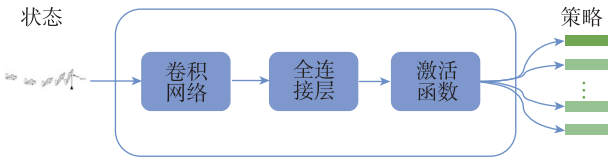


图 2 策略网络的神经网络结构

Fig. 2 Neural network architecture of policy network

通过模仿学习训练无人机栖落机动飞行策略时，需要大量的演示数据，但过多相似的数据反而会影响训练的效率。参考优先经验回放，对训练的输入样本进行非均匀抽样，这样既能让收敛更快，也能让收敛时的平均回报更高。先对所有的样本根据重要度降序排列，每个样本抽取的概率为

$$p_j \propto 1/\mathrm{rank}(j) \quad (16)$$

其中，$\mathrm{rank}(j)$ 为样本的序号。

由于演示的数据必然只会覆盖状态空间的一部分，也没有采取所有可能的动作，因此有很多状态动作从未被采取过，所以对模仿学习时的损失函

数增加了一个监督损失

$$L_1(Q)=\max_{a\in A}[Q(s,a)+l(a_e,a)]-Q(s,a_e) \quad (17)$$

式中，a_e 为演示数据中所采取的动作，当 $a=a_e$ 时，$l(a_e,a)=0$，否则为正。增加的监督损失能够诱导无人机智能体的行为靠近演示数据。

3.3　策略梯度优化

在第二阶段，本文使用 PPO 算法训练策略网络，以提高无人机栖落机动策略的泛化能力。该策略通过与仿真环境的互动获得收益，并使用反馈回来的奖励函数优化策略网络。在从第一阶段得到的策略网络的基础上学习，能够提高该阶段的强化学习样本利用率以及学习效率，从而获得更加通用的策略。由于无人机栖落飞行系统需要连续的预测动作，因此采用行动器-评判器的框架更新策略。用带参数 ω 的神经网络近似价值函数，记为 $V(s,\omega)$。则策略参数和价值函数参数的更新公式如下

$$\theta_{t+1}=\theta_t+\alpha_\theta\delta_t\frac{\nabla\pi(A_t\mid s_t,\theta_t)}{\pi(A_t\mid s_t,\theta_t)} \quad (18)$$

$$\omega_{t+1}=\omega_t+\alpha_\omega\delta_t\nabla V(s_t,\omega) \quad (19)$$

式中，δ_t 为 t 时刻的回报和以价值函数作为基准线的差值

$$\delta_t=r_{t+1}+\gamma V(s_{t+1},\omega)-V(s_t,\omega) \quad (20)$$

采用 PPO 算法优化此策略网络，可以实现多次的小批量更新，并且能够更好地应对复杂的环境情况。将经过模仿学习预先训练好的 θ_1 加载到 θ 中，代替随机初始化策略网络参数，有助于减少 PPO 在早期阶段探索所耗费的时间。

为了更好地利用更新策略时所得到的轨迹数据，可以使用重要性采样构建目标函数[17]

$$J(\theta)=\mathrm{E}\left[\frac{\pi(a_t\mid s_t,\theta)}{\pi(a_t\mid s_t,\theta_{\mathrm{old}})}A_t\right] \quad (21)$$

式中，θ_{old} 是更新之前的策略参数。

为了防止前后更新的策略差异较大，PPO 算法通过使用超参数裁剪目标函数的形式来解决这类问题。

$$J^{\mathrm{clip}}(\theta)=\mathrm{E}[\min(p_t(\theta)A_t,\mathrm{clip}(p_t(\theta),1-\varepsilon,1+\varepsilon)A_t)] \quad (22)$$

式中，$p_t(\theta)$ 为策略的概率比 $\dfrac{\pi(a_t\mid s_t,\theta)}{\pi(a_t\mid s_t,\theta_{\mathrm{old}})}$，$\mathrm{clip}(\cdot)$ 将 $p_t(\theta)$ 的变化范围限制在 $[1-\varepsilon,1+\varepsilon]$。通过选取未裁剪与裁剪后的目标函数的较小值，使得策略的更新变得更加稳定[18]。由于本文中阐述的无人机栖落机动为分幕式任务，研究采用广义优势估计来描述优势函数

$$A_t=\delta_t+\gamma\lambda\delta_{t+1}+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1} \quad (23)$$

式中，λ 为广义优势估计的平滑系数。

4　仿真结果与分析

4.1　仿真实验参数

为验证提出的轨迹规划算法的有效性，本节进行了仿真实验研究。仿真中所采用的无人机动力学方程如式(1)，气动参数如式(2)～式(4)，无人机的物理参数如表 1 所示。初始时间为 $t_0=0$s，离散化采样时间为 $\Delta t=0.01$s，栖落机动轨迹优化最大完成时间为 $t_f=2$s。

表 1　无人机的各项物理参数

Tab. 1　Physical parameters of UAV

参数	参数值
m/kg	0.8
I_y/(kg·m^2)	0.1
S_e/m^2	0.054
S/m^2	0.25
l_e/m	0.235
ρ/(kg/m^3)	1.225
g/(m/s^2)	9.8

在仿真过程中，无人机的理想初始状态为 $s_0=[10,0,0.2544,0,0,0]$，初始控制量 $a_0=[3.7698,-0.15]$。为了检验算法在初始状态不确定的情况下的轨迹优化效果，设定初始情况下的水平位置与理想状态的偏差范围在 0.5m 以内。无人机在栖落机动飞行过程中以及终点位置的约束参数如表 2 所示，并且希望终点的速度、俯仰角、水平以及纵向位移的偏差越小越好。IRL 算法的超参数设置如表 3 所示。

4.2　基于 IRL 的仿真实验

IRL 在训练无人机过程中的奖励变化曲线如图 3 所示。其中，蓝色曲线为没有经过模仿学习，直接用策略梯度优化算法的奖励变化曲线；红色曲线为经过模仿学习再用策略梯度优化(即 IRL)的奖励变化曲线；黄色曲线则为不包含迎角奖励模块的 IRL 奖励变化曲线。

表 2 状态以及控制量约束参数

Tab. 2 Constraints of states and control variable parameters

参数	参数值
v_f/(m/s)	3.5
v_{max}/(m/s)	25
μ_{max}/rad	$\pi/4$
α_{max}/rad	$\pi/2$
q_{max}/(rad/s)	3.5
x_f/m	14.9
x_{max}/m	15
h_f/m	1.6
h_{max}/m	5
σ_v/(m/s)	0.5
σ_x/m	0.1
σ_h/m	0.1

表 3 算法超参数

Tab. 3 Algorithm hyperparameters

超参数	参数值
α_θ	0.0001
α_ω	0.0001
ε	0.2
γ	0.997
λ	0.95

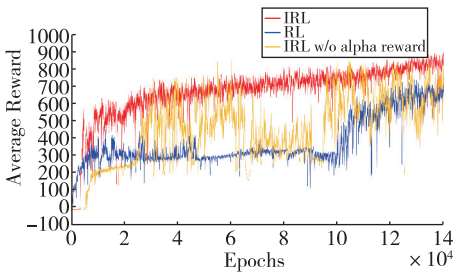


图 3 训练时的奖励曲线

Fig. 3 Reward curves during training

由图 3 可知，RL 的方法需要在早期对状态动作空间进行大量的探索，而 IRL 则利用专家演示数据对神经网络进行预训练，并与行动器网络共享网络权重，大大加快了学习的进程，并且 IRL 在训练到 8 万幕时，平均的奖励函数值就已达到了预期的目标。在 IRL 方法的奖励函数中加入迎角奖励模块，则能够进一步加快早期学习的过程，并能够减少无人机在中期对不确定状态的试探。

在采用 IRL 对无人机的栖落飞行训练完后，进行仿真测试，测试集为 1000 幕。在训练达到 14 万幕时，仿真测试成功率达到可接受的范围。在训练幕数达到 20 万幕时，仿真测试的成功率能够达到 97.5%。图 4 给出了 100 幕不同初始条件下成功和失败案例的降落点分布图。图 5 则分别给出了不同初始条件下无人机在栖落机动飞行过程中的不同状态量变化曲线，且无人机栖落机动飞行过程中的状态量满足设定的过程约束和终点约束。

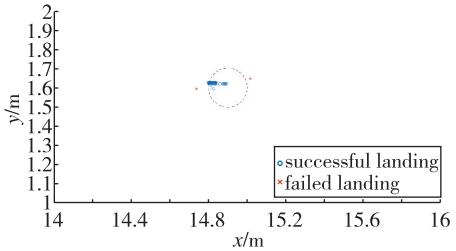
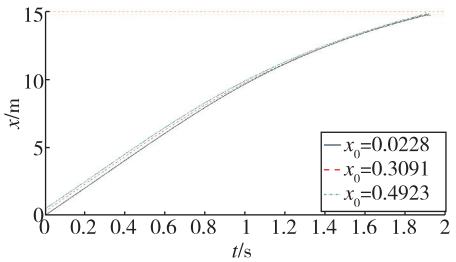
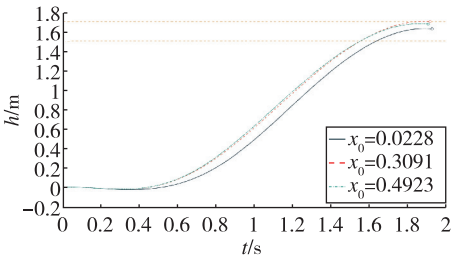


图 4 降落点分布图

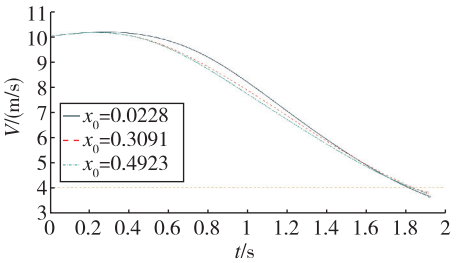
Fig. 4 The distribution of landing



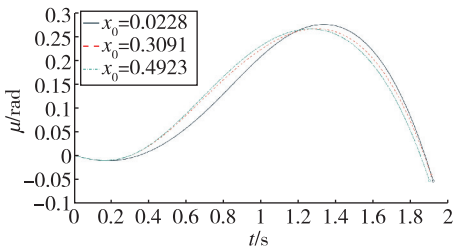
(a)水平位置变化曲线



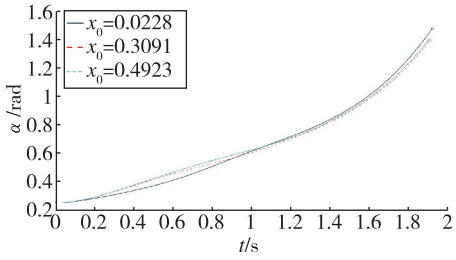
(b)高度变化曲线



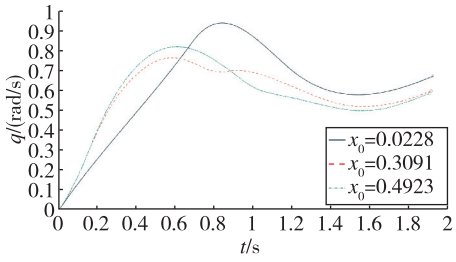
(c)速度变化曲线



(d)航迹倾斜角变化曲线



(e)迎角变化曲线



(f)俯仰角速率变化曲线

图 5　状态量变化曲线

Fig. 5　The curves of control variables

5　结论

1)基于 IRL 的控制策略设计方法,只在模仿学习阶段需要专家经验轨迹(可以离线生成),在强化学习阶段和在线运行中都不需要参考轨迹。利用该方法得到的控制器不依赖于参考轨迹,可以根据当前的飞行状态直接得到控制量,具有良好的适应性。

2)栖落机动轨迹规划的仿真结果表明,采用 IRL 设计的控制器不仅能够实现无人机的栖落机动,同时经过离线学习后,能够根据不同的状态实时在线调整轨迹,保证轨迹满足预先给定的要求。

本文针对固定翼无人机栖落运动的非线性问题,在深度强化学习的基础上考虑了模仿控制策略。后续工作中,将会考虑根据自身条件做出决策的无人机个体性的学习方案。今后也将进一步研究在各种外部风扰动情况下无人机的栖落机动飞行,以增强面对各种复杂环境的适应能力。

参考文献

[1] Green P R, Cheng P. Variation in kinematics and dynamics of the landing flights of pigeons on a novel perch[J]. Journal of Experimental Biology, 1998, 201(24): 3309-3316.

[2] Hoburg W, Tedrake R. System identification of post stall aerodynamics for UAV perching[C]// Proceedings of AIAA Infotech@Aerospace Conference, 2009.

[3] Puopolo M G, Reynolds R, Jacob J D. Comparison of three aerodynamic models used in simulation of a high angle of attack UAV perching maneuver[C]// Proceedings of 51st AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, 2013.

[4] Rao D M K K, Tang H, Go T H. A parametric study of fixed-wing aircraft perching maneuvers[J]. Aerospace Science and Technology, 2015, 42: 459-469.

[5] 袁亮, 何真, 王月. 变体无人机栖落机动建模与轨迹优化[J]. 南京航空航天大学学报, 2018, 50(2): 266-275.
Yuan Liang, He Zhen, Wang Yue. Modeling and trajectory optimization of perching maneuvers morphing UAV [J]. Journal of Nanjing University of Aeronautics & Astronautics, 2018, 50(2): 266-275(in Chinese).

[6] He Z, Li D, Lu Y. Disturbance compensation based piecewise linear control design for perching maneuvers [J]. IEEE Transactions on Aerospace and Electronic Systems, 2019, 55(1): 192-204.

[7] 王无天, 何真, 岳珵. 飞行器栖落机动的轨迹跟踪控制及吸引域优化计算[J]. 北京航空航天大学学报, 2021, 47(2): 414-423.
Wang Wutian, He Zhen, Yue Cheng. Trajectory tracking control and optimal computation of attraction domain for aircraft in perching maneuvers[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47 (2): 414-423(in Chinese).

[8] 万慧雯, 何真, 曹瑞. 无人机栖落机动的一种离线鲁棒预测控制算法[J]. 南京航空航天大学学报, 2019, 51(6): 785-794.
Wan Huiwen, He Zhen, Cao Rui. An off-line robust predictive control algorithm for UAV in perching maneuver[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2019, 51(6): 785-794(in Chinese).

[9] Woodbury T, Dunn C, Valasek J. Autonomous soar-

ing using reinforcement learning for trajectory generation [C]//Proceedings of 52nd AIAA Aerospace Sciences Meeting. Maryland, 2014.

[10] Reddy G, Wong-Ng J, Celani A, et al. Glider soaring via reinforcement learning in the field[J]. Nature, 2018, 562: 236-239.

[11] Clarke S G, Hwang I. Deep reinforcement learning control for aerobatic maneuvering of agile fixed-wing aircraft[C]// Proceedings of AIAA Scitech 2020 Forum. FL, 2020.

[12] Hester T, Vecerik M, Pietquin O, et al. Deep Q-learning from demonstrations[C]// Proceedings of AAAI Conference on Artificial Intelligence, 2018.

[13] Vecerik M, Hester T, Scholz J, et al. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards[J]. arXiv preprint arXiv: 1707.08817, 2017.

[14] Puopolo M G, Reynolds R, Jacob J D. Comparison of three aerodynamic models used in simulation of a high angle of attack UAV perching maneuver[C]// Proceedings of 51st AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, 2013.

[15] Rao D M K K, Go T H. Optimization, stability analysis, and trajectory tracking of perching maneuvers [J]. Journal of Guidance Control and Dynamics, 2014, 37(3): 879-888.

[16] Liang X, Wang T, Yang L, et al. Cirl: controllable imitative reinforcement learning for vision-based self-driving[C]// Proceedings of European Conference on Computer Vision (ECCV), 2018: 584-599.

[17] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv: 1707.06347, 2017.

[18] Chen Y, Ma L. Rocket powered landing guidance using proximal policy optimization[C]// Proceedings of 2019 4th International Conference on Automation, Control and Robotics Engineering, 2019: 1-6.

（编辑：孟彬）