基于强化学习的小型无人直升机有限时间收敛控制设计

鲜 斌†, 林嘉裕

(天津大学 电气自动化与信息工程学院，天津 300072)

摘 要: 针对小型无人直升机精确动力学模型难以获取以及姿态控制易受未知外界风扰影响的问题,设计一种基于强化学习(reinforcement learning, RL)与super twisting相结合的非线性控制算法. 利用直升机在线飞行数据,训练执行者-评价者(actor-critic, AC)网络以逼近系统建模不确定部分. 为了抑制未知外界风扰,提高系统鲁棒性,同时补偿AC网络逼近误差,设计基于super twisting的鲁棒控制算法. 进而,利用Lyapunov稳定性分析方法证明无人直升机姿态误差能在有限时间内收敛到零. 最后对所提出的算法进行实验验证,实验结果表明,所提出算法具有良好的控制效果,对系统不确定性和外界扰动具有良好的鲁棒性.

关键词: 无人直升机；强化学习；鲁棒控制；未知外部扰动；有限时间收敛；实验验证

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0328 **开放科学(资源服务)标识码(OSID):**

引用格式: 鲜斌, 林嘉裕. 基于强化学习的小型无人直升机有限时间收敛控制设计[J]. 控制与决策, 2020, 35(11): 2646-2652.

Finite time control based on reinforcement learning for a small-size unmanned helicopter

XIAN Bin†, LIN Jia-yu

(School of Electrical and Information Engineering，Tianjin University，Tianjin 300072，China)

Abstract: This paper presents a nonlinear control law based on the combination of reinforcement learning (RL) and super twisting methodology for the attitude control of a small-size unmanned helicopter, which is subjected to modeling uncertainties and unknown external disturbances. The proposed control law only uses input and output data of the helicopter to train the actor-critic (AC) neural networks to compensate for modeling uncertainties. Then a nonlinear robust controller based on super twisting methodology is developed to compensate for the unknown external disturbances. The Lyapunov based stability analysis is used to prove that the attitude error of a unmanned helicopter can converge to zero in finite time. Finally, the proposed control law is verified on a self-built hardware in the loop testbed. The experimental results show that the proposed control law can achieve good control performance together with good robustness for modeling uncertainties and wind disturbances.

Keywords: helicopter; reinforcement learning; robust control; unknown external disturbances; finite time convergence; experimental verification

0 引 言

近年来,小型无人直升机凭借垂直起降能力、空中悬停能力以及具有灵活飞行的特点被视为工业级无人机的重要发展方向[1]. 然而,直升机具有非线性、强动态耦合以及其生成推力的空气动力学特性存在的参数和模型不确定性,难以建立精确的动力学模型,使得其控制设计极具挑战性[2].

传统线性控制算法大多基于线性化模型来设计,只能稳定于平衡点附近,且处理系统模型不确定性的能力有限,因此,研究人员开展了大量非线性控制算法的研究. 文献[3]设计自适应反步控制器,实现了无人直升机的姿态和高度控制. 文献[4]设计了滑模控制器以降低外部扰动对系统的影响,但滑模控制容易产生抖振现象. super twisting控制因其能够抑制抖振现象,且具备有限时间收敛的特性[5],广泛应用于无人直升机的控制设计中[6].

收稿日期: 2019-03-21；**修回日期:** 2019-06-20.
基金项目: 国家自然科学基金项目(91748121, 90916004, 60804004).
责任编委: 易建强.
†通讯作者. E-mail: xbin@tju.edu.cn.

针对小型无人直升机难以获取精确动力学模型的问题,强化学习等智能控制方法得到了广泛应用. 文献[7]基于策略搜索的强化学习算法实现了小型无人直升机的特技飞行. 文献[8]基于确定性策略搜索的强化学习方法,实现了四旋翼无人机的自主悬停控制. 但以上方法缺乏严格的稳定性证明.

强化学习强调智能体在与环境的交互过程中在线地进行学习,通过每次动作后环境的回报来修正自身行动策略,从而实现最优化决策[9]. 动态规划(dynamic programming, DP)是解决最优控制问题的有效方法,但此方法常用于离线训练,并且在系统复杂时容易引发“维数灾难”问题. 为应用DP方法,文献[10]提出了基于执行者-评价者(AC)结构的自适应动态规划(adaptive dynamic programming, ADP)方法,以在线获得系统的近似最优控制策略. 然而,对于实际系统,外界干扰总是存在的,单纯地使用ADP方法很难克服外界扰动的影响,因此,以上因素促使智能控制与非线性控制的结合. 文献[11]采取神经网络与反步法相结合的方法,通过仿真验证了所提出的轨迹跟踪控制设计的有效性. 文献[12]与文献[13]分别利用神经网络与强化学习在线估计系统不确定性,结合鲁棒控制算法,实现无人直升机的镇定抗扰飞行控制,并在理论上证明了姿态误差的半全局渐近收敛.

基于以上分析,本文将采用基于强化学习(RL)与super twisting相结合的非线性控制算法. 首先设计基于ADP方法的RL控制器,用于补偿建模不确定性;然后,通过基于super twisting的鲁棒控制器来抑制未知外界风扰的影响,以保证RL训练过程中系统的稳定性. 本文的创新性在于: 1)对比文献[4]等对模型依赖性较强的控制算法,本文所设计的算法对模型依赖性降低,减少了无人直升机建模不确定性对系统控制性能的影响; 2)文献[11]等研究多数仅通过数值仿真验证控制设计,尚未进行实时实验验证,而本文将控制算法应用于无人直升机半实物实验平台进行实时实验,取得了良好的控制效果,提高了算法应用于实际的可靠性; 3)文献[12]和文献[14]通过实验验证了控制算法的有效性,但理论上仅证明了姿态跟踪误差的半全局渐近收敛,而本文利用Lyapunov稳定性分析方法,从理论上证明了在受外界未知扰动和模型不确定性影响下无人直升机的姿态跟踪误差能在有限时间内收敛到零,提高了无人直升机控制的快速响应能力.值得一提的是,很少有研究成果能达到这个程度.

1　小型无人直升机动力学模型

基于文献[14],小型无人直升机的动力学模型可写成如下形式:

$$M(\eta)\ddot{\eta}+C(\eta,\dot{\eta})\dot{\eta}+\tau_d=S^{-\mathrm{T}}(AD\delta+B). \tag{1}$$

其中: $M(\eta)\in\boldsymbol{R}^{3\times3}$为惯性矩阵; $C(\eta,\dot{\eta})\in\boldsymbol{R}^{3\times3}$为科氏力矩阵; $\tau_d(t)$为外部未知扰动; $S(t)$为角速度变换矩阵; $A\in\boldsymbol{R}^{3\times3}$、$B\in\boldsymbol{R}^{3\times1}$为旋翼动力学相关矩阵; $D\in\boldsymbol{R}^{3\times3}$为旋翼挥舞角动力学相关矩阵; $\eta(t)=[\phi(t),\theta(t),\psi(t)]^{\mathrm{T}}$为姿态角向量, $\phi(t)$为滚转角, $\theta(t)$为俯仰角, $\psi(t)$为偏航角; $\dot{\eta}(t)$、$\ddot{\eta}(t)$为$\eta(t)$的一阶、二阶导数向量; $\delta(t)=[\delta_{\mathrm{lat}}(t),\delta_{\mathrm{lon}}(t),\delta_{\mathrm{ped}}(t)]^{\mathrm{T}}$为控制输入向量, $\delta_{\mathrm{lat}}(t)$为横向周期变距,是横滚舵机输入, $\delta_{\mathrm{lon}}(t)$为纵向周期变距,是俯仰舵机输入, $\delta_{\mathrm{ped}}(t)$为尾桨总距,是偏航舵机输入. 文献[13-15]给出了$M(\eta)$、$C(\eta,\dot{\eta})$、$S(t)$、A、B、D的具体表达式.

式(1)中的动力学模型为简化后的动力学模型,为解决模型中存在不确定性的问题,将$M(\eta)$、$C(\eta,\dot{\eta})$、B分别写为如下形式:

$$\begin{cases} M=M_0+M_{\Delta},\\ C=C_0+C_{\Delta},\\ B=B_0+B_{\Delta}. \end{cases} \tag{2}$$

其中: M_0、C_0、B_0分别为$M(\eta)$、$C(\eta,\dot{\eta})$、B的最佳估计矩阵, M_{Δ}、C_{Δ}、B_{Δ}为估计误差矩阵. 为方便后续控制设计与分析,定义$\Omega(t)=S^{-\mathrm{T}}AD$,并假设$M_{\Delta}\in L_{\infty}$、$C_{\Delta}\in L_{\infty}$、$B_{\Delta}\in L_{\infty}$及$\Omega(t)\in L_{\infty}$,则式(1)可写为如下形式:

$$M_0\ddot{\eta}+C_0\dot{\eta}+\tau_d=\Omega\delta+S^{-\mathrm{T}}B_0+p(\eta,\dot{\eta},\ddot{\eta}), \tag{3}$$

其中$p(\eta,\dot{\eta},\ddot{\eta})=S^{-\mathrm{T}}B_{\Delta}-M_{\Delta}\ddot{\eta}-C_{\Delta}\dot{\eta}$,表示模型中存在的不确定性.

为实现无人直升机的姿态角控制,定义系统姿态跟踪误差$e_1(t)=[e_{1\phi}(t),e_{1\theta}(t),e_{1\psi}(t)]^{\mathrm{T}}\in\boldsymbol{R}^{3\times1}$及线性滑模面$e_2(t)\in\boldsymbol{R}^{3\times1}$如下:

$$\begin{cases} e_1=\eta_d-\eta,\\ e_2=\dot{e}_1+ke_1. \end{cases} \tag{4}$$

其中: $k=\mathrm{diag}\{[k_1,k_2,k_3]^{\mathrm{T}}\}\in\boldsymbol{R}^{3\times3}$为对称正定增益矩阵, $\eta_d(t)=[\phi_d(t),\theta_d(t),\psi_d(t)]^{\mathrm{T}}\in\boldsymbol{R}^{3\times1}$为期望的有界跟踪轨迹.

本文的控制目标为: 设计强化学习控制律$\hat{N}(x)$及其参数更新律,保证神经网络权重估计误差一致最终有界,并设计控制输入$\delta(t)$,使无人直升机姿态角向量$\eta(t)$跟踪期望轨迹$\eta_d(t)$,并保证跟踪误差信号能在有限时间内收敛到零.

2 控制器设计

为提高系统模型不确定性近似的准确度,本文采用执行网-评价网的结构形式设计强化学习控制律,采用super twisting方法来设计鲁棒控制器. 为方便后续控制设计,定义如下状态值函数作为系统的性能指标函数:

$$J(e_1)=\int_t^{\infty} r(e_1(s),\tau(s))\mathrm{d}s. \tag{5}$$

其中:$\tau=\Omega\delta$为无人直升机质心所受力矩;$r(e_1(s),\tau(s))=e_1^{\mathrm{T}}Qe_1+\tau^{\mathrm{T}}R\tau$为根据无人机姿态跟踪误差和质心所受力矩定义的回报函数,且$Q\in\boldsymbol{R}^{3\times3}$,$R\in\boldsymbol{R}^{3\times3}$为正定对称常数矩阵. 根据最优控制理论,定义哈密尔顿函数[16]为

$$H(e_1,\tau)=r(e_1(s),\tau(s))+\nabla J\dot{e}_1, \tag{6}$$

其中∇J为状态值函数梯度,且$\nabla J=\dfrac{\partial J}{\partial e_1}$.

定义最优控制策略τ^*对应的最优状态值函数为

$$J^*(e_1)=\min\int_t^{\infty} r(e_1(s),\tau^*(s))\mathrm{d}s. \tag{7}$$

当控制量为最优控制策略τ^*时,最优状态值函数$J^*(e_1)$满足如下哈密尔顿方程:

$$H(e_1,\tau^*)=r(e_1(s),\tau^*(s))+\nabla J^*\dot{e}_1=0. \tag{8}$$

考虑到式(8)求解困难,一般设计执行网络与评价网络来逼近该方程的近似解,从而得到最优控制策略[16].

2.1 评价网络设计

利用如下神经网络来表示最优状态值函数$J^*(e_1)$:

$$J^*(e_1)=W_c^{\mathrm{T}}\varphi_c(e_1)+\varepsilon_c. \tag{9}$$

其中:$W_c(t)$为评价网络理想权重矩阵,$\varphi_c(\cdot)$选择双曲正切函数$\tanh(\cdot)$作为神经网络的激励函数,$\varepsilon_c\in\boldsymbol{R}^{3\times1}$为评价网络逼近误差.

为了实现对最优状态值函数的逼近,设计如下神经网络:

$$\hat{J}(e_1)=\hat{W}_c^{\mathrm{T}}\varphi_c(e_1). \tag{10}$$

其中:$\hat{W}_c(t)$是对理想权重$W_c(t)$的估计,评价网络的权重估计误差为$\tilde{W}_c=W_c-\hat{W}_c$. 这里定义Bellman误差变量$e_c(t)$为

$$e_c=\hat{W}_c^{\mathrm{T}}\nabla\varphi_c\dot{e}_1+r=-\tilde{W}_c^{\mathrm{T}}\nabla\varphi_c\dot{e}_1+\varepsilon_H, \tag{11}$$

其中ε_H为辅助信号,且$\varepsilon_H=W_c^{\mathrm{T}}\nabla\varphi_c\dot{e}_1+r$.

为使残差的平方$E_c=\dfrac{1}{2}e_c^{\mathrm{T}}e_c$最小,设计$\hat{W}_c(t)$的更新律如下:

$$\dot{\hat{W}}_c=-a_c\frac{\partial E_c}{\partial\hat{W}_c}=-a_c\frac{\xi_1(\xi_1^{\mathrm{T}}\hat{W}_c+r)}{(\xi_1^{\mathrm{T}}\xi_1+1)^2}. \tag{12}$$

其中:$\xi_1=\nabla\varphi_c\dot{e}_1$;$a_c$为评价网络的学习率,是正常数. 为便于分析,此处定义$\xi_2=\xi_1/\xi_3$,$\xi_3=\xi_1^{\mathrm{T}}\xi_1+1$,且满足$\xi_{2m}\leqslant\|\xi_2\|\leqslant\xi_{2M}$,$\xi_{3m}\leqslant\|\xi_3\|\leqslant\xi_{3M}$. 则

$$\dot{\tilde{W}}_c=-a_c\xi_2\xi_2^{\mathrm{T}}\tilde{W}_c+a_2\xi_2\frac{\varepsilon_H}{\xi_3}. \tag{13}$$

2.2 执行网络设计

由式(4)可知,$e_2(t)$与$e_1(t)$具有相同的收敛性. 对$e_2(t)$求一阶时间导数并代入式(3),可得

$$\dot{e}_2=N(x)-M_0^{-1}\Omega\delta+M_0^{-1}\tau_d. \tag{14}$$

其中:$x=[\eta^{\mathrm{T}},\dot{\eta}^{\mathrm{T}},\ddot{\eta}^{\mathrm{T}},\dot{e}_1^{\mathrm{T}}]^{\mathrm{T}}\in\boldsymbol{R}^{12\times1}$为状态变量;辅助函数$N(x)$为模型的不确定部分,表达式为

$$N(x)=\ddot{\eta}_d+k\dot{e}_1-M_0^{-1}[S^{-\mathrm{T}}B_0+p(\eta,\dot{\eta},\ddot{\eta})-C_0\dot{\eta}].$$

利用如下神经网络来表示$N(x)$:

$$N(x)=W_a^{\mathrm{T}}\varphi_a(x)+\varepsilon_a. \tag{15}$$

其中:$W_a(t)$表示理想权重矩阵,$\varphi_a(\cdot)$选择双曲正切函数$\tanh(\cdot)$作为执行网络的激励函数,$\varepsilon_a\in\boldsymbol{R}^{3\times1}$为执行网络逼近误差.

为实现对模型的不确定部分$N(x)$的逼近,设计如下神经网络:

$$\hat{N}(x)=\hat{W}_a^{\mathrm{T}}\varphi_a(x). \tag{16}$$

其中:$\hat{W}_a(t)$是对执行网络理想权重的估计,执行网络权重估计误差为$\tilde{W}_a=W_a-\hat{W}_a$. 这里定义反馈误差信号$e_a(t)$为

$$e_a=\hat{W}_a^{\mathrm{T}}\varphi_a(x)+k_z\nabla\varphi_c^{\mathrm{T}}\hat{W}_c, \tag{17}$$

其中k_z为执行网络的增益参数,且$k_z>0$. 为使反馈误差信号的平方$E_a=\dfrac{1}{2}e_a^{\mathrm{T}}e_a$最小,设计$\hat{W}_a(t)$的更新律[17]为

$$\dot{\hat{W}}_a=-a_a\varphi_a(\hat{W}_a^{\mathrm{T}}\varphi_a+k_z\nabla\varphi_c^{\mathrm{T}}\hat{W}_c)^{\mathrm{T}}, \tag{18}$$

其中a_a为执行网络的学习率,为正常数. 将$\tilde{W}_a=W_a-\hat{W}_a$代入式(18),可得

$$\dot{\tilde{W}}_a=-a_a\varphi_a(\tilde{W}_a^{\mathrm{T}}\varphi_a+k_z\nabla\varphi_c^{\mathrm{T}}\tilde{W}_c+\varepsilon_z)^{\mathrm{T}}, \tag{19}$$

其中ε_z为辅助信号,且$\varepsilon_z=W_a^{\mathrm{T}}\varphi_a+k_z\nabla\varphi_c^{\mathrm{T}}W_c$.

根据神经网络的性质,给出如下假设.

假设1 执行网络与评价网络的理想权重向量有界,隐藏层激励函数满足$\varphi_{am}\leqslant\|\varphi_a\|\leqslant\varphi_{aM}$,$\varphi_{cm}\leqslant\|\varphi_c\|\leqslant\varphi_{cM}$. 由于激励函数选择双曲正切函数,可知$\|\nabla\varphi_c\|\leqslant\varphi_{cdM}$,则$\|\varepsilon_H\|\leqslant\varepsilon_{HM}$,$\|\varepsilon_z\|\leqslant\varepsilon_{zM}$. φ_{am}、φ_{aM}、φ_{cm}、φ_{cM}、φ_{cdM}、ε_{HM}以及ε_{zM}均为正常数[17].

2.3 强化学习收敛性分析

定理1 对于开环系统(14),若执行网络学习率a_a、评价网络学习率a_c以及后面的式(23)中变量满

足式(24),则式(16)的强化学习控制律以及式(12)和(18)的更新律,能使执行网络与评价网络的权重估计误差$\tilde{W}_a(t)$、$\tilde{W}_c(t)$达到一致最终有界.

证明　定义Lyapunov候选函数为

$$L(t)=L_1(t)+L_2(t). \tag{20}$$

其中: $L_1(t)=\frac{1}{2a_c}\tilde{W}_c^{\mathrm{T}}\tilde{W}_c, L_2(t)=\frac{1}{2a_a}\tilde{W}_a^{\mathrm{T}}\tilde{W}_a$. 则可知$L(t)$为正定函数. 对式(13)求一阶时间导数,有

$$\begin{aligned}\dot{L}_1(t)=&\frac{1}{a_c}\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c=\\&\frac{1}{a_c}\Big[-a_c\tilde{W}_c^{\mathrm{T}}\xi_2\Big(\xi_2^{\mathrm{T}}\tilde{W}_c^{\mathrm{T}}-\frac{\varepsilon_H}{\xi_3}\Big)\Big]\leqslant\\&-\xi_{2m}^2\|\tilde{W}_c\|^2+\frac{a_c}{2}\xi_{2M}^2\|\tilde{W}_c\|^2+\frac{1}{2a_c}\frac{\varepsilon_{HM}^2}{\xi_{3m}^2}.\end{aligned} \tag{21}$$

对式(19)求一阶时间导数,有

$$\begin{aligned}\dot{L}_2(t)=&\frac{1}{a_a}\tilde{W}_a^{\mathrm{T}}\dot{\tilde{W}}_a=\\&\frac{1}{a_a}[-a_a\tilde{W}_a^{\mathrm{T}}\varphi_a(\tilde{W}_a^{\mathrm{T}}\varphi_a+k_z\nabla\varphi_c^{\mathrm{T}}\tilde{W}_c+\varepsilon_z)^{\mathrm{T}}]\leqslant\\&-\varphi_{am}^2\|\tilde{W}_a\|^2+a_a^2\varphi_{aM}^2\|\tilde{W}_a\|^2+\\&\frac{k_z}{2a_a}\varphi_{cdM}^2\|\tilde{W}_c\|^2+\frac{1}{2a_a}\varepsilon_{zM}^2.\end{aligned} \tag{22}$$

将式(21)、(22)代入(20),可得

$$\begin{aligned}\dot{L}(t)=&\dot{L}_1(t)+\dot{L}_2(t)\leqslant\\&-\Big(\xi_{2m}^2-\frac{a_c}{2}\xi_{2M}^2-\frac{k_z}{2a_a}\varphi_{cdM}^2\Big)\|\tilde{W}_c\|^2-\\&(\varphi_{am}^2-a_a\varphi_{aM}^2)\|\tilde{W}_a\|^2+\varepsilon_{LM},\end{aligned} \tag{23}$$

其中$\varepsilon_{LM}=\frac{1}{2a_c}\frac{\varepsilon_{HM}^2}{\xi_{3m}^2}+\frac{1}{2a_a}\varepsilon_{zM}^2$,为正常数. 当满足如下条件时:

$$\begin{cases}a_a<\dfrac{\varphi_{am}^2}{\varphi_{aM}^2},\\[2mm] a_c<\dfrac{2\xi_{2m}^2}{\xi_{2M}^2}-\dfrac{k_z\varphi_{cdM}^2}{\xi_{2M}^2},\\[2mm] \|\tilde{W}_c\|>\sqrt{\dfrac{\varepsilon_{LM}}{\xi_{2m}^2-\frac{a_c}{2}\xi_{2M}^2-\frac{k_z}{2a_a}\varphi_{cdM}^2}},\\[2mm] \|\tilde{W}_a\|>\sqrt{\dfrac{\varepsilon_{LM}}{\varphi_{am}^2-a_a\varphi_{aM}^2}},\end{cases} \tag{24}$$

可得$\dot{L}(t)<0$. 根据Lyapunov理论,执行网络与评价网络权重估计误差$\tilde{W}_a(t)$、$\tilde{W}_c(t)$能达到一致最终有界. □

2.4　鲁棒控制器设计

基于式(14),设计系统的控制输入为

$$\delta=\Omega^{-1}M_0\Big[\alpha|e_2|^{\frac{1}{2}}\mathrm{sgn}(e_2)+\beta\int_0^t\mathrm{sgn}(e_2)\mathrm{d}\tau+\hat{W}_a^{\mathrm{T}}\varphi_a(x)\Big]. \tag{25}$$

其中: $\alpha=\mathrm{diag}\{[\alpha_\phi,\alpha_\theta,\alpha_\psi]^{\mathrm{T}}\}\in\boldsymbol{R}^{3\times3}$、$\beta=\mathrm{diag}\{[\beta_\phi,\beta_\theta,\beta_\psi]^{\mathrm{T}}\}\in\boldsymbol{R}^{3\times3}$为super twisting控制增益矩阵, $\mathrm{sgn}(\cdot)$为标准的符号函数.

将式(25)代入(14)并整理,可得

$$\begin{cases}\dot{e_2}=-\alpha|e_2|^{\frac{1}{2}}\mathrm{sgn}(e_2)+y+\omega(t),\\ \dot{y}=-\beta\mathrm{sgn}(e_2).\end{cases} \tag{26}$$

其中: $y=-\beta\int_0^t\mathrm{sgn}(e_2)\mathrm{d}\tau, \omega(t)=\tilde{W}_a^{\mathrm{T}}\varphi_a(x)+\varepsilon_a+d, d=M_0^{-1}\tau_d$.

考虑到实际环境中存在干扰,以及为方便后续分析,给出如下假设.

假设2　扰动$d(t)$是有界的,且满足$\|d(t)\|\leqslant d_M, d_M$为正常数.

假设3　由定理1可知,执行网络与评价网络权重估计误差一致最终有界,从而根据文献[5],假设$\|\omega(t)\|\leqslant\mu_1+\mu_2\sqrt{|e_2|+y^2}$, μ_1、μ_2均为非负常数.

3　系统稳定性分析

为方便后续证明,引入以下引理,引理证明参见文献[18].

引理1　对于给定的对称矩阵$\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{12}^{\mathrm{T}}&\Sigma_{22}\end{bmatrix}$,以下3个条件等价:

1) $\Sigma>0$;

2) $\Sigma_{11}>0, \Sigma_{22}-\Sigma_{12}^{\mathrm{T}}\Sigma_{11}^{-1}\Sigma_{12}>0$;

3) $\Sigma_{22}>0, \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}}>0$.

定理2　对于对称矩阵

$$Q_s=\begin{bmatrix}\alpha+2\beta p_{12}&-\frac{1}{2}(1-\alpha p_{12})+\beta p_{22}\\ -\frac{1}{2}(1-\alpha p_{12})+\beta p_{22}&-p_{12}\end{bmatrix},$$

存在正定对称矩阵

$$P=\begin{bmatrix}p_{11}&p_{12}\\ p_{12}^{\mathrm{T}}&p_{22}\end{bmatrix}.$$

其中: $p_{11}=1, p_{12}=-\sqrt{\frac{a}{2\gamma\Gamma}}, p_{22}=\frac{a}{2\Gamma}, \gamma>1, a$、$\Gamma$均为正常数. 若super twisting增益$\alpha$、$\beta$大小满足后面的式(29)和(30)所示不等式,则$Q_s$为正定矩阵.

证明　根据引理1,由$-p_{12}>0$可知,使Q_s正定的条件为

$$\begin{aligned}&-p_{12}(\alpha+2\beta p_{12})-\frac{1}{4}(1-\alpha p_{12})^2-\\&\beta^2p_{22}^2+\beta(1-\alpha p_{12})p_{22}>0.\end{aligned} \tag{27}$$

令$\zeta=-\alpha p_{12}, \gamma=\frac{p_{22}}{p_{12}^2}, a=\beta p_{22}$,则$\alpha=\zeta\sqrt{\frac{2\gamma\Gamma}{a}}, \beta=$

2Γ. 式(27)可写为

$$\zeta-\frac{2}{\gamma}a>a^2+\frac{1}{4}(1+\zeta)^2-(1+\zeta)a. \tag{28}$$

根据文献[5],当$\gamma>1$时,曲线$\zeta-\dfrac{2}{\gamma}a=a^2+\dfrac{1}{4}(1+\zeta)^2-(1+\zeta)a$存在. 为便于讨论,选取$\gamma=100$,利用Matlab工具画出曲线如图1所示.

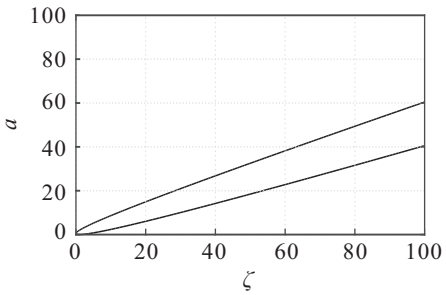

图1 $\zeta-\dfrac{2}{\gamma}a=a^2+\dfrac{1}{4}(1+\zeta)^2-(1+\zeta)a$**曲线**

由图1可知,当(ζ,a)处于曲线内侧时,$\zeta>0$. 取定ζ,由根与系数的关系可知,当γ满足下式时:

$$\begin{cases}\gamma>1,\ \zeta>1;\\ \gamma>\dfrac{1}{\zeta},\ 0<\zeta<1.\end{cases} \tag{29}$$

方程$\zeta-\dfrac{2}{\gamma}a=a^2+\dfrac{1}{4}(1+\zeta)^2-(1+\zeta)a$存在两个根,即

$$a_1=\frac{1}{2}\Big\{(1+\zeta)-\frac{2}{\gamma}\Big[1+\sqrt{\zeta\gamma^2-(1+\zeta)\gamma+1}\Big]\Big\},$$

$$a_2=\frac{1}{2}\Big\{(1+\zeta)-\frac{2}{\gamma}\Big[1-\sqrt{\zeta\gamma^2-(1+\zeta)\gamma+1}\Big]\Big\}.$$

因此,当$a_1<a<a_2$时,super twisting增益α、β大小满足式(29)和如下不等式:

$$\begin{cases}\zeta\sqrt{\dfrac{2\gamma\Gamma}{a_2}}<\alpha<\zeta\sqrt{\dfrac{2\gamma\Gamma}{a_1}},\\ \beta>0,\end{cases} \tag{30}$$

因此Q_s为正定矩阵. □

定理3 对于开环系统(14),当满足假设2,且super twisting增益α、β大小满足定理2中条件时,式(25)的控制输入能够使姿态角跟踪误差信号$e_1(t)$、$e_2(t)$在有限时间内收敛到零.

证明 定义Lyapunov候选函数为

$$V(t)=\xi^{\mathrm{T}}P\xi. \tag{31}$$

其中:$P=\begin{bmatrix}p_{11} & p_{12}\\ p_{12}^{\mathrm{T}} & p_{22}\end{bmatrix}$,$\xi(t)=[|e_2|^{\frac{1}{2}}\mathrm{sgn}(e_2),y]^{\mathrm{T}}$. $V(t)$为二次正定函数,且径向无界,满足如下不等式:

$$\lambda_{\min}(P)\|\xi\|_2^2\leqslant V(t)\leqslant\lambda_{\max}(P)\|\xi\|_2^2. \tag{32}$$

其中:$\lambda_{\min}(P)$和$\lambda_{\max}(P)$分别代表矩阵P的最小、最大特征值,$\|\xi\|_2$代表欧几里得范数. 令$\|\xi\|_2^2=\xi_a^2+\xi_b^2$,可得

$$|\xi_a|=|e_2|^{\frac{1}{2}}\leqslant\|\xi\|_2\leqslant\lambda_{\min}^{-\frac{1}{2}}(P)V(t)^{\frac{1}{2}}, \tag{33}$$

$$\|\xi\|_2\geqslant\lambda_{\max}^{-\frac{1}{2}}(P)V(t)^{\frac{1}{2}}. \tag{34}$$

对状态变量$\xi(t)$求导,可得

$$\dot{\xi}=\frac{1}{|\xi_a|}\begin{bmatrix}-\dfrac{\alpha}{2}|e_2|^{\frac{1}{2}}\mathrm{sgn}(e_2)+\dfrac{1}{2}y+\dfrac{1}{2}\omega(t)\\ -\beta\mathrm{sgn}(e_2)+0+0\end{bmatrix}. \tag{35}$$

令$A_1=\begin{bmatrix}-\dfrac{\alpha}{2} & \dfrac{1}{2}\\ -\beta & 0\end{bmatrix}$,$B_1=\begin{bmatrix}1\\0\end{bmatrix}$,$\tilde{\omega}=\dfrac{1}{2}\omega(t)$,则由式(35)可得

$$\begin{cases}\dot{\xi}=\dfrac{1}{|\xi_a|}(A_1\xi+B_1\tilde{\omega}),\\ \dot{\xi}^{\mathrm{T}}=\dfrac{1}{|\xi_a|}(\xi^{\mathrm{T}}A_1^{\mathrm{T}}+\tilde{\omega}^{\mathrm{T}}B_1^{\mathrm{T}}).\end{cases} \tag{36}$$

对$V(t)$求一阶时间导数,可得

$$\begin{aligned}\dot{V}(t)=2\dot{\xi}^{\mathrm{T}}P\xi=\\ \frac{1}{|\xi_a|}[\xi^{\mathrm{T}}(A_1^{\mathrm{T}}P+PA_1)\xi+2\tilde{\omega}^{\mathrm{T}}B_1^{\mathrm{T}}P\xi]=\\ -\frac{1}{|\xi_a|}\{\xi^{\mathrm{T}}Q_s\xi-[\omega(t)\quad 0]P\xi\}.\end{aligned} \tag{37}$$

其中:矩阵

$$Q_s=-(A_1^{\mathrm{T}}P+PA_1)=\begin{bmatrix}\alpha+2\beta p_{12} & -\dfrac{1}{2}(1-\alpha p_{12})+\beta p_{22}\\ -\dfrac{1}{2}(1-\alpha p_{12})+\beta p_{22} & -p_{12}\end{bmatrix},$$

$p_{22}>0$, $p_{12}<0$. 由定理2可知,Q_s为正定矩阵,则

$$\lambda_{\min}(Q_s)\|\xi\|_2^2\leqslant\xi^{\mathrm{T}}Q_s\xi\leqslant\lambda_{\max}(Q_s)\|\xi\|_2^2. \tag{38}$$

根据假设2的条件以及三角不等式,可得

$$\begin{aligned}[\omega(t)\quad 0]P\xi=\omega(t)[\|e_2\|^{\frac{1}{2}}\mathrm{sgn}(e_2)+p_{12}y]\leqslant\\ (\mu_1+\mu_2\|\xi\|_2)\cdot\rho\|\xi\|_2,\end{aligned} \tag{39}$$

其中$\rho=\sqrt{1+p_{12}^2}$. 若常数$\mu_2<\lambda_{\min}(Q_s)/\rho$,则将式(38)、(39)代入(37),经过整理可得

$$\dot{V}(t)\leqslant-\frac{\|\xi\|_2}{\|\xi_a\|}\{[\lambda_{\min}(Q_s)-\mu_2\rho]\|\xi\|_2-\mu_1\rho\}.$$

再根据式(33)、(34)对$\dot{V}(t)$进一步缩放,可得

$$\begin{aligned}\dot{V}(t)\leqslant\\ -[\lambda_{\min}(Q_s)-\mu_2\rho](k_v+1-k_v)\|\xi\|_2+\mu_1\rho\leqslant\\ -k_v\epsilon V(t)^{\frac{1}{2}},\ \forall\|\xi\|_2\geqslant\vartheta.\end{aligned} \tag{40}$$

其中:k_v、ϵ、ϑ均为正常数,$\epsilon=\dfrac{\lambda_{\min}(Q_s)-\mu_2\rho}{\lambda_{\max}^{\frac{1}{2}}(P)}$,$\vartheta=\dfrac{\mu_1\rho}{(1-k_v)[\lambda_{\min}(Q_s)-\mu_2\rho]}$,$k_v$满足$0<k_v<1$.

根据文献[5],系统状态在T_F时间内收敛到集合

$$\Pi_\rho=\{e_2\in\boldsymbol{R}^3|V(e_2)\leqslant\lambda_{\max}(P)\vartheta^2\}. \tag{41}$$

当$\mu_1 = 0$且μ_2满足$\mu_2 < \dfrac{\lambda_{\min}(Q_s)}{\rho}$时,可知$\vartheta = 0, \epsilon > 0$. 根据芭芭拉定理,可得系统姿态跟踪误差信号$e_1(t)$、$e_2(t)$满足

$$\lim_{t\to T_F} e_2(t) = \lim_{t\to T_F} e_1(t) = 0.$$

即系统姿态跟踪误差信号$e_1(t)$、$e_2(t)$能在有限时间内收敛到零. □

注1　关于系统收敛时间T_F,通过求解式(41)的微分不等式可得

$$\sqrt{V(t)} - \sqrt{V(t_0)} \leqslant -\frac{k_v\epsilon}{2}t, \tag{42}$$

其中t_0为系统处于初始状态时刻,当$t = T_F$时,系统到达集合Π_ρ. 因此,$V(T_F) = \lambda_{\max}(P)\vartheta^2$,从而$\sqrt{V(T_F)} = \lambda_{\max}^{\frac{1}{2}}(P)\vartheta$. 由$\dot{V}(t) < 0$可知,$V(t)$是单调递减函数,故对于任意$t \geqslant T_F$,有

$$\lambda_{\max}(P)\vartheta^2 \leqslant V(t) \leqslant V(T_F) = \lambda_{\max}(P)\vartheta^2. \tag{43}$$

将式(43)代入(42),可得

$$\lambda_{\max}^{\frac{1}{2}}(P)\vartheta - \sqrt{V(t_0)} \leqslant -\frac{k_v\epsilon}{2}T_F. \tag{44}$$

进一步，可计算得到收敛时间T_F为

$$T_F = \frac{2}{k_v\epsilon}[\sqrt{V(t_0)} - \lambda_{\max}^{\frac{1}{2}}(P)\vartheta] = \frac{2\lambda_{\max}^{\frac{1}{2}}(P)}{k_v[\lambda_{\min}(Q_s) - \mu_2\rho]}[\sqrt{V(t_0)} - \lambda_{\max}^{\frac{1}{2}}(P)\vartheta]. \tag{45}$$

4　实验验证

为了验证本文所设计控制器的有效性,采用自制的无人直升机半实物实验平台进行实时镇定抗扰实验,并在相同条件下与传统滑模控制器进行对比实验. 实验持续时间约160 s,实验开始先手动操作飞行,约18 s切换自动飞行模式,无人机根据期望轨迹飞行. 在90 s后,加入持续定向风扰,无人机进行抗扰飞行,实验结果如图2～图4所示. 两组实验设定的期望轨迹均为$\eta_d(t) = [0,0,0]^{\mathrm{T}}$,无人直升机模型参数参见文献[14].

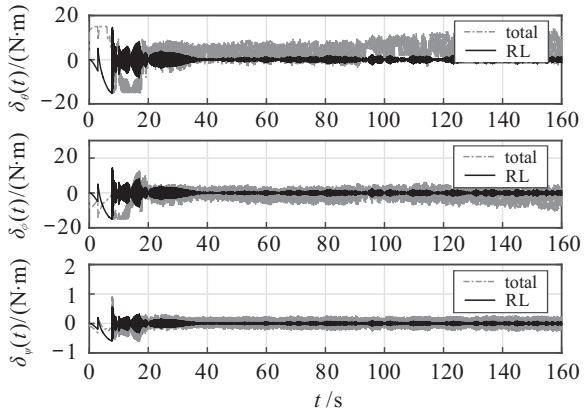

图 2　强化学习鲁棒控制-镇定抗扰实验控制量曲线

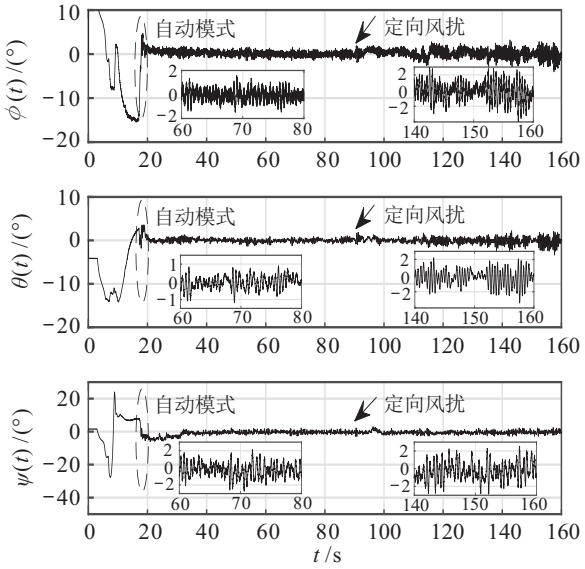

图 3　强化学习鲁棒控制-镇定抗扰实验姿态角曲线

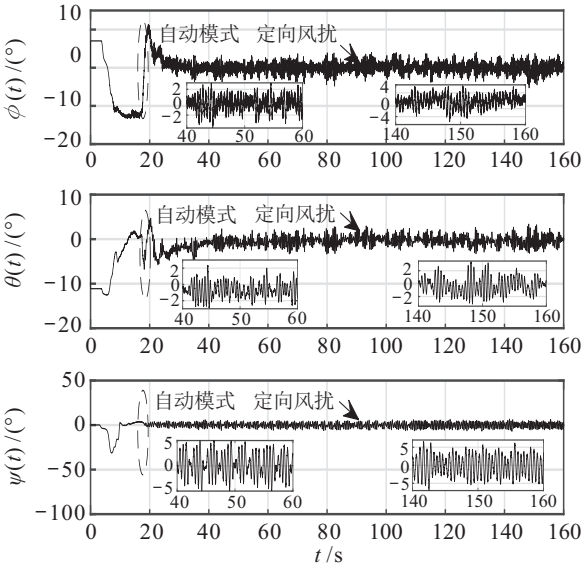

图 4　传统滑模控制-镇定抗扰实验姿态角曲线

4.1　强化学习鲁棒控制器

实验中先调整鲁棒控制器增益参数,使直升机具有一定的鲁棒性后,加入强化学习控制项,逐步调整网络增益参数,确保控制器的收敛性,同时微调鲁棒控制器增益以优化控制效果. 最终鲁棒控制器参数选取为: $k = \mathrm{diag}\{[220,240,12]^{\mathrm{T}}\}$, $\alpha = \mathrm{diag}\{[75,75,1.5]^{\mathrm{T}}\}$, $\beta = \mathrm{diag}\{[80,80,1.5]^{\mathrm{T}}\}$. AC网络隐层节点选取10个,权重$W_a \in \boldsymbol{R}^{10\times 3}$、$W_c \in \boldsymbol{R}^{10\times 3}$初值设置为0.01,评价网络中$Q$、$R$矩阵取为单位矩阵. AC网络参数选取为: $a_a = \mathrm{diag}\{[1.2,1.2,0.012]^{\mathrm{T}}\}$, $a_c = \mathrm{diag}\{[1.5,1.5,0.015]^{\mathrm{T}}\}$, $k_z = \mathrm{diag}\{[0.1,0.1,0.1]^{\mathrm{T}}\}$.

为分析强化学习所产生的控制作用,分别画出强化学习部分的控制输入以及总控制输入曲线如图2所示. 由图2可知,直升机刚进入自动模式时,状态还未稳定,此时强化学习作用较为明显,达到约40 %的控制占比. 进入稳态后,强化学习控制占比逐渐降低,

约占2%. 加入风扰后,由于状态受干扰,强化学习控制占比提高,约占10%. 由此验证了强化学习控制律对模型不确定性估计的有效性.

4.2 传统滑模控制器

设计滑模面及控制律为

$$e_2 = \dot{e}_1 + ke_1,\ u = -k_{\rm smc}{\rm sgn}(e_2). \tag{46}$$

选取实验效果较好的一组控制器增益参数为$k = {\rm diag}\{[220, 240, 12]^{\rm T}\}, k_{\rm smc} = {\rm diag}\{[55, 58, 2.8]^{\rm T}\}$,得到图4所示实验结果.

由图3可以看出:强化学习鲁棒控制器约2 s实现镇定飞行,滚转角和偏航角精度控制在$\pm 2.1^\circ$以内,俯仰角精度控制在$\pm 1.2^\circ$以内;风扰状态下滚转角和俯仰角精度控制在$\pm 3^\circ$以内,偏航角精度控制在$\pm 2.1^\circ$以内. 由图4可以看出:传统滑模控制器约12 s实现镇定飞行,滚转角和俯仰角精度控制在$\pm 3.5^\circ$以内,偏航角精度控制在$\pm 5^\circ$以内;风扰状态下滚转角和俯仰角精度控制在$\pm 4^\circ$以内,偏航角精度控制在$\pm 5.5^\circ$以内. 由此可见,本文算法具有更快的收敛性,并且对风扰有着更好的鲁棒性.

5 结 论

本文针对小型无人直升机难以获取精确动力学模型以及姿态控制易受未知外界风扰影响的问题,设计了基于强化学习的非线性鲁棒控制器. 利用AC网络逼近系统不确定性,基于super twisting的鲁棒控制器抑制外界未知风扰. 基于Lyapunov方法从理论上证明了姿态跟踪误差能在有限时间内收敛到零. 同时,利用自制的半实物实验平台与传统滑模控制进行了对比实验,所得结果验证了强化学习控制律对模型不确定性估计的有效性、所设计的算法在有限时间内收敛的快速性以及对未知外界风扰的鲁棒性.

参考文献(References)

[1] Sheng S Z, Wang D B, Jiang B, et al. Longitudinal and lateral adaptive control without attitude feedback for a new prototype unmanned helicopter[J]. Control and Decision, 2010, 25(8): 1215-1219.

[2] Zhou H B, Pei H L, He Y B, et al. Trajectory tracking control of unmanned helicopter via filtering backstepping[J]. Control and Decision, 2012, 27(4): 613-617.

[3] Sun X Y, Fang Y C, Sun N. Backstepping-based adaptive attitude and height control of a small-scale unmanned helicopter[J]. Control Theory & Applications, 2012, 29(3): 381-388.

[4] Odelga M, Chriette A, Plestan F. Control of 3 dof helicopter: A novel autopilot scheme based on adaptive sliding mode control[C]. 2012 American Control Conference (ACC). Montréal: IEEE, 2012: 2545-2550.

[5] Moreno J A, Osorio M. Strict Lyapunov functions for the super twisting algorithm[J]. IEEE Transactions on Automatic Control, 2012, 57(4): 1035-1040.

[6] Fang X, Wu A, Shang Y J, et al. Multivariable super twisting based robust trajectory tracking control for small unmanned helicopter[J]. Mathematical Problems in Engineering, 2015, 2015: 1-13.

[7] Ng A Y, Jordan M I. Shaping and policy search in reinforcement learning[D]. California: Department of Computer Sciences, University of California. Berkeley, 2003.

[8] Hwangbo J, Sa I, Siegwart R, et al. Control of a quadrotor with reinforcement learning[J]. IEEE Robotics and Automation Letters, 2017, 2(4): 2096-2103.

[9] Liu D R, Li H L, Wang D. Data-based self-learning optimal control: Research progress and prospects[J]. Acta Automatica Sinica, 2013, 39(11): 1858-1870.

[10] Werbos P J. Consistency of HDP applied to a simple reinforcement learning problem[J]. Neural Networks, 1990, 3(2): 179-189.

[11] Nodland D, Zargarzadeh H, Jagannathan S. Neural network-based optimal adaptive output feedback control of a helicopter UAV[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(7): 1061-1073.

[12] Xian B, Zhang H N. Nonlinear robust control for a small unmanned helicopter based on neural network[J]. Control and Decision, 2018, 33(4): 627-632.

[13] Xian B, Gao J C, Zhang Y, et al. Sliding mode tracking control for miniature unmanned helicopters[J]. Chinese Journal of Aeronautics, 2015, 28(1): 277-284.

[14] An H, Xian B. Attitude reinforcement learning control of an unmanned helicopter with verification[J]. Control Theory & Applications, 2019, 36(4): 516-524.

[15] Cai G W, Chen B M, Lee T H. Unmanned rotorcraft systems[M]. London: Springer Science and Business Media, 2011: 32-40.

[16] Cui L L, Liu J, Zhang Y. Near-optimal control of a class of unknown nonlinear systems based on single network ADP[J]. Control and Decision, 2013, 28(9): 1423-1426.

[17] Song R Z, Lewis F, Wei Q L, et al. Multiple actor-critic structures for continuous-time optimal control using input-output data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(4): 851-865.

[18] Zhou K M, Mao J Q, Zhong Y S, et al. Robust and optimal control: Vol.138[M]. Beijing: National Defense Industry Press, 2002: 548-555.

作者简介

鲜斌(1975−), 男, 教授, 博士生导师, 从事非线性系统控制、无人机系统、实时控制系统及其应用等研究, E-mail: xbin@tju.edu.cn;

林嘉裕(1994−), 男, 硕士生, 从事旋翼无人机的自主定位与非线性控制的研究, E-mail: linjiayu@tju.edu.cn.

(责任编辑: 李君玲)