控制与决策
Control and Decision

基于强化学习的固定翼飞机姿态控制方法

付宇鹏, 邓向阳, 何明, 朱子强, 张立民

引用本文:
付宇鹏,邓向阳,何明,朱子强,张立民. 基于强化学习的固定翼飞机姿态控制方法[J]. 控制与决策, 2023, 38(9): 2505−2510.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2230

基于强化学习的固定翼飞机姿态控制方法

付宇鹏[1], 邓向阳[1†], 何 明[2], 朱子强[1], 张立民[1]

(1. 海军航空大学 航空作战勤务学院，山东 烟台 264001；2. 陆军工程大学 指挥控制工程学院，南京 210007)

摘 要: 研究基于强化学习的飞机姿态控制方法,控制器输入为飞机纵向和横向状态变量以及姿态误差,输出为升降舵和副翼偏转角度指令,实现不同初始条件下飞机姿态角快速响应,同时避免使用传统PID控制器和不同飞行状态下的参数调节. 根据飞机姿态变换特性,通过设置分立的神经网络模型提高算法收敛效率. 为贴近实际的固定翼飞机控制,仿真基于JSBSim的F-16飞机空气动力学模型,利用OpenAI gym搭建强化学习仿真环境,以任意角速度、角度和空速作为初始条件,对姿态控制器中的动作网络和评价网络进行训练. 仿真结果表明,基于强化学习的姿态控制器响应速度快,动态误差小,并能避免大过载等边界条件.

关键词: 强化学习；近端策略优化算法；姿态控制；固定翼；PID；JSBSim

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2230

引用格式: 付宇鹏,邓向阳,何明,等. 基于强化学习的固定翼飞机姿态控制方法[J]. 控制与决策, 2023, 38(9): 2505-2510.

Reinforcement learning based attitude controller design

FU Yu-peng[1], *DENG Xiang-yang*[1†], *HE Ming*[2], *ZHU Zi-qiang*[1], *ZHANG Li-min*[1]

(1. School of Aviation Support, Naval Aeronautical University，Yantai 264001，China；2. Command and Control Engineering Colledge, People's Liberation Army Engineering University，Nanjing 210007，China)

Abstract: This article presents an attitude controller based on reinforcement learning (RL). The inputs of the actor network are states of attitude angle, angular rates etc, where the output is the angle control command of elevator and aileron, achieving the rapid response of the attitude angle with variable initial conditions, avoiding the application of the conventional PID controller and the parameter adjustment. According to the states transfer characteristics, by setting the splitting neural network model, the efficiency of algorithms is improved. In order to be close to the actual fixed-wing aircraft model, the simulation is based on the JSBSim F-16 aerodynamic model, using the OpenAI gym to build the simulation environment for reinforcement learning. With arbitrary angular speed, angle, and airspeed as initial conditions, the actor and critic networks are trained. The simulation results show that the RL based attitude controller has faster response and less dynamic error compared with the conventional PID controller.

Keywords: reinforcement learning；PPO；attitude control；fixed-wing；PID；JSBSim

0 引 言

随着强化学习算法和算力不断发展,智能体通过强化学习有能力完成如飞行控制、航迹追踪、任务规划、机动决策、博弈对抗或协同等复杂任务. 在这些任务系统中,无论采用端到端还是分层网络架构,首先要保证飞机模型的稳定控制,使飞机能够随时准确达到所需的正常过载迎角或姿态[1-2]. 传统控制器通常采用经典的proportion integration differentiation (PID)控制、模糊控制等方法,具有良好的工程适应性和控制效果[3]. 但无论是经典控制理论还是现代控制理论,对于六自由度(6-DoF)飞机的多输入变量-多输出变量的复杂模型[4],都需要对被控模型进行数学建模并设计其参数,理论复杂,计算量大,这无疑会对非飞行专业的研究人员造成一定的负担. 此外,针对不同任务或任务过程的控制器参数优化较为困难. 文献[5-7]提出了多变量PID神经元的控制方法,能够实现多输入多输出变量的解耦,但该方法很难适用于复杂动力学模型控制系统,因此提出至今极少实际应用于飞行控制系统.

目前,强化学习在飞机控制领域取得了一定进展. 针对不同模型和不同状态下,控制器结构设计、参数选择和优化,文献[8-9]利用神经网络实现了自适

收稿日期: 2021-12-24；**录用日期:** 2022-04-27.
基金项目: 泰山学者工程专项基金项目(ts201511020).
†通讯作者. E-mail: xiangyang.deng@qq.com.

应控制器参数调节. 文献[10]利用deep deterministic policy gradient(DDPG)算法训练得到的控制器能够实现对旋翼飞机位置控制. 文献[1]利用经验池Actor-Critic算法,通过选择原子动作实现了固定翼飞机轨迹跟踪. 但目前基于强化学习设计姿态控制器难度仍然较大,普遍建立在传统控制器基础上或采用较简单的气动模型[11].

本文提出一种基于强化学习的姿态控制器设计方法,结合姿态转移的动态特点搭建双通道独立动作网络,并选择奖励函数实现控制器设计,快速完成控制器设计. 控制器由俯仰通道和滚转通道的动作网络构成,根据不同飞机状态输出最优响应的升降舵和副翼偏转指令,从而实现姿态快速变换. 通过仿真验证了面对复杂受控模型,强化学习算法能够实现飞机稳定控制,有利于开展下一步任务层次的研究工作.

1 姿态控制器设计

飞机的姿态主要由发动机推力、升降舵、副翼、方向舵控制. 当推力和各操纵面变化时,模型根据对应的气动参数改变飞机合力和合力矩. 根据飞机气动方程[12-14],不同轴向的力或力矩同时受到空速、攻角、操作面偏转角的影响,因此不可避免地存在耦合,但解耦算法复杂度高,计算量大. 为了简化控制器设计难度,传统飞行控制器中一般认为纵向(俯仰)和横向(滚转)通道相对耦合较小,大都采用单通道串级PID控制.

1.1 基于PID的控制器设计

基于串级PID的传统姿态控制器原理如图1所示,包括纵向俯仰通道和横向滚转通道.

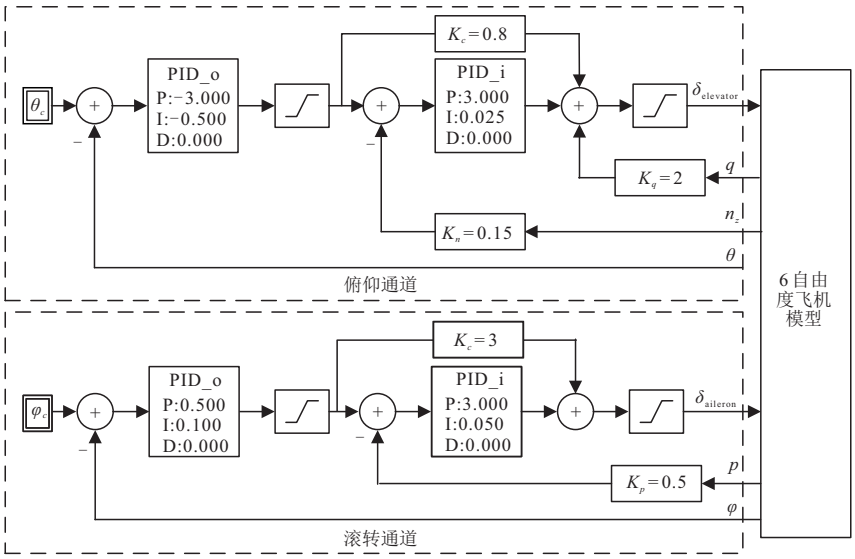


图1 串级PID姿态控制器

纵向控制器内环为控制增稳系统,包括角速度、过载反馈环路和控制指令前馈通路[8],其中q为俯仰角速度,n_z为法向过载. 角速率环路可采用简单的比例控制,通过反馈增加系统阻尼,从而提高稳定性. 法向过载与角运动密切相关,具有短周期特性,响应速度仅次于角速度反馈,因此基于过载稳定的控制方法具有更强的稳定性和鲁棒性. 过载指令经过前馈增益k_c输入给角速度环路以提高飞机姿态变化率. 外环为俯仰角反馈环路,其中θ为俯仰角. 横向控制器内环同样为角速度反馈和控制指令前馈,外环为滚转角反馈环路,其中ϕ为滚转角,p为滚转角速度.

传统PID控制器仅以单通道状态作为反馈,并未考虑不同姿态角、空速等带来的合力、合力矩的变化,因此在复杂姿态调整过程中可能存在响应滞后、过冲大的缺点. 为了增强控制器的鲁棒性和响应特性,控制器参数需要根据当前状态调整,无疑增加了设计人员的工作量.

1.2 基于强化学习的控制器设计

为了实现姿态角快速响应以及变量解耦的目的,提出基于强化学习的姿态控制器,结构如图2所示. 对于强化学习智能体而言,在进行数据采样时,环境(environment)即为包括飞机空气动力学模型、升降舵和副翼偏转角速度、角度限制在内的响应模块,这些模块在JSBSim平台中可直接复用其控制组件,有助于减少决策算法层研究人员的工作量.

在控制器网络结构设计上,文献[11]采用多输入多输出网络,存在两通道参数相互影响、奖励函数设置灵活性较差的问题,面对复杂模型网络不易收敛. 考虑到固定翼飞机气动力特性,俯仰通道和滚转通道影响因素和响应速度不同,本设计参考传统PID控

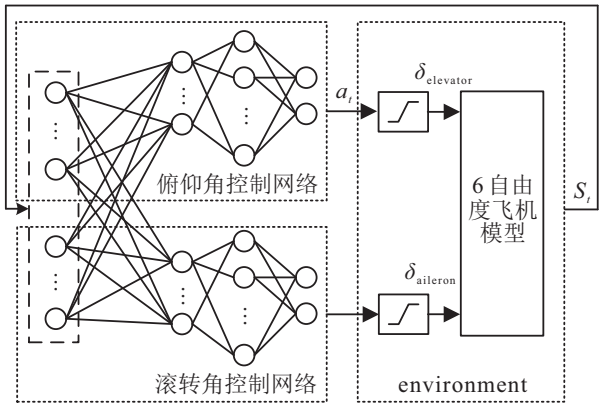


图 2　基于强化学习的姿态控制器

制器的实现方法,控制器同样将动作网络分为俯仰角和滚转角控制网络. 不同之处在于,各通道输出指令本质上是通过气动力和气动力矩反映在各通道状态上,从而影响气动模型的状态转移函数,造成通道间的耦合,而不同通道的动作网络通过共享状态变量,经过一层感知机处理,实现对当前状态特征提取,再经过动作网络实现任意姿态下响应曲线的优化.

飞机姿态的状态转移过程可以认为是如下马尔科夫决策过程(MDP),即环境当前状态转移到下一状态的概率仅与当前状态和动作有关:

$$P(s_t+1)=E(S_{t+1}=s_{t+1}|S_t=s_t,A_t=a_t),\quad(1)$$

其中 $s_t=(s_t^p,s_t^q)$、$a_t=(a_t^p,a_t^q)$ 分别代表俯仰和滚转通道状态和动作的集合. 在状态转移时智能体将获得环境反馈的奖励 r_t,本设计将其表示为

$$r_t=(r_t^p(s_t,a_t^p),r_t^q(s_t,a_t^q)),\quad(2)$$

其为两通道奖励之和. 俯仰动作和滚转动作满足各自的策略网络,即 $a_t^p\sim\mu(s_t),a_t^q\sim\nu(s_t)$. 回合期望回报表示为

$$\begin{aligned}R_{\mu,\nu}=&E_{a^p\sim\mu(\cdot|s),a^q\sim\nu(\cdot|s)}\Big[\sum_{t=0}^{\mathrm{T}}\gamma^t r_t(s_t,a_t^p,a_t^q)\Big]=\\&E_{a^p\sim\mu(\cdot|s)}\Big[\sum_{t=0}^{T}\gamma^t r_t^p(s_t,a_t^p)\Big]+\\&E_{a^q\sim\nu(\cdot|s)}\Big[\sum_{t=0}^{\mathrm{T}}\gamma^t r_t^q(s_t,a_t^q)\Big].\end{aligned}\quad(3)$$

每个通道所获得的最大回报为

$$R^{p^*}=\max_{\mu}(\max_{\nu}R_{\mu,\nu}),\ R^{q^*}=\max_{\nu}(\max_{\mu}R_{\mu,\nu}),\quad(4)$$

即目标策略优化期望达到最大回报,此时输出动作为

$$\begin{aligned}a_t^p&\in\arg\max_{\mu}(\max_{\nu}R_{\mu,\nu}),\\a_t^q&\in\arg\max_{\nu}(\max_{\mu}R_{\mu,\nu}).\end{aligned}\quad(5)$$

观察式(4)和(5),各个通道相互作为对手,目标是得到自身最大奖励,因此各通道动作网络训练过程本质上是一个对抗过程. 但相比于鲁棒对抗或博弈对抗强化学习[15],区别在于本文相互协作提高奖励,因此在网络训练过程中不应交替训练对抗网络,而应同时更新网络参数.

2　算法设计与实现

近端策略优化算法(proximal policy optimization, PPO)是2017年提出的强化学习算法,因其易用性和良好的性能成为OpenAI默认的强化学习算法以及目前各新型算法的对比基准[16-18]. 本文采用PPO算法作为强化学习网络训练算法.

PPO算法是一种Actor-Critic方法,Actor网络输出动作,Critic网络输出价值函数 $V(s_t)$,于是优势函数[19]定义为

$$\hat{A}_t=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1},\quad(6)$$

$$\delta_t=r_t+\gamma V(s_{t+1})-V(s_t).\quad(7)$$

其中: δ_t 为TD-error; $\hat{A}_t$ 为时刻 t 的优势函数的估计,一般采用目前应用较多的广义优势估计(generalized advantage estimator, GAE)方法; γ 和 λ 为GAE(γ,λ)函数的两个重要参数, γ 决定了价值函数的最大值, λ 用来平衡方差和偏差.

PPO算法中梯度通过微分函数得到,设置目标函数 $\mathcal{L}(\theta)$ 为

$$\mathcal{L}(\theta)=\hat{E}_t[\log\pi_\theta(a_t|s_t)\hat{A}_t].\quad(8)$$

其中 π_θ 是以 θ 为参数的随机策略网络. 为了提高训练效果,TRPO算法[17]提出了目标函数的概念,即

$$\mathcal{L}^{\mathrm{CPI}}(\theta)=\hat{E}_t\Big[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta\mathrm{old}}(a_t|s_t)}\hat{A}_t\Big].\quad(9)$$

其中 $\mathcal{L}^{\mathrm{CPI}}(\theta)$ 为当前网络与旧网络输出策略概率比.

PPO算法对目标函数 $\mathcal{L}^{\mathrm{CPI}}(\theta)$ 的更新进行幅度限制,剪裁概率比,从而降低目标函数的波动,即

$$\begin{aligned}&\mathcal{L}^{\mathrm{CLIP}}(\theta)=\\&\hat{E}_t[\min(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta),1+\epsilon,1-\epsilon)\hat{A}_t)].\end{aligned}\quad(10)$$

基于此,PPO算法流程如下所示.

---

for iteration $=1,2,\ldots,M$ do
　for actor $=1,2,\ldots,N$ do
　　环境中运行策略 $\pi_{\theta\mathrm{old}}$ T 步
　　计算优势函数 $\hat{A}_1,\ldots,\hat{A}_T$
　end for
　利用mini-batch数据计算目标函数 $\mathcal{L}(\theta)$
　更新 $\theta_{\mathrm{old}}\leftarrow\theta$
end for

为了避免传统策略梯度算法中采样数据利用

率低的缺点,PPO算法中智能体采用旧策略$\pi_{\theta\text{old}}$与环境交互获得训练数据并计算优势函数$\hat{A}_t$,利用mini-batch多次更新目标函数$\mathcal{L}(\theta)$,完成Actor网络和Critic网络参数θ的更新.

图3给出了以俯仰通道为例的网络训练框图,根据算法流程,Actor网络与环境进行多次交互,根据当前状态s_t生成动作a_t;环境根据动作a_t响应产生下一时刻状态s_{t+1}和当前奖励r_t,该四元组(s_t, a_t, r_t, s_{t+1})存放在mini-batch中. 当mini-batch中数据量达到尺寸要求时,批量计算K次优势函数$\hat{A}_t$并更新网络参数(图中虚线). 训练采用分布式仿真,N个智能体同时探索状态空间,有效利用计算资源降低训练时间. 模型中,动作网络和评价网络输入状态为

$$s = \{\Delta\theta, \epsilon_p \Sigma_{\Delta\theta}, \alpha, p, n_z, \Delta\phi, \epsilon_q \Sigma_{\Delta\phi}, q, v_k\}. \quad (11)$$

其中:俯仰通道状态维度为6,包括俯仰角误差$\Delta\theta = |\theta - \theta_c|$、累积误差$\Sigma_{\Delta\theta}$、攻角$\alpha$、俯仰角速度$p$、过载$n_z$;累计误差$\Sigma_{\Delta\theta}$作为输入应进行比例$\epsilon$缩放并限幅,避免累计误差过大影响学习效率;滚转通道状态向量维度为4,包括滚转角误差$\Delta\phi = |\phi - \phi_c|$、累积误差$\Sigma_{\Delta\phi}$、滚转角速度$q$、空速$v_k$;俯仰通道和滚转通道动作网络输出维度均为2,即$a_t^p = \{\mu_p, \sigma_p\}$,$a_t^q = \{\mu_q, \sigma_q\}$,分别为升降舵和副翼控制指令的均值和方差,服从高斯分布,当网络部署时仅输出均值.

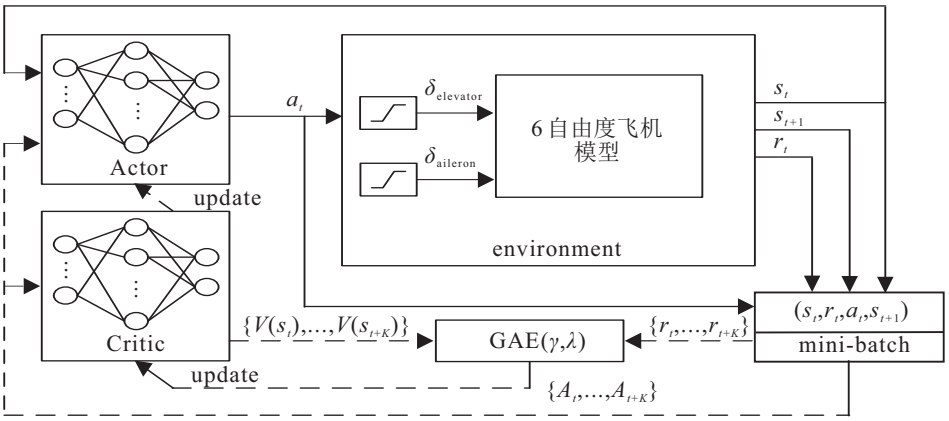


图3 基于强化学习的姿态控制器

在不同姿态角误差条件下,通过设置合理的奖励函数能够引导智能体到达期望的目标姿态角,例如可以要求滚转角和俯仰角满足最短时间或按照特定的响应曲线进行. 本文模仿PID控制器的响应函数,将两通道的奖励函数分别设置为

$$r_t^p = \begin{cases} r_t^p - \eta_n, \ n_z > 4 \text{ or } n_z < -8; \\ -\left(|\Delta\theta| + \dfrac{\epsilon_p T_s}{T_i^p}|\Sigma_{\Delta\theta}|\right) - \eta_q\left(|\Delta\phi| + \dfrac{\epsilon_q T_s}{T_i^q}|\Sigma_{\Delta\phi}|\right), \\ \quad \text{otherwise.} \end{cases} \quad (12)$$

$$r_t^q = -\left(|\Delta\phi| + \frac{\epsilon_q T_s}{T_i^q}|\Sigma_{\Delta\theta}|\right) - \eta_p\left(|\Delta\theta| + \frac{\epsilon_p T_s}{T_i^p}|\Sigma_{\Delta\theta}|\right). \quad (13)$$

对比PID控制器实际物理意义,T_i^p和T_i^q可以看作系统的积分时间常数,T_s表示采样时间. 奖励函数中各变量分别与输入状态s中角度误差和角度累积误差对应,因此网络模型训练后可通过微调累积误差缩放比例ϵ优化系统响应特性或适应不同采样时间,训练时设置$\epsilon = 1/T_s$. 根据JSBSim模型设计经验,俯仰通道积分时间通常小于5 s,滚转通道积分时间通常小于2 s. 结合传统PID控制器参数整定方法,为了保证控制器稳定,将T_i^p和T_i^q分别设置为8 s和5 s;η_p和η_q表示俯仰通道和滚转通道的耦合系数,约束姿态角在二维空间中拟合出最优响应曲线,本文中为0.5;特别地,考虑到实际飞机法向过载n_z的限制条件,当$n_z > 4$或$n_z < -8$时,奖励函数引入惩罚项η_z,使智能体避免进入大过载区域.

算法中网络结构和超参数设计如表1所示,两通道A-C网络结构相同,均采用表1中所示的全连接结构,隐藏层激活函数为ReLu函数,动作网络均值激活函数为tanh函数,方差激活函数为softmax函数,Loss function均采用Adam方法更新梯度[20].

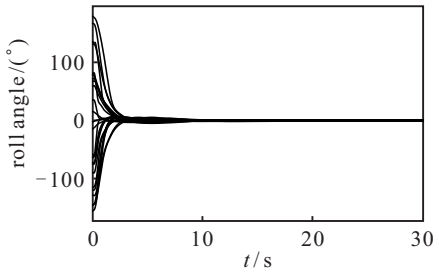
表1 算法参数设置

名称	值	名称	值
Actor结构	9×128×128×2	Critic结构	9×128×128×1
Actor学习率	1e-5	Critic学习率	1e-5
每回合最大时长	30 s	mini batch size	64
γ	0.9	Actor更新次数	15
λ	0.95	Critic更新次数	10

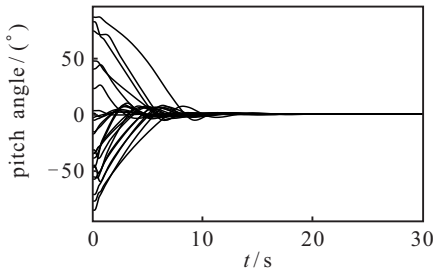
3　系统仿真

本文强化学习训练环境采用OpenAI gym平台,飞机空气动力学模型基于JSBSim开源平台F-16气动模型[12-14],其中气动系数使用NASA公布的实验数据. 飞机模型具有高阶非线性和延时的特点,变量之间相互耦合,随机初始化会导致网络不易收敛. 因此在实际训练中,姿态角初始化均值为0°,初始空速为200 m/s,归一化方差均为0.1. 随着迭代次数增大,方差逐渐增大,使智能体逐步探索状态空间,同时学习率逐步降低,避免梯度过大.

网络收敛后还应以时序响应作为判断依据. 图4给出了初始任意姿态角、角速度和空速条件下,飞机改平飞时的姿态角响应曲线. 由于滚转和俯仰通道模型复杂度不同,滚转角控制模型和气动模型相对简单,因此学习速度更快,相反俯仰通道模型相对复杂,学习速度较慢,当最终俯仰通道网络收敛时,滚转通道网络可能过拟合导致输出动作非全局最优,因此设置η_p和η_q为0.5. 由图4可以看到,两通道网络同时收敛,时域上姿态角的响应更符合预期,在全角度范围内姿态角能快速响应,滚转角基本在5 s内稳定,俯仰角由于过载限制需要15 s左右稳定. 根据李雅普诺夫稳定性理论[21],利用神经网络拟合动力学系统渐近收敛,图4中系统在初始状态大幅度扰动的情况下,控制系统均能保证姿态角恢复并实现零误差,具有稳定性.



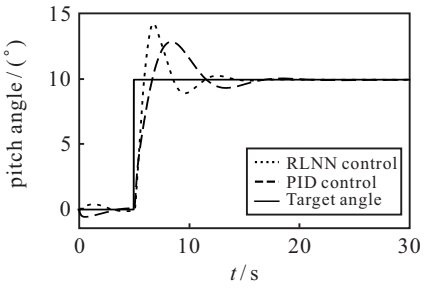
(a)　滚转角响应曲线



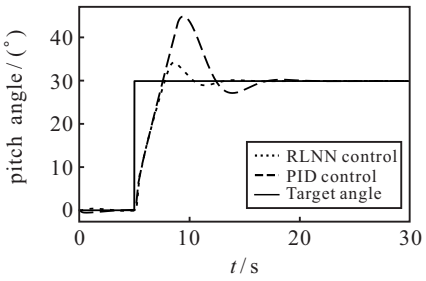
(b)　俯仰角响应曲线

图 4　任意姿态角改平飞响应曲线

为了验证控制器工作性能,本节进行传统PID控制器(PID control)与强化学习控制器(RLNN control)跟踪目标姿态角时的响应对比. 在滚转角保持0°的条件下,图5给出了由RLNN控制器和PID控制器跟踪阶跃函数的俯仰角响应曲线. 在10°阶跃响应时,两者差别较小,RLNN控制器超调较大但响应速度更快;当30°阶跃响应时,在上升阶段两者均为相同角速度,而RLNN控制器稳定时间更短,差别主要与奖励函数设置和累计误差限制幅度大小有关. 仿真发现,累计误差限幅越小超调越小,但响应速度会下降.



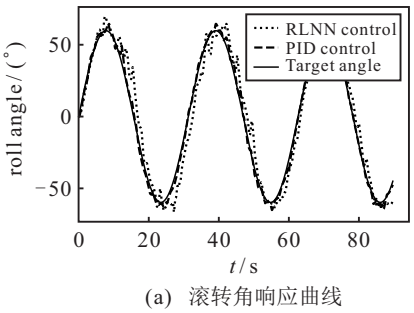
(a)　10°阶跃信号响应曲线



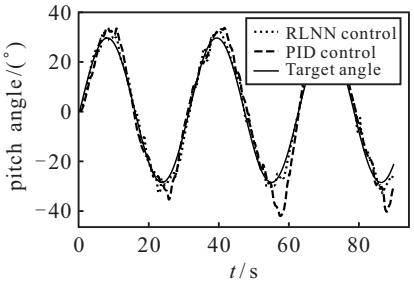
(b)　30°阶跃信号响应曲线

图 5　俯仰角阶跃信号响应曲线

为了验证控制器的鲁棒性,图6给出了扰动条件下姿态角跟踪正弦信号响应曲线,其中sine函数周期为10π,幅度分别为30°和60°. 在地面坐标系下引入随机突风、传感器噪声和发动机推力扰动,均服从高斯分布;在30 s仿真时间时放起落架和襟翼并



(a)　滚转角响应曲线



(b)　俯仰角响应曲线

图 6　姿态角正弦信号响应曲线

清空油箱,以改变飞机重心和气动力. 仿真结果表明,RLNN控制器和PID控制器在扰动条件下均具有鲁棒性,RLNN控制器面对扰动响应更快.

通过上述比较分析可以表明,基于强化学习的神经网络模型能够高效地优化各种飞机状态,如不同角速度、角度、空速、扰动和机动动作等条件下的响应,根据当前状态输出控制指令,从而实现姿态角快速跟踪,避免传统PID控制器在不同姿态下参数调节过程,有效减少设计人员工作量.

4 结 论

本文提出了一种基于强化学习的飞机姿态控制器,实现了任意初始状态下目标姿态角的跟踪. 相比于传统PID控制器,所提出的RLNN控制器核心思想在于能够通过网络模拟PID控制器的响应,遍历状态空间并找到当前状态下的数值最优解.

在未来的工作中,基于此姿态控制器,面向对抗场景下,采用强化学习自我博弈的机动决策系统将被进一步研究和实现,决策网络输入双方态势,由机动策略网络对态势进行评估并输出目标姿态、速度等控制维度,通过底层控制网络增加飞机模型的稳定性,降低训练难度,同时控制网络参数随着训练加深不断优化,以适应对抗场景.

参考文献(References)

[1] Wang C, Yan C, Xiang X J, et al. A continuous actor-critic reinforcement learning approach to flocking with fixed-wing UAVs[C]. The 11th Asian Conference on Machine Learning. Piscataway: IEEE, 2019: 1-8.

[2] Huang X, Luo W Y, Liu J R. Attitude control of fixed-wing UAV based on DDQN[C]. Chinese Automation Congress. Hangzhou, 2019: 4722-4726.

[3] Visioli A. Practical PID control[M]. Berlin: Springer Science & Business Media, 2006: 209-250.

[4] Wang J, Gao Z H. The automatic flight simulation of waypoint flight[J] . Flight Dynamics, 2008, 26(1): 75-78.

[5] Shu H L. PID neural network for decoupling control of strong coupling multivariable time-delay systems[J]. Control Theory and Applications, 1998, 15(6): 920-924.

[6] Yu D J, Long W Z, He J B. PID neural network decoupling control of multi-variable system and its application[C]. Proceedings of the 6th International Conference on Information Engineering for Mechanics and Materials. Huhhot, 2016: 277-282.

[7] Xu H, Lai J G, Yu Z H, et al. Based on neural network PID controller design and simulation[C]. Proceedings of the 2012 2nd International Conference on Computer and Information Applications. Paris, 2012: 508-511.

[8] Peng Z H, Wang D, Zhang H W, et al. Distributed neural network control for adaptive synchronization of uncertain dynamical multiagent systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(8): 1508-1519.

[9] Chen M, Ge S S, How B V E. Robust adaptive neural network control for a class of uncertain MIMO nonlinear systems with input nonlinearities[J]. IEEE Transactions on Neural Networks, 2010, 21(5): 796-812.

[10] Sufiyan D, Win L T S, Win S K H, et al. A reinforcement learning approach for control of a nature-inspired aerial vehicle[C]. International Conference on Robotics and Automation. Montreal, 2019: 6030-6036.

[11] Zhen Y, Hao M R, Sun W D. Deep reinforcement learning attitude control of fixed-wing UAVs[C]. The 3rd International Conference on Unmanned Systems. Harbin, 2020: 239-244.

[12] Russell R S. Nonlinear f_{16} simulations using Simulink and Matlab[R]. Minneapolis: University of Minnesota, 2003.

[13] Sonneveldt L. Nonlinear F_{16} model description[R]. Delft: Delft University of Technology, 2010.

[14] Baspinar B, Koyuncu E. Aerial combat simulation environment for one-on-one engagement[C]. AIAA Modeling and Simulation Technologies Conference. Kissimmee, 2018: 0432.

[15] Lerrel Pinto, James Davidson. Robust adversarial reinforcement learning[C]. Proceedings of the 34th International Conference on Machine Learning. Sydney, 2017: 1-10.

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

[17] Hill Ashley. Stable baselines[DB/OL]. (2018-06-02) [2021-10-16]. https://github.com/hill-a/stable-baselines.

[18] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization[C]. ICML. California: University of California, 2015: 1889-1897.

[19] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation[J/OL]. 2015, arXiv: 1506.02438.

[20] Kingma D P, Ba J. Adam: A method for stochastic optimization[J/OL]. 2017, arXiv: 1412.6980v9.

[21] Zhang S. Dynamical analysis and control of fractional-order neural networks based on Lyapunov method[D]. Beijing: Beijing Jiaotong University, 2017.

作者简介

付宇鹏(1991−), 男, 讲师, 博士, 从事强化学习、智能空战等研究, E-mail: 230169616@seu.edu.cn;

邓向阳(1981−), 男, 副教授, 博士, 从事计算智能、大数据方法等研究, E-mail: xiangyang.deng@qq.com;

何明(1978−), 男, 教授, 博士, 从事无人机集群协同控制等研究, E-mail: ming_he_2020@126.com;

朱子强(1996−), 男, 助教, 硕士, 从事计算机网络、强化学习等研究, E-mail: 994255320@qq.com;

张立民(1966−), 男, 教授, 博士生导师, 从事信息系统仿真、智能处理等研究, E-mail: iamzlm@163.com.