基于 LSTM-PPO 算法的无人作战飞机近距空战机动决策

丁　维，王　渊，丁达理，谢　磊，周　欢，谭目来，吕丞辉
（空军工程大学航空工程学院，西安，710038）

摘要　近距空战中环境复杂、格斗态势高速变化，基于对策理论的方法因数据迭代量大而不能满足实时性要求，基于数据驱动的方法存在训练时间长、执行效率低的问题。对此，提出了一种基于深度强化学习算法的 UCAV 近距空战机动决策方法。首先，在 UCAV 三自由度模型的基础上构建飞行驱动模块，形成状态转移更新机制；然后在近端策略优化算法的基础上加入 Ornstein-Uhlenbeck 随机噪声以提高 UCAV 对未知状态空间的探索能力，结合长短时记忆网络(LSTM)增强对序列样本数据的学习能力，提升算法的训练效率和效果。最后通过设计 3 组近距空战仿真实验，并与 PPO 算法作性能对比，验证所提方法的有效性和优越性。

关键词　无人作战飞机；空战机动决策；深度强化学习；近端策略优化；长短时记忆网络
DOI　10.3969/j.issn.1009-3516.2022.03.004
中图分类号　V271.4　　**文献标志码**　A　　**文章编号**　1009-3516(2022)03-0019-07

Maneuvering Decision of UCAV in Close Air Combat Based on LSTM-PPO Algorithm

DING Wei, WANG Yuan, DING Dali, XIE Lei, ZHOU Huan, TAN Mulai, LYU Chenghui
(Aviation Engineering School, Air Force Engineering University, Xi'an 710038, China)

Abstract　With the increasing military application of unmanned combat aircraft (UCAV), unmanned combat will become the main combat mode in the future air battlefield. In close-range air combat, the environment is complex and the combat situation changes rapidly. The method based on game theory cannot meet the real-time requirements due to the large amount of data iteration, and the data-driven method has the problems of long training time and low execution efficiency. To solve this problem, a UCAV maneuver decision method based on deep reinforcement learning algorithm is proposed in this paper. Firstly, the flight drive module is constructed on the basis of UCAV three-degree-of-freedom model to form the state transition updating mechanism. Then, on the basis of PPO algorithm, ornstein-uhlenbeck (OU) random noise was added to improve UCAV's ability to explore unknown state space, and LSTM was combined to enhance UCAV's ability to learn sequence sample data, so as to improve the training efficiency and effect of the algorithm. Finally, the effectiveness and superiority of the proposed method are verified by designing three groups of close-range air combat simulation experiments and comparing the performance with PPO

收稿日期：2021-11-20
基金项目：陕西省自然科学基金(2020JQ-481)
作者简介：丁　维(1996—)，男，安徽铜陵人，硕士生，研究方向为无人飞行器作战系统与技术。E-mail:3271378690@qq.com.

引用格式：丁维，王渊，丁达理，等. 基于 LSTM-PPO 算法的无人作战飞机近距空战机动决策[J]. 2022, 23(3): 19-25. DING Wei, WANG Yuan, DING Dali, et al. Maneuvering Decision of UCAV in Close Air Combat Based on LSTM-PPO Algorithm[J]. Journal of Air Force Engineering University (Natural Science Edition), 2022, 23(3): 19-25.

algorithm.

Key words unmanned combat aerial vehicles; air combat maneuver decision; deep reinforcement learning; proximal policy optimization; short and long duration memory network

随着无人作战飞机(unmanned combat aerial vehicles,UCAV)的自主化、智能化水平不断提高,由其自主完成空战任务获取制空权已成为未来战场发展的必然趋势。其中,空战机动决策方法一直是自主空战领域研究的重要一环[1-2]。目前无人机空战机动决策常用方法主要分为两类,一类是基于对策理论的方法,另一类是基于数据驱动的方法。基于对策理论的方法应用在近距空战机动决策上主要有微分对策法[3]、矩阵对策法和影响图法[4],基于数据驱动的近距空战机动决策方法主要有神经网络及强化学习方法。

文献[5]将微分对策法应用于空战追逃问题,构建了微分对策模型,现阶段虽然应用较为广泛,但其计算量太大、实时性差,且其目标函数设定非常困难,因此不适用于复杂的空战环境;文献[6]应用矩阵对策法获得我机最优选择策略的大致范围,虽然算法容易理解,但是其结果精度不高且实时性较差,因此较难应用于无人机自主空战中;文献[7]将影响图法应用于机动决策,虽然能有效引导 UCAV 战斗,但是模型结构复杂,计算繁琐且实时性较差,很难求解出较复杂的决策问题。对于基于数据驱动空战机动决策方法而言,文献[8]应用神经网络方法,虽然鲁棒性强、实时性好,但是需要大量样本进行训练且产生的数据不真实;文献[9]运用强化学习方法由环境反馈出的信息来展开学习,虽然无需提供训练样本,但是却存在训练时间长、执行效率低的缺点。

针对 UCAV 近距空战机动决策问题,本文首先在 UCAV 三自由度模型的基础上构建飞行驱动模块,以此来实现深度强化学习过程中与环境的不断交互,并形成一种状态转移更新机制。在算法层面,针对现有常用方法存在的无法满足实时性、收敛速度慢、容易陷入局部最优等不足,本文以近端策略优化(proximal policy optimization,PPO)算法[10]为基础,充分发挥神经网络离线训练的可塑性和在线使用的实时性,通过引入 OU 随机噪声进一步提升算法在训练过程中的探索性能,引入长短时记忆网络(long short term memory,LSTM)[11]将空战状态转化为高维感知态势,加强网络对时序性空战数据的学习能力,从而提出基于长短时记忆-近端策略优化(long short term memory-proximal policy optimization, LSTM-PPO)算法的 UCAV 近距空战机动决策方法。通过设计不同的近距空战仿真实验,并与 PPO 算法作性能对比,验证该方法的有效性和优越性。

1 空战环境设计

1.1 UCAV 三自由度模型设计

UCAV 三自由度模型是对 UCAV 运动状态的具体描述,为了降低控制量之间的耦合关系,并充分考虑平台气动特性对飞行状态的影响,使模型更加贴近实际,飞行轨迹更为真实,增加其工程利用价值,其三自由度质点运动、动力学模型如下:

$$\begin{cases} x=v\cos\gamma\cos\psi \\ y=v\cos\gamma\sin\psi \\ z=v\sin\gamma \\ v=\dfrac{T\cos\alpha-D}{m}-g\sin\gamma \\ \gamma=\dfrac{(L+T\sin\alpha)\cos\mu}{mv}-\dfrac{g}{v}\cos\gamma \\ \psi=\dfrac{(L+T\sin\alpha)\sin\mu}{mv\cos\gamma} \end{cases} \tag{1}$$

式中:(x,y,z)分别代表速度 v 在坐标系各个轴上的分量;γ 为航迹倾角;ψ 为偏航角;m 为 UCAV 总体质量;g 为重力加速度;(α,μ,T)为模型的控制量,分别表示当前时刻 UCAV 的攻角、滚转角及推力;L 和 D 分别表示升力参量和阻力参量,具体可以表示为:

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

式中:S 为气动参考面积;$\rho=1.225\mathrm{e}^{\frac{h}{9300}}$代表大气密度;$C_L$ 和 C_D 分别代表升力系数和阻力系数,其值可以通过拟合来得到。

1.2 状态转移更新机制设计

为实现算法与空战环境不断交互,从而输出控制量对 UCAV 的运动进行控制,将上述三自由度模型设置成飞行驱动模块。假设 UCAV 与敌机使用相同的平台模型,通过飞行驱动模块实现敌我双方空战状态的更新,即通过当前时刻状态与控制量实时计算出下一时刻 UCAV 与敌机所处的新状态,以此形成一种状态转移更新机制,见图 1。

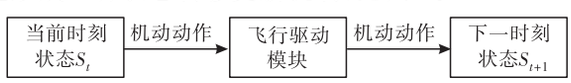

图 1 状态转移更新机制

1.3　奖励函数设计

基于强化学习的近距空战机动决策的目标是找到一个最优机动策略使 UCAV 完成攻击占位，从而使完成当前任务的累计奖励最大。奖励是评价策略的唯一量化指标，决定智能体最终学到策略的优劣，并直接影响算法的收敛性和学习速度。UCAV 通过深度强化学习进行空战决策时，除完成任务的奖励外，中间过程无法获得奖励，存在着稀疏奖励[12]的问题，因此在复杂的空战任务中不仅需要设计完成任务的胜负奖励，对于每一回合中每一步的辅助奖励设计也至关重要。为了有助于验证算法的有效性，本文以机动决策难度较大的使用近距空空导弹后半球攻击策略为例，分别设计角度、高度、距离奖励函数。

1.3.1　角度奖励函数

空战过程中，不同的角度奖励体现了 UCAV 使用不同的机载武器对应目标的前半球、后半球或全向攻击等不同的战术战法。综合考虑 UCAV 的进入角 q_u 和目标的进入角 q_t 以及导弹最大离轴发射角 $\theta^u_{\max}$，设计后半球攻击策略角度奖励函数 r_A 如下：

$$r_A=\begin{cases}1, q_u\leqslant\theta^u_{\max}\text{且}\dfrac{\pi}{2}\leqslant q_t\\ 0, q_t\leqslant\dfrac{\pi}{2}\\ -1,\text{其他}\end{cases}\tag{3}$$

1.3.2　距离奖励函数

两机的相对距离 R 也是导弹发射条件以及双方态势影响因素之一。距离奖励函数的设定要考虑 UCAV 机载武器最大发射距离 $L^u_{\max}$ 以及最小发射距离 $L^u_{\min}$，其可根据文献[13]中的方法实时解算出来。距离奖励函数 r_R 设定如下：

$$r_R=\begin{cases}1, L^u_{\min}\leqslant R\leqslant L^u_{\max}\\ -1,\text{其他}\end{cases}\tag{4}$$

式中：相对距离

$$R=\sqrt{(x_e-x_u)^2+(y_e-y_u)^2+(z_e-z_u)^2}$$

1.3.3　高度奖励函数

高度奖励的设置应充分考虑不同武器的作战性能，主要体现为通过高度奖励使 UCAV 与敌机的高度差保持在理想范围内，充分发挥武器性能。设计高度奖励函数 r_H 如下：

$$r_H=\begin{cases}1, \Delta H_{\text{down}}\leqslant\Delta H\leqslant\Delta H_{\text{up}}\\ -1,\text{其他}\end{cases}\tag{5}$$

式中：ΔH 代表 UCAV 与目标的相对高度；ΔH_{up} 和 ΔH_{down} 分别表示理想高度差的上下限。

1.3.4　胜负奖励函数

空战胜负判定主要分为 3 种情况：①飞行高度过低导致坠毁；②态势占据劣势被敌机击中回合失败；③占据态势优势满足导弹发射条件，空战胜利。

胜负回报奖励函数设计如下：

$$r_{\text{end}}=\begin{cases}100,\text{空战胜利}\\ -100,\text{坠毁}\\ -50,\text{失败}\end{cases}\tag{6}$$

其中 end 为 UCAV 胜负判定结果，可以表示为：

$$\text{end}=\begin{cases}\text{win}, |q_1|\geqslant\dfrac{\pi}{2}\,\&\,L^u_{\min}\leqslant R\leqslant L^u_{\max}\,\&\\ \Delta H_{\text{down}}\leqslant\Delta H\leqslant\Delta H_{\text{up}}\\ \text{loss, others}\end{cases}\tag{7}$$

1.3.5　单步综合奖励设计

空战中需要综合考虑角度、距离、高度对空战态势的影响，即在空战中设置每一步的综合奖励。综合奖励的设计是将角度、距离、高度等因素设置权重值，并与胜负奖励函数相加计算单步综合奖励。具体设计如下：

$$r_{\text{total}}=W_1r_A+W_2r_R+W_3r_H+r_{\text{end}}\tag{8}$$

式中：W_1、W_2、W_3 分别表示角度、距离、高度奖励对应的权重，在近距空空导弹后半球攻击策略中由于对角度奖励要求较高，因此设置 $W_1=0.5$，$W_2+W_3=0.5$。

2　LSTM-PPO 算法

2.1　PPO 算法

PPO 算法是由学者 Schulman 提出的一种新型的深度强化学习算法，在策略梯度算法的基础上以演员-评论家（actor-critic，AC）算法为架构演化而来，可以应用在连续的状态和动作空间中[14]。它和其他基于深度强化学习算法相比优势如下：①将新旧策略的更新步长限制在一个合理区间上，让其策略变化不要太剧烈，这样就解决了策略梯度算法无法解决的步长难以选择的问题；②PPO 算法的参数更新方式能够保证其策略一直上升即在训练过程中值函数单调不减；③利用重要性采样原理来离线更新策略，避免浪费更新完的数据。

PPO 算法的目标函数为：

$$J_{\text{ppo}(\theta)}=\hat{E}_t[\min(r_t(\theta)\hat{A}_1,\text{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t)]\tag{9}$$

其中：

$$\hat{A}=r_t+\gamma r_{t+1}+\cdots+\gamma^{T-t+1}\gamma_{T-1}+\gamma^{T-t}V(s_T)-V(s_t)\tag{10}$$

式中：$r_t(\theta)$ 为新旧策略的比值；$\hat{A}_t$ 为每一步的优势函数；r_t 为每一步所获得的奖励；γ 为折扣系数；$V(s_t)$ 为状态值函数；$clip$ 与 ε 分别为截断函数与截断常数，通过将新旧策略的比值限制在 $1-\varepsilon$ 与 $1+\varepsilon$

之间来增强训练效果,避免策略出现突变。

2.2　LSTM 网络

LSTM 网络的每个单元可以被划分为遗忘门 $\boldsymbol{f}_t$、输入门 $\boldsymbol{i}_t$、以及输出门 $\boldsymbol{o}_t$[15] 见图 2 所示。

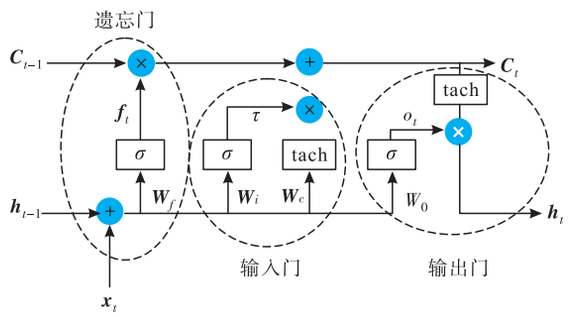

图 2　LSTM 单元结构图

其中,遗忘门主要利用 sigmoid 函数,决定上一时刻网络的输出 $\boldsymbol{h}_{t-1}$ 和上一时刻网络的单元状态 $\boldsymbol{C}_{t-1}$ 是否继续存在于当前时刻网络的单元状态 $\boldsymbol{C}_t$ 中。遗忘门计算公式如下:

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot g[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{11}$$

式中:$\boldsymbol{W}_f$ 为权值矩阵;$\boldsymbol{b}_f$ 为偏置量;$\boldsymbol{x}_t$ 为当前网络的输入;g 表示向量拼接。

输入门利用 sigmoid 函数输出的信息与 tach 函数输出的信息相乘,决定当前时刻的输入 $\boldsymbol{x}_t$ 有多少要传到单元状态 $\boldsymbol{C}_t$ 中。输入门计算公式如下:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot g[\boldsymbol{h}_{t-1},\boldsymbol{x}_1]+\boldsymbol{b}_i)\tanh(\boldsymbol{W}_c\cdot g[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{12}$$

输出门也是利用 sigmoid 函数与 tach 函数输出的信息相乘,决定单元状态 $\boldsymbol{C}_t$ 中有多少可以传到当前输出 $\boldsymbol{h}_t$ 中。输出门的计算公式如下:

$$\boldsymbol{h}_t=\sigma(\boldsymbol{W}_0\cdot g[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_0)\cdot\tanh(\boldsymbol{C}_t) \tag{13}$$

2.3　OU 随机噪声

在训练过程中,平衡算法的探索能力和开发能力至关重要,探索的目的在于寻找到更优的策略。作为引入的随机噪声,OU 噪声在时序上具备较高斯噪声更好的相关性,能够较好地探索具备动量属性的环境,在进一步提升动作决策随机性的同时可以更好地约束探索的区间,减少超出阈值机动的产生。图 3 为基于 OU 随机噪声探索策略示意图。OU 噪声的微分方程形式如下:

$$\mathrm{d}x_t=-\theta(x_t-\mu)\mathrm{d}t+\sigma\mathrm{d}W_t \tag{14}$$

式中:x_t 表示状态;W_t 代表维纳过程;θ、μ、σ 均为参数。

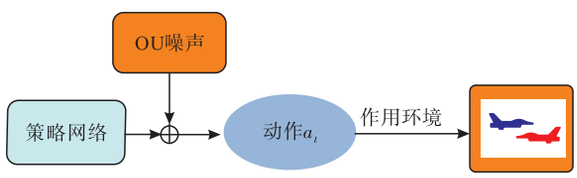

图 3　基于 OU 噪声探索策略

2.4　LSTM-PPO 算法

为了增强 PPO 算法的探索性,本文通过在输出动作上加入 OU 随机噪声来提升 UCAV 对未知状态空间的探索能力。又因为空战环境具有高动态、高维度的博弈性和复杂性,因此单纯采用 PPO 算法中的全连接神经网络来逼近策略函数和价值函数已无法满足其复杂性的需求。本文的策略网络及价值网络使用 LSTM 网络架构,首先引入 LSTM 网络从高维空战态势中提取特征,输出有用的感知信息,增强对序列样本数据的学习能力,再通过全连接神经网络来逼近策略函数及价值函数。LSTM-PPO 算法的架构见图 4。

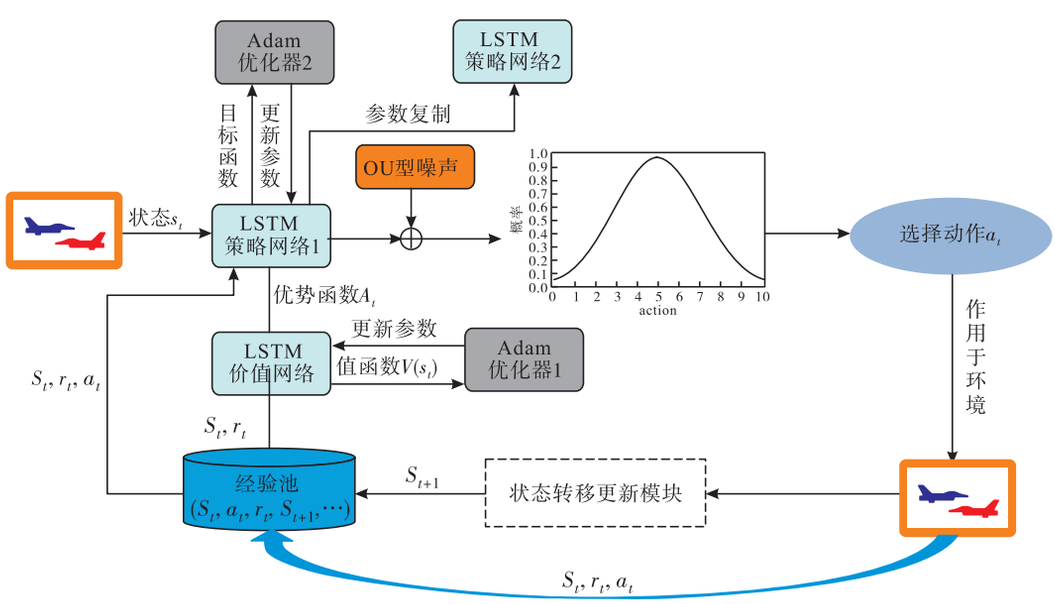

图 4　LSTM-PPO 算法架构图

2.4.1　策略网络设计

针对策略网络部分，输入层设置 12 个节点，对应着 UCAV 和敌机的 12 个状态量 $s=[x,y,z,v,\gamma,\psi,x_e,y_e,z_e,v_e,\gamma_e,\psi_e]$，其中 (x,y,z) 表示 UCAV 的坐标，v 为 UCAV 的速度，γ、ψ 分别代表 UCAV 的航迹倾角及偏航角，(x_e,y_e,z_e) 表示敌机的坐标，v_e 为敌机的速度，γ_e、ψ_e 分别表示敌机的航迹倾角及偏航角；隐藏层分别设置 LSTM 网络层及全连接层，LSTM 网络层设置 3 个网络单元，全连接层设计为 3 层，均采用 tach 为激活函数；输出层有 3 个节点，分别对应着 UCAV 滚转角变化量 $\Delta\mu_t$、攻角变化量 $\Delta\alpha_t$ 及推力变化量 ΔT_t，采用 softmax 为激活函数。策略网络结构图见图 5。

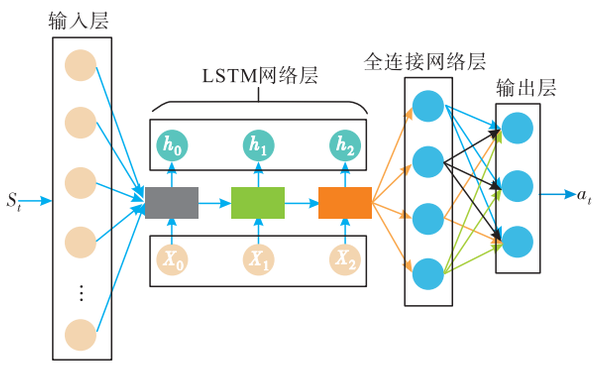

图 5　策略网络结构图

2.4.2　价值网络设计

针对价值网络部分，输入层设置了 15 个节点，对应着 UCAV 和敌机的 12 个状态量 $s=[x,y,z,v,\gamma,\psi,x_e,y_e,z_e,v_e,\gamma_e,\psi_e]$ 及当前策略网络生成的控制量变化量 $a_t=[\Delta\mu,\Delta\alpha,\Delta T_t]$ 的合并；隐藏层中的 LSTM 网络层设置 3 个网络单元，全连接层设计为 3 层，均采用 tach 为激活函数；输出层设置一个节点，对应着状态值函数，采用 Linear 为激活函数。价值网络结构图见图 6。

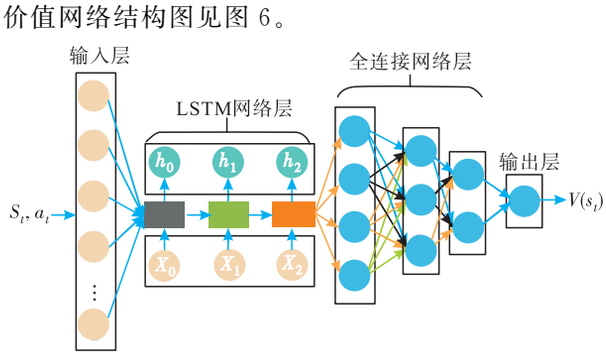

图 6　价值网络结构图

3　仿真实验

3.1　场景设计

本文以 UCAV 与敌机一对一近距空战为背景进行仿真分析，设置 3 组仿真实验，分别为敌机采取随机机动策略，基于专家规则库的机动策略和基于优化算法的机动策略。设每个 epoch 包含 200 个训练回合，每回合的仿真步长设为 30 步，每一步的决策时间为 0.05 s，UCAV 与敌机对抗 900 个 epoch 后停止学习。UCAV 的速度为 300 m/s，航迹倾角和航迹偏角均为 0°，敌机的速度为 250 m/s，航迹倾角为 0°，航迹偏角为 180°。参数设置如表 1 所示，利用表 1 中的参数结合 LSTM-PPO算法对所设计的空战场景进行仿真。

表 1　参数设置

参数	值
策略网络学习率 A_LR	0.000 1
价值网络学习率 C_LR	0.000 2
批量大小 Batch	32
最大回合数 EP_MAX	1 000
每回合最大步数 EP_LEN	200
折扣因子 Gamma	0.9
动作更新步数 A_UPDATE_STEPS	10
价值更新步数 C_UPDATE_STEPS	10

3.2　仿真结果

3.2.1 实验 1：敌机采取随机机动策略

该策略下，针对敌机选择缓慢向上爬升的随机机动动作，UCAV 首先平飞再通过缓慢爬升接近敌机，形成后半球攻击态势并使敌机进入我机导弹攻击区，进而取得空战胜利。图 7 为 UCAV 与敌机空战对抗轨迹图。

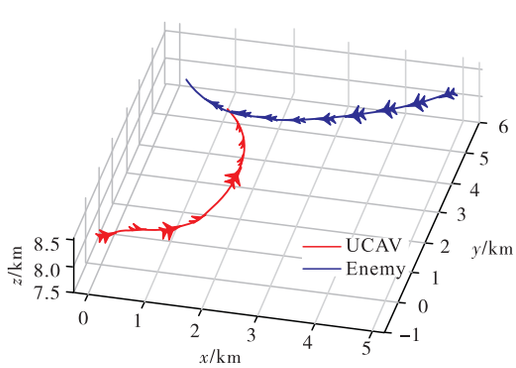

图 7　空战对抗轨迹图(实验 1)

图 8 为反映两机对抗相对优势的累计奖励曲线，横坐标每个 epoch 包含了 200 个训练回合，纵坐标为 200 个训练回合所获得累计奖励的平均值。从图中可以看出，训练初期由于 UCAV 学习不到任何有效策略导致坠毁或被敌机击落，使得累计奖励不断减小，到了训练中期由于我机能够保持平飞，避免了训练前期坠毁的情况，因此累计奖励值逐步增大，最终在约 400 个 epoch 的训练下能够学习到有效的机动动作，形成后半球攻击态势，此时累计奖励值收敛。

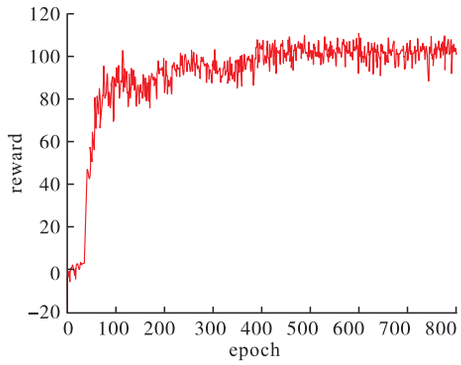

图 8 累计奖励曲线(实验 1)

3.2.2 实验 2:敌机采取基于专家规则库的机动策略

该策略下,针对敌机采取迂回盘旋机动动作[16],我方 UCAV 首先通过缓慢爬升接近敌机,再采取突然俯冲机动跟随敌机,当敌机采取左转缓慢俯冲动作欲完成逃逸时,UCAV 通过小过载爬升机动形成后半球攻击态势,并使敌机进入我机导弹攻击区进而取得空战胜利。图 9 为该场景下的空战对抗轨迹图。

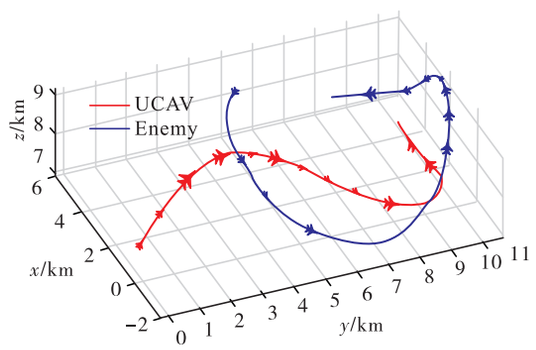

图 9 空战对抗轨迹图(实验 2)

从图 10 的累计奖励曲线中可以看出,初始阶段由于我机对环境认知不足,学习不到较好策略导致出现高惩罚值行为,之后通过训练逐步掌握了能够尾随敌机的策略,最终在约 600 个 epoch 的训练下策略不再大幅变化,此时奖励值收敛。

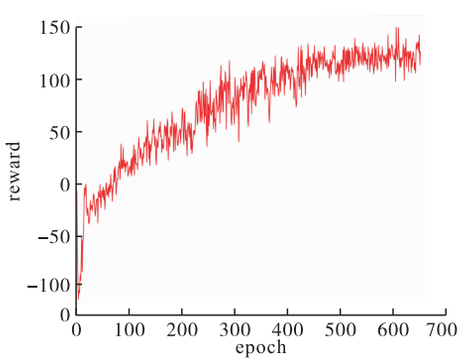

图 10 累计奖励曲线

3.2.3 实验 3:敌机采取基于优化算法的机动策略

由于敌机具有一定的策略[17],因此对抗博弈程度较实验 1 剧烈很多。开始由于 UCAV 高度处于劣势,因此敌机欲采取筋斗机动完成逃逸,此时 UCAV 交替执行平飞与爬升机动接近敌机并与敌机抢占高度优势。当敌机抵达最高点开始向下俯冲,UCAV 完成爬升获得高度优势后,UCAV 跟随敌机进行俯冲,从而在获得后半球角度优势的情况下达到武器发射条件,最终取得空战胜利。图 11 为该场景下的空战对抗轨迹图。

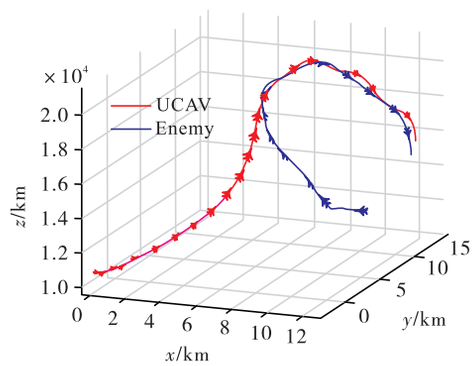

图 11 空战对抗轨迹图(实验 3)

从图 12 的曲线变化趋势可以看出由于敌机飞行具有一定的策略,因此收敛速度比较慢且奖励值曲线波动较为剧烈,体现出了空战任务的复杂性,在大约 720 个 epoch 的训练下累计奖励值收敛,完成学习。

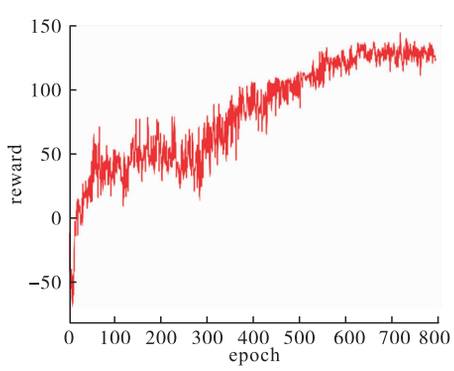

图 12 累计奖励曲线(实验 3)

3.3 算法对比分析

将 PPO 算法和 LSTM-PPO 算法设置相同的超参数,并使用相同的空战环境,经过 900 个 epoch 训练后选取前 800 个 epoch 进行测试。以平均奖励值、收敛时间、空战获胜概率作为衡量两种算法性能的重要指标,进行两种算法在实验 1 和实验 2 下的性能对比分析,见表 2~3。可以看出,LSTM-PPO 算法平均奖励值和获胜概率均大于 PPO 算法,收敛速度 LSTM-PPO 算法快于 PPO 算法。

表 2 实验 1 算法性能的对比

算法	平均奖励值	收敛时间	获胜概率
PPO	19.82	约 450 个 epoch	0.816
LSTM-PPO	24.87	约 400 个 epoch	0.887

表 3　实验 2 算法性能的对比

算法	平均奖励值	收敛时间	获胜概率
PPO	20.68	约 700 个 epoch	0.803
LSTM-PPO	26.82	约 600 个 epoch	0.876

4　结语

由于空战环境复杂、格斗态势高速变化，因此本文针对 UCAV 与敌机一对一近距空战引入了基于 LSTM-PPO 算法的 UCAV 机动决策方法，设计了敌机采取随机机动策略、基于专家规则库的机动策略和基于优化算法的机动策略 3 组仿真实验。3 组实验下的结果表明，不论敌机采取何种机动策略，UCAV 均可以很好地感知空战态势，做出合理的机动动作，进而取得空战胜利。与 PPO 算法作性能对比也可以发现，基于 LSTM-PPO算法的 UCAV 空战机动决策方法具有获得平均奖励值大、收敛速度快、获胜概率高的优点。

参考文献

[1] 周新民，吴佳晖. 无人机空战决策技术研究进展[J]. 国防科技，2021，42(3)：148-154.

[2] 周思雨，王庆超，王子健，等. 基于 copeland 集结算法的协同空战机动决策方法[J]航空计算技术，2020，50(6)：43-46.

[3] 李伟. 基于微分对策理论的无人战机空战决策方法研究[J]. 北京航空航天大学学报，2019，45(4)：722-734.

[4] MCGREW J S，HOW J P，WILLIAMS B，et al. Decision Theoretical Approach to Pilot Simulation[J]. Journal of Aircraft，1999，27(4)：632-641.

[5] 傅莉，王晓光. 无人战机近距空战微分对策建模研究[J]. 兵工学报，2012，33(10)：1210-1216.

[6] 邓可，彭宣淇，周德云. 基于矩阵对策与遗传算法的无人机空战决策[J]. 火力与指挥控制，2019，44(12)：188-193.

[7] 梅丹，吴文海，徐家义. 影响图的空战机动决策方法[J]. 火力与指挥控制，2008，33(5)：1641-1654.

[8] 郭万春，解武杰，尹晖，等. 基于改进双延迟深度确定性策略梯度法的无人机反追击机动决策[J]. 空军工程大学(自然科学版)，2021，22(4)：15-21.

[9] VAN HASSELT H，GUEZ A，SILVER D. Deep Reinforcement Learning with Double Q-Learning[C]// Proceedings Intelligence，Phoenix，US：[s. n.]，2016：2094-2100.

[10] SCHULMAN J，LEVINE S，MORITZ P，et al. Trust Region Policy Optimization[C]// Proceedings of the 31st International Conperence on Machine Learning. Lille，France：JMLR，2015：1889-1897.

[11] KUANG J W，YANG H Z，et al. Dynamic Prediction of Car Disease Using Improved LSTM [J]. International Journal of Crowd Science，2019，3(1)：14-25.

[12] 赖俊，饶瑞. 深度强化学习在室内无人机目标搜索中的应用[J]. 计算机工程与应用，2020，56(17)：156-160.

[13] 王杰，丁达理，许明，等. 基于目标逃逸机动预估的空空导弹可发射区[J]. 北京航空航天大学学报，2019，45(4)：722-734.

[14] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Playing Atari with Deep Reinforcement Learning[J]. ArXiv Preprint ArXiv：1312.5602，2013.

[15] WANG P X，WANG H E，ZHANG H C，et al. A Hybrid Markov and LSTM Model for in Door Location Rrediction[J]. IEEE Access，2019，7：1852-1859.

[16] 傅莉，谢福怀，孟光磊，等. 基于滚动时域的无人机空战决策专家系统[J]. 北京航空航天大学学报，2015，41(11)：1994-1999.

[17] 高阳阳，余敏建，韩其松，等. 基于改进共生生物搜索算法的空战机动决策[J]. 北京航空航天大学学报，2019，45(3)：429-436.

(编辑：徐敏)