基于改进强化学习的无人机规避决策控制算法

Tajmihir Islam Teethi，卢　虎，闵　欢，卞志昂
（空军工程大学信息与导航学院，陕西 西安 710077）

摘　要：针对当前无人机常用的“建图＋规划”避障方法依赖于地图，构图模型参数适应性不强等问题，将无人机自主避障问题转化为强化学习框架下的决策控制问题，提出基于改进强化学习的无人机规避决策控制算法，并设计了适用于无人机导航避障控制任务的 D3QN 结构。实验结果表明，所设计的 D3QN 结构相比经典 DQN 结构可提升约 25％的训练效率，经过训练之后的 D3QN 网络能根据视觉信息做出可靠的规避控制决策，能用于指导无人机在未知无图等典型场景中进行无碰撞的探索飞行或导航飞行。

关键词：视觉避障；强化学习；马尔可夫决策；深度 Q 网络
中图分类号：TP29　　**文献标志码**：A　　**文章编号**：1008-1194(2022)03-0068-06

An Improved Reinforcement Learning Method for Drone Avoidance Decision Control

Tajmihir Islam Teethi, LU Hu, MIN Huan, BIAN Zhiang,
(College of Information and Navigation, Air Force Engineering University, Xi'an 710077, China)

Abstract: In order to solve the problem that the current "map + planning" obstacle avoidance method commonly used by drones relies on the map and the lack of adaptability of the composition model parameters, the paper transformed the problem of autonomous obstacle avoidance by drones into a decision-making control problem under the framework of reinforcement learning. The paper proposed a drone avoidance decision control algorithm based on improved reinforcement learning, and designed a D3QN (dueling double deep Q-learning network, D3QN) structure suitable for drone navigation and obstacle avoidance control tasks. Experimental results showed that the designed D3QN structure could improve training efficiency by about 25% compared with the classic DQN structure. After training, the D3QN network could make reliable evasion control decisions based on visual information. The research results of the paper could be used for the collision-free exploration flight or navigation flight of the drones in typical scenes such as unknown and imageless.

Key words: visual obstacle avoidance; reinforcement learning; Markov decision, deep Q network

0　引言

无人机自主飞行、自主导航是提高无人系统智能化水平的重要基础。在与真实世界的交互过程中，避障是无人机应当具有的最基本的功能之一。目前，很多无人机厂商，如中国的大疆、法国的 PARROT 等公司也都将自主避障能力作为其无人机产品的一项重要技术指标。当前，无人机的自主避障主要是通过机载传感器获取障碍物的距离、位置、速度等有效信息，再根据障碍物信息自主规划出合理的路径，从而保证其在运行的过程中避开障碍[1-2]。

传统的无人机自主避障技术主要由障碍感知与避障规划两大功能模块组成。障碍感知是指无人机通过机载传感器实时获取周边障碍物的信息。避障所常用的传感器主要包括超声波传感器、激光雷达、双目视觉传感器等。SLAM(simultaneous localization and mapping)技术可以为避障提供更加丰富全面的地图环境信息，因此在近几年的研究中，SLAM 技术被广泛应用于移动机器人自主导航避障[3-4]。但基于 SLAM 的避障

＊ 收稿日期：2022-01-27
作者简介：Tajmihir Islam Teethi(1988—)，女，孟加拉国达卡人，硕士研究生。

方法仍需要手动调试大量的构图模型参数以达到良好的建图效果和可靠的路径规划,且在一架无人机上调试好的一套构图参数,由于平台、传感器载荷等的性能差异并不完全适用于另一架无人机。

当前,人工智能正在飞速发展,基于强化学习的避障方法通过训练深度神经网络进行端到端动作决策,使得无人机避障无须"额外"的建图过程,取而代之的是一种即时自主的行为,并且避免了复杂的建模和大量的参数调整,而且因其不需要建图的特性,此类基于学习的避障方法也能更好地适应于未知无图的应用场景[5-6]。

1 强化学习的马尔可夫决策表示

强化学习作为机器学习的一大分支[7],其基本思想是,智能体在完成某项任务时,通过动作与环境进行交互,环境在动作的作用下会返回新的状态和奖励值,奖励值越高说明该动作越好,反之则说明该动作应该被舍弃,经过数次迭代之后,智能体最终会学到完成某项任务的最优策略。强化学习基本的原理框架如图 1 所示。

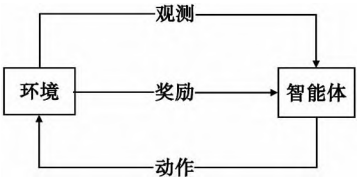


图 1 强化学习基本原理框架

Fig. 1 The principle framework of reinforcement learning

通常的强化学习问题可用马尔可夫决策过程(Markov decision processes,MDPs)来描述,马尔可夫决策过程由一个五元组$\langle S,A,P,R,\gamma\rangle$构成,其中 S 和 A 分别代表智能体的状态集和动作集;$\boldsymbol{P}$ 为状态转移矩阵,定义 $P_{ss'}^{a}=\boldsymbol{P}[S_{t+1}=s'\mid S_t=s,A_t=a]$,可以理解为智能体在当前时刻实施动作 $A_t=a$ 之后,智能体从当前状态 $S_t=s$ 转移到下一时刻状态 $S_{t+1}=s'$ 的概率为 $P_{ss'}^{a}$,式中的小写字母表示具体的动作或状态,$s,s'\in S,a\in A$;R 为奖励函数,$R(s,a)$ 表示智能体在状态 s 下执行动作 a 获得的奖励值;γ 为折扣因子,用于描述智能体关注长远奖励的程度,γ 的取值范围为 $\gamma\in(0,1]$,γ 的取值越大,智能体越能考虑长远的利益。如图 2 所示,通常采用马尔可夫链来表示一个马尔可夫决策过程。

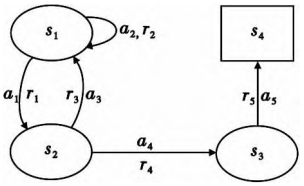


图 2 马尔可夫决策过程示例

Fig. 2 An example of Markov decision process

在强化学习中,从初始状态 S_1 到终止状态的序列过程 $S_1,S_2,\cdots,S_T$,被称为一个片段,一个片段的累积奖励定义为式(1),式中,r_τ 为智能体在 τ 时刻从环境获得的即时奖励值,T 为智能体达到终止状态时的时刻

$$G_t=\sum_{\tau=t}^{T}\gamma^{\tau-t}r_\tau。\tag{1}$$

除此之外,强化学习还在马尔可夫决策过程的基础上,定义了智能体的策略 $\pi(a\mid s)$,策略 π 表示的是智能体在状态 s 下的动作的概率分布,其定义为:

$$\pi(a\mid s)=P[A_t=a\mid S_t=s]。\tag{2}$$

强化学习的目的就是通过不断试错来改善智能体的策略 $\pi(a\mid s)$,以最大化其获得的累积奖励,因此引入了值函数来评价某个策略获得的累积奖励。一般来说值函数分为两种:状态值函数(V 函数)和状态动作值函数(Q 函数)。V 函数的定义是,从状态 s 开始,使用策略 π 得到的期望奖励值,其定义式

$$V_\pi(s)=\mathrm{E}[G_t\mid S_t=s]。\tag{3}$$

Q 函数的定义为,从状态 s 开始,执行动作 a,然后使用策略 π 得到的期望奖励值,其定义式

$$Q_\pi(s,a)=\mathrm{E}[G_t\mid S_t=s,A_t=a,\pi]。\tag{4}$$

最终得到 V 函数的贝尔曼期望方程

$$V_\pi(s)=\mathrm{E}[r_t+\gamma V_\pi(S_{t+1})\mid S_t=s]。\tag{5}$$

贝尔曼期望方程将 V 函数的求取分为了两部分,一部分是当前的即时奖励 r_t,另一部分是后继状态 S_{t+1} 的 V 值。同理,也可以推导出 Q 函数的贝尔曼期望方程

$$Q_\pi(s,a)=\mathrm{E}[r_t+\gamma Q_\pi(S_{t+1},A_{t+1})\mid S_t=s,A_t=a,\pi]。\tag{6}$$

定义最优值函数为所有策略中最大的值函数,即

$$V_*(s)=\max_\pi V_\pi(s),\tag{7}$$

$$Q_*(s,a)=\max_\pi Q_\pi(s,a)。\tag{8}$$

2 基于改进强化学习的无人机规避决策控制算法

2.1 无人机视觉避障的马尔可夫决策模型

无人机视觉避障的强化学习问题可以表述为无

人机通过视觉传感器与环境交互的马尔可夫决策过程：无人机获取当前时刻 t 的视觉图像 s_t，根据策略 $\pi(a|s)$ 执行动作 a_t，观测环境反馈的奖励值 r_t，然后转移到后继状态 s_{t+1}，其中 $t\in(0,T]$，$a_t\in A$，A 为智能体的动作集，T 为每个交互片段的终止时刻。

强化学习算法通过改善智能体的行为策略 $\pi(a|s)$，最大化其获得的累积奖励 $G_t=\sum_{\tau=t}^{T}\gamma^{\tau-t}r_\tau$，式中折扣因子 $\gamma\in(0,1]$，γ 的取值越大，智能体越能考虑长远的利益，训练难度也越大。根据第二章强化学习基础，在策略 $\pi(a|s)$ 下，用于评估动作好坏的状态动作值函数(Q 函数) 被定义为：

$$Q_\pi(s_t,a_t)=\mathrm{E}[G_t|s_t,a_t,\pi]。\tag{9}$$

根据贝尔曼期望方程，当前 Q 值可以进一步通过当前奖励和后继状态的 Q 值求出：

$$Q_\pi(s_t,a_t)=\mathrm{E}[r_t+\gamma Q_\pi(s_{t+1},a_{t+1})|s_t,a_t,\pi]。\tag{10}$$

智能体的动作决策依据是每个动作的最优 Q 值：

$$Q_*(s_t,a_t)=\max_\pi Q_\pi(s_t,a_t)=\max_\pi \mathrm{E}[G_t|s_t,a_t,\pi]。\tag{11}$$

因此，将 Q 函数的贝尔曼期望方程进一步转化为贝尔曼最优方程的形式，即当前的最优 Q 值可以通过当前奖励和后继状态的最优 Q 值中的最大值求出：

$$Q_*(s_t,a_t)=\mathrm{E}[r_t+\gamma\max_{a_{t+1}}Q_*(s_{t+1},a_{t+1})|s_t,a_t,\pi]。\tag{12}$$

在求得每个状态动作对 (s_t,a_t) 的最优 Q 值之后，智能体便可以在不同的输入状态下进行最优动作决策，从而生成最优策略 $\pi_*(a|s)$，其决策的核心思想是贪婪思想，即选择输入状态下最大的最优 Q 值所对应的动作作为最优动作：

$$\pi_*(a|s)=\begin{cases}1,a=\mathrm{argmax}_{a\in A}Q_*(s,a)\\0,\text{其他}\end{cases}。\tag{13}$$

2.2 深度 Q 网络的改进方法

经过 2.1 节的建模分析，我们将无人机的自主避障问题转化为了求取每个状态动作对的最优 Q 值问题，在基于 Q 值函数的强化学习算法中，求取最优 Q 值可以依靠 Q 学习算法[8]。但是 Q 学习算法无法很好地解决状态空间维度爆炸的问题，于是研究者们提出了 DQN(deep Q network)，采用深度神经网络逼近求取 Q 值[9]，即：$Q(s,a)=Q(s,a,\theta)$，式中 $Q(s,a,\theta)$ 是深度神经网络的预测 Q 值，θ 为网络权重参数，再结合 Q-learning 算法的迭代更新思想，使得神经网络预测的 Q 值不断逼近最优 Q 值：$Q_*(s,a)=Q_*(s,a,\theta)$。DQN 算法虽然很好地解决了 Q-learning 算法面临的状态空间维度过大的问题，但是其自身也存在过估计和训练速度慢的问题，因此在后续的发展中，文献[10] 提出采用 double Q-learning 方法评估值函数，用于解决 DQN 的过估计问题，被称为 DDQN(double DQN)；文献[11] 提出竞争网络结构(dueling network) 来加速 DQN 的训练过程。

在强化学习问题中，状态是智能体选择动作的重要依据，状态的设置可以是智能体对环境的观测，也可以是智能体的自身状态。在无人机避障过程中，无人机需要感知与障碍物之间的距离，因此选择无人机视觉传感器采集的深度图作为无人机的状态。

为了能使无人机更好地做出合理的决策，设计无人机的状态为连续抓取深度图组成的深度图堆，如图 3 所示。这样设计的好处在于使状态中既包含了深度信息又隐含了无人机的运动信息，考虑到无人机运行时的实时性，最终决定以连续抓取 4 帧深度图来组成一个深度图堆。

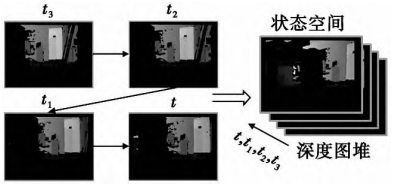


图 3 无人机状态空间设置

Fig. 3 State space configuration of drone

动作空间是无人机能够执行的具体动作，为了使网络经过训练能得到更加可靠的避障策略，需要以无人机能及时规避障碍为目标，合理地设计无人机的动作空间。本文所设计的离散动作空间如图 4 所示，分为前进和转向两大动作组。前进动作组控制无人机的前进速度，其中包含快速前进和慢速前进 2 个动作：$v\in(4,2)$ m/s。转向动作组控制无人机偏航角速率，其包含快速左转、左转、停止转向、右转、快速右转 5 个动作：$\mathrm{yaw}_{\mathrm{rate}}\in(\pi/6,\pi/12,0,-\pi/12,-\pi/6)$ rad/s。动作空间总共包含 7 个动作，通过前进动作和转向动作的组合共能生成 10 种动作指令，基本包含了无人机常见的机动方式。

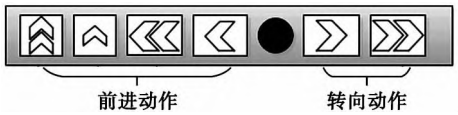


图 4 无人机离散动作空间

Fig. 4 Discrete action space of drone

为了提高训练过程的稳定性和学习效率,本节结合 double Q-learning[10] 和 dueling network[11] 方法,设计了用于无人机视觉避障的 D3QN(dueling double DQN)网络,如图 5 所示。

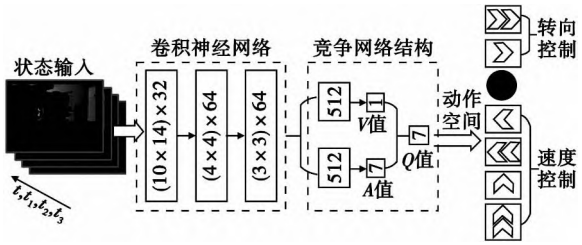


图 5 无人机视觉避障 D3QN 网络结构

Fig. 5 D3QN network structure for drone visual obstacle avoidance

本文所设计的 D3QN 网络的输入是连续 4 帧的深度图,尺寸为 160×128×4,经过 3 层卷积神经网络提取特征后,按照 dueling network 分为两个数据流,再通过两层全连接层,网络的最终输出是动作空间内各个动作的 Q 值。网络的损失函数为:

$$L=(Y_t^{\mathrm{DDQN}}-Q(s_t,a_t,\theta))^2。\tag{14}$$

网络规避训练算法如下:

1:初始化在线网络权重参数 θ,初始化目标网络权重参数 $\theta^-=\theta$;

2:初始化记忆回放单元 D;

3:For episode=1, M do;

4:读取初始状态 s_t;

5:For $t=1$, T do;

6:计算当前状态下所有动作的 Q 值 $Q(s_t,a,\theta),a\in A$;

7:根据小概率 ε 选择随机动作 $a_t\in A$,否则选择动作 $a_t=\mathrm{argmax}_{a\in A}Q(s_t,a,\theta)$;

8:无人机执行动作 a,观测奖励值和后继状态 r_t 和后继状态 s_{t+1};

9:将五元组 $\{s_t,a_t,r_t,s_{t+1},\mathrm{reset}\}$ 存入 D,reset 用于判断 s_{t+1} 是否终止状态;

10:状态转移 $s_t=s_{t+1}$;

11:从记忆回放单元随机采样 n 个样本数据 $\{s_t,a_t,r_t,s_{t+1},\mathrm{reset}\}_i,i=1,\cdots,n$;

12:计算 $Y_t^{\mathrm{DDQN}}=$

$$\begin{cases}r_t,\text{if reset=True}\\ r_t+\gamma Q(s_{t+1},\mathrm{argmax}_{a_{t+1}\in A}Q(s_{t+1},a_{t+1},\theta),\theta^-),\text{otherwise}\end{cases};$$

13:计算 $L=(Y_t^{\mathrm{DDQN}}-Q(s_t,a_t,\theta))^2$,执行梯度下降法更新在线网络参数 θ;

14:每 C 步更新目标网络参数 $\theta^-=\theta$。

3 实验验证与分析

3.1 仿真平台搭建

为了验证本文所提出的视觉自主避障算法的可行性与有效性,在 AirSim 仿真平台[12] 上开展了无人机避障仿真实验。

无人机视觉自主避障的训练环境,为 40 m×40 m×30 m 的方盒世界,如图 6 所示,其全局坐标系位于方盒的中心,无人机的初始位置设置于坐标系的原点,然后在其中布置了三种不同形状的障碍物,在训练环境中以算法训练无人机感知障碍、规避障碍的能力。随后搭建了如图 7 所示的泛化测试环境,测试场景 1 在训练环境的基础上,对原来的 3 个柱形障碍物进行了移动,测试场景 2 则是在方盒世界中加入了更多的障碍物。

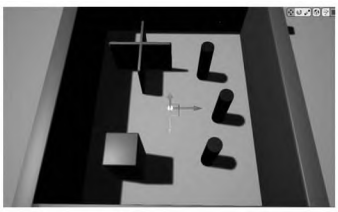


图 6 无人机避障训练环境

Fig. 6 UAV obstacle avoidance training environment

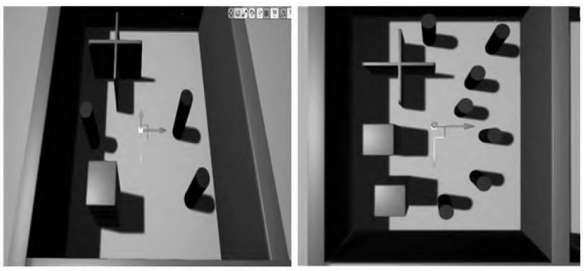


图 7 泛化测试场景

Fig. 7 Generalization test scenario

为了测试基于 D3QN 的无人机导航避障算法能力，搭建了如图 8 所示的未知测试场景：无人机按照从初始位置→目标位置的路径执行多航点任务，导航途中面临多个障碍物，以模拟复杂城市低空复杂场景。

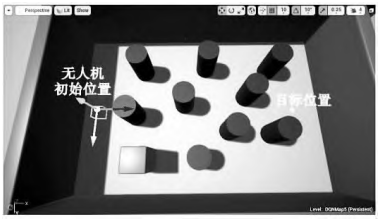


图 8 未知测试场景

Fig. 8 Unknown test scenario

3.2 网络性能对比

首先，为了分析所设计的 D3QN 网络的优势，分别采用了 D3QN、DDQN、DQN 三种不同网络在训练环境中进行训练，网络训练的硬件条件为 CPU：2.70 GHz×8，GPU：RTX2080ti 11 GB，三种网络训练的总片段数均设置为 1 000，每次从记忆回放单元采样 32 个样本数据进行梯度下降，训练过程中的奖励值曲线如图 9 所示。

可以看出，D3QN 模型最先开始收敛（约 600 片段），DDQN 和 DQN 收敛较慢（约 800 片段），D3QN 的训练速度相比 DDQN 和 DQN 提升了约 25%，并且平均每个片段的累积奖励高于 DDQN 和 DQN 模型；DDQN 相比于 DQN，两者的收敛速度相差不大，但 DDQN 的平均奖励水平高于 DQN。这可能是由于 D3QN 和 DDQN 模型都运用了 Double Q-learning，改善了 DQN 的过估计问题，给予智能体更多的探索机会，使其能够获得更高的奖励值。综合对比来看，D3QN 模型的训练效率最高，达到了预期的改进效果。

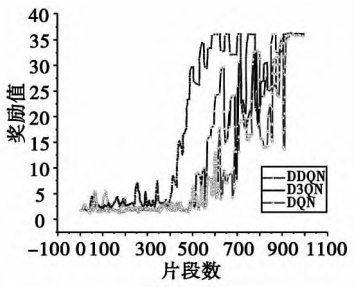
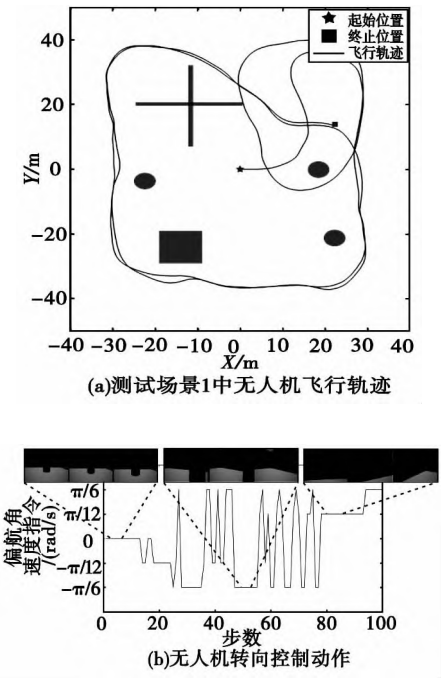


图 9 不同网络训练奖励曲线对比

Fig. 9 Comparison of reward curves for different network training

3.3 规避决策控制算法泛化测试

为了进一步测试训练好的 D3QN 网络的泛化性能，接下来分别在泛化测试场景 1 和 2 中加载 D3QN 网络模型，并运行算法进行实际飞行测试，简明起见仅给出场景 1 的飞行过程中记录的运动轨迹以及无人机的转向控制动作，如图 10 所示。



(a)测试场景1中无人机飞行轨迹

(b)无人机转向控制动作

图 10 泛化测试结果

Fig. 10 Generalization test results

在场景 1 的泛化测试中，无人机没有事先对环境构建全局地图，由记录的运行轨迹可以看出，无人机在未知的新环境下也能进行无碰撞的自主飞行，通过学习得到的规避决策能力具有较好的自适应性，在图 10(b)记录的转向控制动作中虽然出现了较多的跳变现象，但是其不影响整体的避障性能。

综上可以看出，本文算法训练出的避障策略，对环境的改变具有较好的自适应能力，训练后的 D3QN 网络也表现出了较好的泛化性能。

3.4 导航避障算法测试

为了进一步测试把规避决策应用到具体任务中的表现性能，在搭建的图 8 的未知测试场景中对算法进行了测试，测试结果如图 11 所示。

从仿真测试结果可以看出，无人机在执行航点导航任务的过程中，能够判断出前方是否存在障碍，并能做出合理决策，及时进行规避，在避开障碍之后继续朝着设置的航点飞行。

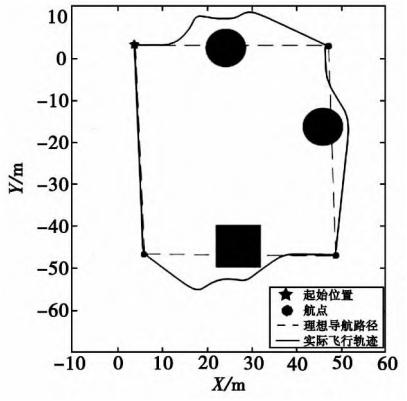


图 11 飞行轨迹(多航点任务)导航避障算法测试飞行轨迹

Fig. 11 Flight trajectory (multi-waypoint task) navigation obstacle avoidance algorithm test flight trajectory

相比于基于地图和规划的避障方法,基于学习的避障方法直接根据图像作出相应决策,其优势在于不依赖地图,能较好地适用于未知无图的环境。但是在建图良好的情况下,基于地图和规划的避障方法可以依据规划好的路径,控制无人机以尽可能快的速度飞行,其动作连续,控制更加精准,最大运行速度可达 10 m/s。而在本章的仿真实验中,考虑到图像处理、网络计算量以及动作空间的离散性,无人机自主飞行的最大速度为 4 m/s。

4 结论

无人机规避决策的导航控制问题是无人机的核心技术之一,论文研究成果有助于进一步完善无人机智能化、集群化的相关算法与技术,提升无人机中低空飞行的导航控制性能。但本文提出的基于改进强化学习的无人机视觉避障算法,仍有较大性能提升空间。如文中提出的 D3QN 网络只能输出离散动作空间,且只适用于旋翼无人机,为了进一步提升避障控制的精准程度以及算法的通用性,还应当研究旋翼无人机和固定翼无人机飞行控制的共性与区别,设置维度更大的动作空间来组合形成不同的运动模式,或是改用基于策略梯度的深度强化学习算法学习连续化的避障策略;其次是所提避障算法从仿真到真实环境的泛化问题,在仿真器中训练无人机避障时,仿真器所提供的深度图过于理想,不存在任何噪声,实践中应对其进行加噪声处理,从而使仿真器提供的环境更加逼近真实环境。上述问题都有待进一步的深入研究、技术拓展并逐步完善。

参 考 文 献:

[1]MYR-ARTAL R,MONTIEL J, M, M,TARDOS J D. ORB-SLAM:A versatile and accurate monocular SLAM system [J]. IEEE Transactions on Robotics: A publication of the IEEE Robotics and Automation Society,2015,31(5) :1147-1163.

[2]袁瑞廷 . 多旋翼无人机的自主避障、目标跟踪及自主导航定位研究[D]. 南京:南京航空航天大学,2019.

[3]蒋小强,卢虎,闵欢 . 基于连续-离散 MRF 图模型的鲁棒多机器人地图融合方法[J]. 机器,2020,42(1):49-59.

[4]谢文光,吴康,阎芳,等 . 一种面向多无人机协同编队控制的改进深度神经网络方法[J]. 西北工业大学学报,2020,38(2):295-302.

[5]TAI Lei, PAOLO G, LIU Ming. Virtual-to-real deep reinforcement learning: continuous control of mobile robots for mapless navigation[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017.

[6]CEBOLLADA S, LUIS P, MARIA F, et al. A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data[J]. Expert Systems with Applications, 2021, 167: 114195.

[7]SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[C]// Proceedings of the 32nd International Conference on Machine Learning. PMLR, 2015:1889-1897.

[8]闵欢,卢虎,史浩东 . 采用深度神经网络的无人机蜂群视觉协同控制算法[J]. 西安交通大学学报,2020,54(9):173-179.

[9]MNIH VOLODYMYR KAVUKCUOGLU K, SILVER D, et al . Human-level control through deep reinforcement learning[J]. Nature,2015,518(7540) :529-533.

[10]DURYEA E ,GANGER M, HU W. Exploring deep reinforcement learning with multi Q-learning [J]. Intelligent Control and Automation, 2016(7): 129-144.

[11]CHEN Xiliang, CAO Lei, LI Chenxi, et al. Network architecture for deep reinforcement learning[J]. Mathematical Problems in Engineering, 2018.

[12]SHAH S ,DEY D ,LOVETT C ,et al. AirSim: High-fidelity visual and physical simulation for autonomous vehicles[M]. Cham, Switzerland: Springer, 2017.