文章编号:1671-4598(2021)02-0071-05 DOI:10.16526/j.cnki.11—4762/tp.2021.02.016 中图分类号:TP273 文献标识码:A

基于强化学习的四旋翼无人机控制律设计

梁 晨,刘小雄,张兴旺,黄剑雄
(西北工业大学 自动化学院,西安 710072)

摘要:目前四旋翼无人机大部分都采用经典控制方法进行控制律的设计,然而控制参数的选择和对被控对象数学模型的依赖一直是经典控制方法设计中需要克服的问题;针对此问题,采用了一种基于深度强化学习算法 Deep Q Network 的无人机控制律设计方法,以四旋翼姿态角和姿态角速率作为三层神经网络的输入数据,最终输出动作值函数,再根据贪婪策略进行动作的选取,通过与环境的不断交互,智能体根据奖惩信息来更新神经网络的权值,使得智能体朝着获得累积回报最大值的方向选取动作;仿真结果表明在经过强化学习训练之后,四旋翼姿态角能够快速准确地跟踪上参考指令的变化,证明了基于强化学习的四旋翼无人机控制律的可行性,从而避免了传统控制方法对控制参数的选择与控制模型的依赖。

关键词:强化学习(RL);四旋翼无人机;控制律

Design of Control Law for Quadrotor UAV Based on Reinforcement Learning

Liang Chen, Liu Xiaoxiong, Zhang Xingwang, Huang Jianxiong
(College of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: At present, most of the quadrotor UAVs use the classic control method to design the control law. However, the selection of control parameters and the dependence on the mathematical model of the controlled object have always been problems that need to be overcome in the design of the classic control method. Aiming at this problem, a design method of UAV control law based on deep reinforcement learning algorithm Deep Q Network is adopted. The quadrotor attitude angle and attitude angle rate are used as the input data of the three-layer neural network, and finally the action value function is output. Then, the action is selected according to the greedy strategy. Through continuous interaction with the environment, the agent updates the weight of the neural network according to the reward and punishment information, so that the agent selects the action in the direction of obtaining the maximum cumulative return. The simulation results show that after the reinforcement learning training, the quadrotor attitude angle can quickly and accurately track the change of the reference command, which proves the feasibility of the quadrotor UAV control law based on reinforcement learning, thus avoiding the dependence of traditional control methods on the selection of control parameters and control model.

Keywords: reinforcement learning; quadrotor drone; control law

0 引言

近些年来随着科技的提升,旋翼无人机行业发展迅猛,应用场景越来越广阔,由于四旋翼具有可垂直起降、低成本和结构简单的特性,因此在公共安全、民用航拍、消防急救、农业植保以及军事领域具有十分广泛的用途。目前四旋翼无人机正在朝着易携带、多功能和更加安全高效的方向发展。

由于四旋翼是典型欠驱动非线性强耦合系统[1],四旋翼的速度和位置控制都依赖于姿态的控制,因此四旋翼的姿态控制一直是研究的热点之一[2-4]。然而四旋翼在飞行过程中容易受到环境的干扰,旋翼桨叶之间的气动干扰,存在电机快速旋转时产生的陀螺力矩以及旋翼质量分布不均等问题,这使得对四旋翼的精确建模尤为困难,从而导致依赖精确建模的传统控制算法[5]难以达到控制要求。

强化学习又称为增强学习,自从 20 世纪初便被提出来了,经过将近一个多世纪的发展,强化学习与心理学、智能控制、优化理论、计算智能、认知科学等学科有着密切的联系,是一个典型的多学科交叉领域。近些年来,得益于高速计算机、深度学习以及大数据技术的发展,强化学习得到了越来越广泛的关注,尤其是深度强化学习技术,被学术界认为是最有可能实现人工智能的算法,已经成为最受学者们关注的前沿技术之一。

2016 年 AlphaGo 成功战胜了人类顶级棋手,使得深度强化学习得到了广泛的关注,之后的 AlphaGo Zero,更是使得强化学习技术成为人工智能领域最热门的技术之一。强化学习技术不仅在博弈类游戏上取得了巨大成功,而且在控制领域也已有了新的突破,如在两轮车的控制[6]、倒

收稿日期:2020-06-13; **修回日期**:2020-07-08。
基金项目:航空科学基金资助(201905053003);陕西省飞行控制与仿真技术重点实验室资助。
作者简介:梁 晨(1995-),男,陕西大荔人,硕士研究生,主要从事无人机控制律算法方向的研究。
刘小雄(1973-),男,陕西周至人,博士,副教授,主要从事飞机飞行控制、无人机导航、制导与轨迹控制方向的研究。
引用格式:梁 晨,刘小雄,张兴旺,等.基于强化学习的四旋翼无人机控制律设计[J].计算机测量与控制,2021,29(2):71-75,86.

立摆的控制[7]上均取得了较好的进展。本文提出一种将强化学习[8]与神经网络结合起来的端到端的控制方法，该方法只关心系统的输入输出，不关心系统内部过程，通过智能体与环境的不断交互，反馈奖惩信息来优化控制参数，从而避免了对四旋翼进行精确建模等问题。该方法输入为姿态角与姿态角速率，经过神经网络，计算出四旋翼的动作值函数，再通过贪婪策略对动作进行选取，得到四旋翼各个桨叶的拉力。通过强化学习的方法对神经网络进行训练，最终使得神经网络可以收敛。最终通过在强化学习算法工具包 OpenAI Gym 中建立四旋翼的模型，用本文设计的控制算法对该模型进行仿真控制，结果证明了该算法的有效性。

1 四旋翼动力学模型的建立

如图 1 所示，本文对“X”型结构的四旋翼进行动力学模型的建立。在惯性系中应用牛顿第二定律，可得四旋翼飞行器在合外力 F 作用下的线运动和合外力矩 M 作用下的角运动方程：

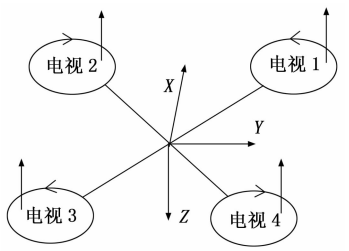


图 1 四旋翼结构图

$$\begin{cases}\sum F=\dfrac{\mathrm{d}}{\mathrm{d}t}(mV)\\ \sum M=\dfrac{\mathrm{d}}{\mathrm{d}t}(L)\end{cases} \tag{1}$$

通过对桨叶动力学模型的分析和电机模型的建立，可以求得桨叶产生的力矩、旋翼惯性反扭力矩以及陀螺效应力矩。根据机体系与地面系的旋转关系可求得欧拉角速率与机体三轴角速率的关系：

$$\begin{cases}\dot{\varphi}=p+(q\sin\varphi\_r\cos\varphi)\tan\theta\\ \dot{\theta}=q\cos\varphi-r\sin\varphi\\ \dot{\psi}=\dfrac{q\sin\varphi+r\cos\varphi}{\cos\theta}\end{cases} \tag{2}$$

当四旋翼姿态变化很小时，通过求解式（1），得到四旋翼飞行器的角运动方程：

$$\begin{cases}\ddot{\varphi}=\dfrac{1}{I_x}[U_\varphi+J_r\dot{\theta}\Omega_G+(I_y-I_z)\dot{\theta}\dot{\psi}]\\ \ddot{\theta}=\dfrac{1}{I_y}[U_\theta+J_r\dot{\varphi}\Omega_G+(I_z-I_x)\dot{\varphi}\dot{\psi}]\\ \ddot{\psi}=\dfrac{1}{I_z}[U_\varphi+(I_x-I_y)\dot{\varphi}\dot{\theta}]\end{cases} \tag{3}$$

其中：φ、θ、ψ 为滚转角、俯仰角、偏航角；p、q、r 为滚转角速率、俯仰角速率、偏航角速率；I_x、I_y、I_z 分别为四旋翼飞行器绕 x、y、z 轴的转动惯量，J_r 为每个桨叶的转动惯量；Ω_G 为陀螺力矩转速；

根据图 1 的四旋翼结构，定义 U_T、U_φ、U_θ、U_ψ 分别为四旋翼高度、滚转通道、俯仰通道以及偏航通道的控制输入，F_1、F_2、F_3、F_4 分别为 4 个桨叶提供的拉力，则有：

$$\begin{bmatrix}U_T\\ U_\varphi\\ U_\theta\\ U_\psi\end{bmatrix}=\begin{bmatrix}1 & 1 & 1 & 1\\ -d & d & d & -d\\ d & d & -d & -d\\ \dfrac{C_M}{C_T} & -\dfrac{C_M}{C_T} & \dfrac{C_M}{C_T} & -\dfrac{C_M}{C_T}\end{bmatrix}\begin{bmatrix}F_1\\ F_2\\ F_3\\ F_4\end{bmatrix} \tag{4}$$

其中：d 表示旋翼转轴到 x 轴或 y 轴的距离；C_M 为反扭力矩系数，C_T 为升力系数。

2 基于强化学习的四旋翼姿态控制结构

图 2 为强化学习的基本图，强化学习是通过智能体与环境的不断交互来更新控制策略的。一开始，智能体随机选择一个动作 A 作用于环境，环境模型通过该动作使得整个系统达到一个新的状态，并且通过回报函数给智能体一个反馈。这样不断循环下去，智能体与环境持续地交互，从而产生更多的数据样本。智能体根据与环境交互而产生的数据样本来改变自身的动作选择策略。通过不断地试错，智能体最终会学到一个最优的策略。

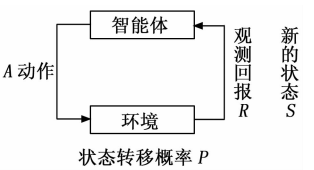


图 2 强化学习基本框图

从强化学习的基本原理中我们可以看出来，强化学习与监督学习和非监督学习有一些本质上的区别。如传统的监督学习中，数据样本是一些带有标签的静止训练集，只要样本数据之间的差异足够明显，就能够训练一个不错的模型。而强化学习则是一个连续决策的过程，在强化学习中，智能体没有直接的指导信息，而是通过环境反馈的立即回报来修正自身的策略，智能体需要不断地与环境进行交互从而实时地获取训练数据，通过这种试错的方式来获得最佳的策略。

本文采用基于值函数逼近的无模型强化学习算法 Deep Q－Network[9]来对四旋翼姿态进行控制。DQN 算法是在 Q－Learning 算法的基础上进行的改进，在 Q－Learning 算法中，维护一张表 Q－Table 它的每一列代表一个动作，每一行代表一个状态，这张表记录了每个状态下采取不同动作所获得的最大长期奖励期望，通过这张表就可以知道每一步的最佳动作是什么。但是在状态空间维度十分庞大甚至连续时，Q－Table 不能存储下所有的状态，因此 DeepMind 对 Q－Learning 进行了改进，便得到了 DQN 算法，其改进主要体现在以下几个方面：

1）DQN 利用卷积神经网络进行值函数的逼近；

2）DQN 利用了经验回放训练强化学习的学习过程；

3）DQN 中独立设置了目标网络来单独处理时间差分算法中的 TD 偏差。

DQN 算法流程如图 3 所示，在 DQN 算法中，有几个比较重要的环节：环境、当前值网络与目标值网络、动作库、经验池以及回报函数。环境模型在上一节中已经建立，其余的将在本节进行建立。

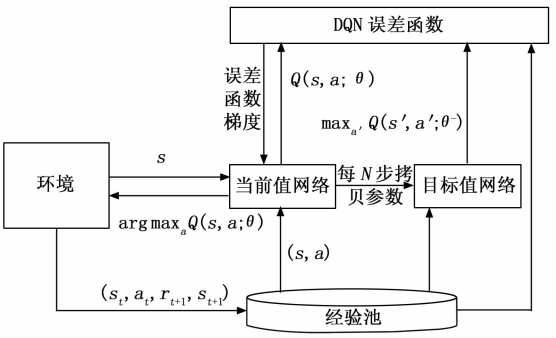


图 3　DQN 算法流程

1）动作库：本文中动作库是俯仰、滚转、偏航三通道的控制量 U_φ 、U_θ 、U_ψ 。

2）值函数神经网络：在 DQN 中，存在两个结构完全相同的逼近值函数的神经网络：一个是当前值函数神经网络，另一个是目标值函数神经网络。当前值函数与目标值函数的差作为部分依据修正神经网络的参数，当前值函数神经网络每一步都更新，而目标值函数神经网络每隔一定的步数更新。本文采用三层 BP 神经网络作为值函数神经网络，输入维度为 2，对应单个通道的角度与角速率，输出维度对应单个通道的动作库维度。

网络参数的更新如下所示：

$$\theta_{t+1}=\theta_t+\alpha[r+\gamma\max_a Q(s',a';\theta^-)-Q(s,a;\theta)]\cdot\nabla Q(s,a;\theta) \tag{5}$$

其中：θ 为网络参数；α 为学习速率；r 为立即回报；γ 为折扣因子；s 和 a 分别为状态与动作。

3）经验池：由于强化学习是建立在马尔科夫决策过程的基础上的，因此通过强化学习得到的样本数据之间存在着相关性，而神经网络的前提是样本之间独立同分布。基于此，建立一个经验池，将通过强化学习得到的数据存在经验池中，训练时从经验池中随机均匀采取一些样本进行训练，以此来打破数据样本之间存在的相关性。

4）回报函数：本文研究内容是四旋翼的姿态控制，以滚转角 φ 单通道为例，设计回报函数如下：

$$r=-(\Delta\varphi)^2-0.1\times\dot{\varphi}^2 \tag{6}$$

当角度偏差或角速度比较大时，对智能体惩罚比较大，反之则惩罚较小。强化学习会对智能体朝着使得累积回报最大的方向进行训练。

3　四旋翼姿态控制律设计

基于以上分析，设计基于无模型强化学习算法 DQN 的四旋翼控制律，模型参数如表 1 所示。

表 1　模型参数

参数	数值	单位
m	1.235	kg
d	0.1591	m
I_x	16.0×10^{-3}	kg/m^2
I_y	16.0×10^{-3}	kg/m^2
I_z	32.0×10^{-3}	kg/m^2
J	$5.695\ 3\times10^{-5}$	kg/m^2
C_T	3.13×10^{-5}	N/s^2
C_M	6.0×10^{-5}	N/s^2

具体设计如下：

四旋翼姿态控制中，一共有 6 个状态量，分别是（φ，θ，ψ，$\dot{\varphi}$，$\dot{\theta}$，$\dot{\psi}$），因此在 DQN 中，神经网络的输入也是 6 个维度。根据前文四旋翼的建模可知，四旋翼的控制输入有油门、滚转、俯仰以及偏航 4 个通道，由于油门通道对姿态的变化不产生影响，因此在姿态的控制中不考虑油门通道。通过在 Matlab 中对该模型进行仿真得到的数据可知，该模型滚转、俯仰、偏航 3 个通道的控制量取值范围分别为（－1.0，1.0）、（－1.0，1.0）以及（－0.8，0.8），因此在本实验中，将每个通道的取值范围进行 20 等分作为 3 个通道的动作库。同时为了降低训练时间，实验中将分别对 3 个通道单独进行控制律设计，最终通过式（4）反解出 F，从而达到控制四旋翼的目的。

由于三通道的控制律结构一致，因此以下仅以滚转通道为例进行介绍。依据第二章四旋翼姿态控制结构，需要建立两个结构相同但参数不同的神经网络，即目标网络和现实网络，定义目标网络的参数为 θ^-，现实网络的参数为 θ。本文采用三层 BP 神经网络对值函数进行拟合，神经网络输入层有两个神经元，分别为俯仰角偏差与俯仰角速率；隐藏层有 10 个神经元；输出层有 20 个神经元，对应俯仰通道的动作值函数。

设置经验池大小为 5 000，经验池中每一个样本存储智能体的上一步状态 s_t、当前状态 s_{t+1}、环境给予的立即回报 r 以及在当前状态下所选取的动作 a，当样本数据大于 5 000 时，可以认为前面的数据已经不具备参考价值，删除掉最早的数据。

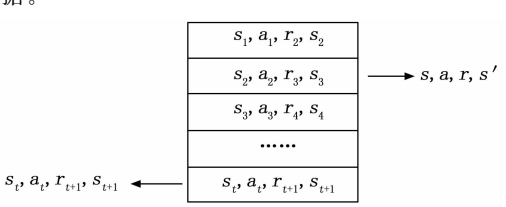


图 4　经验池

规定每次训练时从经验池中随机均匀抽取 50 条样本进行训练。因为训练时系统使用 ε 贪婪策略进行动作的选取，所以为使系统一开始具备较强的探索能力，尽最大可能探索到范围以内的所有状态，不至于收敛到局部最优，实验

中将 ε 贪婪值初始化为 0，每隔 1 000 步，贪婪值增加 0.01，最大增加到 0.95。

在实验中，给定初始角度偏差为 5°，同时初始化角速度为 0°/s，开始训练智能体。训练采取回合制，为防止智能体收敛到局部最优，为每一个回合设定最大仿真步数，当角度偏差超过设定范围或者本回合步数超过最大步数时，本回合结束，开始下一回合训练。实验中，对仿真回合不做限制，但是每一个回合将输出损失与参数模型，实验中可以随时终止训练并将模型导出。

综上所述，以贪婪策略作为滚转通道控制量的选取策略，即滚转通道控制量取值为值函数中值最大的元素所对应的动作，即：

$$U_{\varphi} = \arg\max_{a} Q(s,a;\theta) \tag{7}$$

4 仿真结果与分析

根据第 3 小节所描述的实验方法进行实验，仿真步骤如下所示：

1）初始化一些必要的参数，如动作库、贪婪值、初始角度偏差、经验池大小、每次训练所采的样本数等；

2）开始训练，训练时动作的选取采用 ε 贪婪策略。每回合训练时输出损失与参数模型，当损失达到一定要求之后，停止训练，保存参数模型。对俯仰、滚转、偏航通道分别进行如上所示的训练并保存好参数模型；

3）将参数模型输入到最终的仿真模型中，并通过式（4）反解出每个电机的拉力，从而实现对四旋翼的姿态控制。

仿真电脑 CPU 为 Intel i5－7500，内存 8.00 GB，在 ubuntu16.04 系统下，采用 OpenAI Gym 工具包进行仿真，训练效果如图 5 所示。

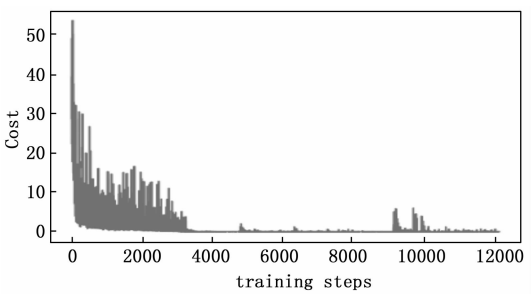


图 5 损失函数曲线

从损失函数曲线中可以看出，在强化学习的训练下，神经网络基本上从 4 000 步开始就已经收敛了，而到 10 000 步左右时出现小幅的波动，这是因为贪婪策略取值最大为 0.95，系统依然有 5%的概率随机选取动作，这会导致目标网络与现实网络之间的偏差增大，进而导致损失函数的增加。实验中，选取了第 4 000 步时的模型参数作为最终的模型参数。

将训练好的模型输入系统控制结构中，并对角度观测量加入 [0.1，0.1] 上的随机噪声，初始化 φ 为 5°，设置期望值为 0 度进行仿真实验。如图 6 所示，系统基本上能够在 1 s 以内达到控制目标，并且稳定在目标值左右。

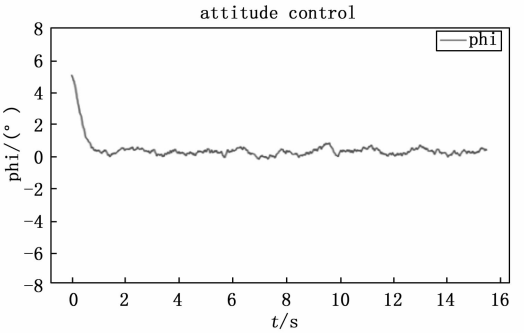


图 6 初始值为正时 $\Delta\varphi$ 的变化

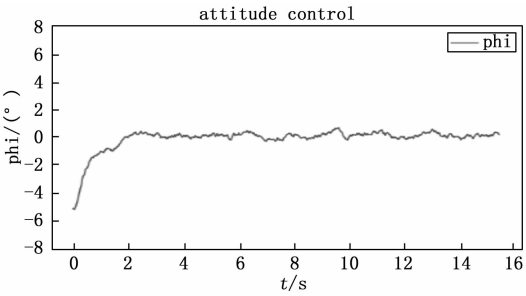


图 7 初始值为负时 $\Delta\varphi$ 的变化

将 $\Delta\varphi$ 的初始值为－5 时训练得到的模型参数，重新输入到飞机模型中，得到曲线如图 7 所示。

由以上分析，将训练好的网络模型输入到飞机模型中，控制器输入为期望角度值，将当前角度值与期望角度值做差，当角度偏差为正时通过第一个网络模型进行控制，当角度偏差为负时通过第二个网络模型进行控制。

图 8 为 φ 的控制效果图，将系统初始角度与期望角度分别设置为 5°和 0°，并在第 5 s、10 s、15 s 时分别改变系统的期望角度，可以看到，系统基本可以跟上控制指令。

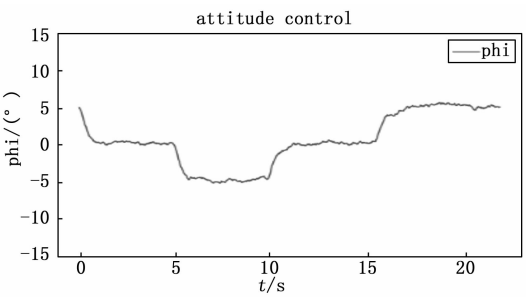
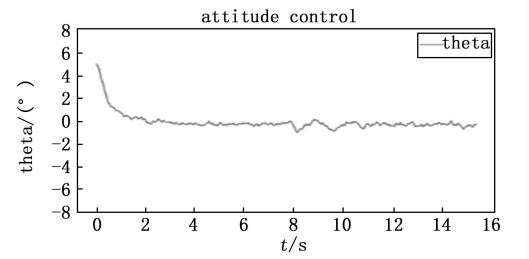


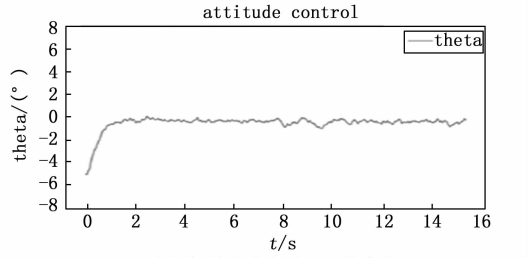
图 8 φ 的控制效果图

同理，俯仰角和偏航角的控制结构与滚转角一致，俯仰通道与偏航通道的控制效果如下：

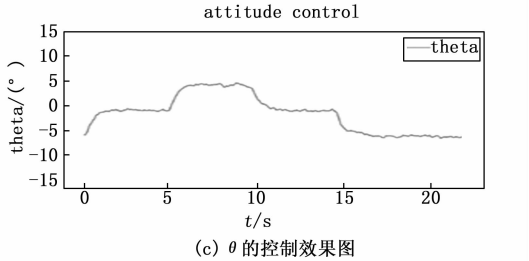
图 9、图 10 分别为俯仰角和偏航角的控制效果图。将俯仰角的初始值和期望值分别设置为－5°和 0°，然后在 5 s、10 s、15 s 时分别改变俯仰角期望值为 5°、0°和－5°，得到曲线如图 9（c）所示。同理将偏航角的初始值与期望值分别设置为 10°和 0°，然后在 5 s、10 s、15 s 时分别改变偏航角的值为 10°、0°和－10°，得到曲线如图 10（c）所示。



(a) 初始值为正时△ θ 的变化

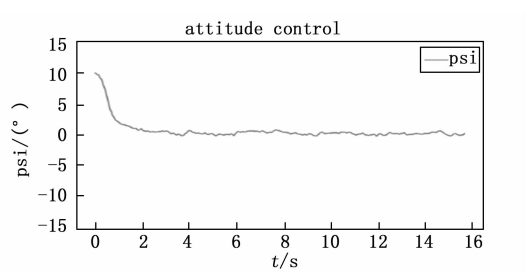


(b) 初始值为负时△ θ 的变化

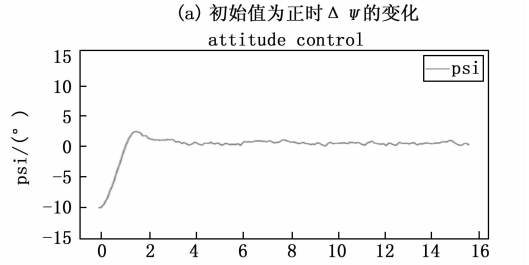


(c) θ 的控制效果图

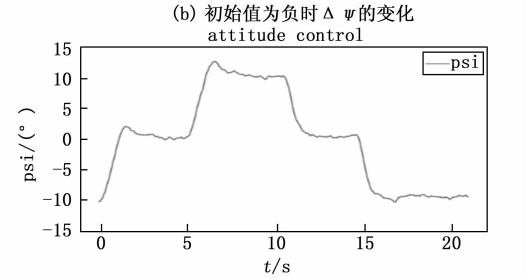
图 9　俯仰角的控制效果图



(a) 初始值为正时 Δ ψ 的变化



(b) 初始值为负时 Δ ψ 的变化



(c) ψ 的控制效果图

图 10　偏航角的控制效果图

综合以上曲线可以看出，在单独控制一个通道时，基于 DQN 的控制律基本可以快速准确地跟踪上指令信号的改变。接着将三通道的控制量通过下式求出各桨叶所提供的拉力，从而达到控制四旋翼飞行器的目的。

$$[F_{t1} \quad F_{t2} \quad F_{t3} \quad F_{t4}]^T = \boldsymbol{A}_c^T(\boldsymbol{A}_c\boldsymbol{A}_c^T)^{-1}[U_T \quad U_\varphi \quad U_\theta \quad U_\psi \quad]^T \tag{8}$$

其中：$\boldsymbol{A}_c$ 为系数矩阵。

图 11 为将各通道控制量经过控制分配之后得到的四旋翼姿态控制效果图。

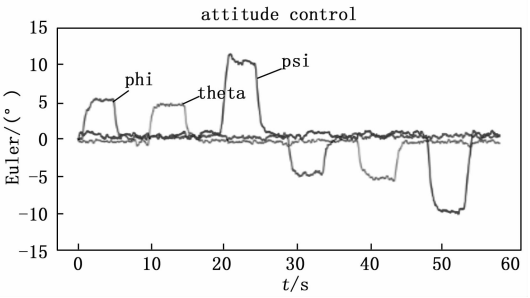


图 11　四旋翼姿态角控制效果

实验中，将四旋翼三轴姿态角均初始化为 0°，角度观测量均加入［－0.1，0.1］上的随机噪声，通过在不同时刻给定四旋翼不同的指令信号，使得四旋翼达到不同姿态，最终效果如图 11 所示。

从图 11 中可以看出，控制器基本上可以使得四旋翼的姿态跟上指令信号，但同时伴随有一定的震荡，还会有一定的误差。这是因为 DQN 中，动作空间并不是连续的，在一些状态下，智能体所需要选择的最优动作并不存在于动作空间中，这时智能体只能选择动作空间中最接近最优动作的动作，这就造成了四旋翼姿态必然会伴随有一定的震荡和误差。因此想要减小震荡并消除误差，这能增加动作空间的维度，将动作空间设置的更加稠密，这样智能体选择的动作就会更加接近最优值。

5　结束语

本文针对“X”型结构的四旋翼非线性运动模型，提出基于无模型强化学习算法 DQN 的四旋翼姿态控制律设计[10]。首先对俯仰角、滚转角以及偏航角分别进行控制律设计，当三通道控制律达到控制要求之后，通过控制分配求出 4 个桨叶的拉力，进而达到控制四旋翼姿态的目的。实验结果表明基于无模型强化学习的控制律能够在不知道被控对象模型的情况下，控制四旋翼实时跟踪上参考指令的变化。

参考文献：

[1] 张　萍. 四旋翼飞行器姿态控制建模与仿真 [J]. 电机与控制应用，2019，46 (12)：70－74.

（下转第 86 页）