ORIGINAL ARTICLE

Open Access

Autonomous maneuver strategy of swarm air combat based on DDPG

Luhe Wang^{1*} , Jinwen Hu¹, Zhao Xu² and Chunhui Zhao¹

Abstract

Unmanned aerial vehicles (UAVs) have been found significantly important in the air combats, where intelligent and swarms of UAVs will be able to tackle with the tasks of high complexity and dynamics. The key to empower the UAVs with such capability is the autonomous maneuver decision making. In this paper, an autonomous maneuver strategy of UAV swarms in beyond visual range air combat based on reinforcement learning is proposed. First, based on the process of air combat and the constraints of the swarm, the motion model of UAV and the multi-to-one air combat model are established. Second, a two-stage maneuver strategy based on air combat principles is designed which include inter-vehicle collaboration and target-vehicle confrontation. Then, a swarm air combat algorithm based on deep deterministic policy gradient strategy (DDPG) is proposed for online strategy training. Finally, the effectiveness of the proposed algorithm is validated by multi-scene simulations. The results show that the algorithm is suitable for UAV swarms of different scales.

Keywords: Deep reinforcement learning, Cooperative air combat, Swarm, Maneuver strategy

1 Introduction

Unmanned aerial vehicle (UAV) with the characteristics of low cost, strong mobility, high concealment and no need of pilot control, have been more and more widely used to replace manned aircraft to perform military tasks such as detection, monitoring and target strike, and is a typical representative of “non-contact” combat equipment [1]. Due to the limitations of single UAV’s mission and combat capability, the swarms and intelligence of unmanned combat have become a research hotspot in recent years. With the increase of the operating range of airborne detection equipment, the scope of modern air combat has gradually developed from line of sight to beyond line of sight [2]. Swarms beyond visual range air combat refers to the situation assessment [3–5], environment awareness [6, 7], and maneuver strategy [8] of UAVs through sensing or detection equipment, and maneuver strategy is the basis of the above tasks.

The existing air combat maneuver strategies can be divided into two categories: rule-based strategy and learning-based strategy. Rule-based strategy mainly select actions according to the given behavior rules in air combat, and not need online training and optimization, including matrix game algorithm [9, 10], expert system [11], influence graph method [12, 13], differential game method [14], etc. The matrix game method is prone to the phenomenon that the overall strategy effect is not good. On the one hand, the score function used in this method is difficult to design, which can not accurately describe the actual air combat. On the other hand, this method has reward delay in the sequential decision making, which does not have the ability of long-term planning. Expert experience is difficult to cover all air combat situations, so is very complex to establish rule base and constraint conditions by using expert system method. Moreover, the UAVs cannot make decision independently by using expert system method. When UAV can not find the appropriate strategy scheme in the rule base, it must introduce human intervention. In [15] and [16], multi-level influence graph can realize multiple-to-one air combat, but it

*Correspondence: luhe_wang@foxmail.com

¹School of Automation, Northwestern Polytechnical University, Xi’an 710129, China

Full list of author information is available at the end of the article

is only suitable for small-scale UAV swarms. Moreover, influence diagram method relies on prior knowledge, and the algorithm reasoning process is cumbersome, which can not meet the requirements of real-time and high dynamic air combat. The two methods introduced in [14] can't be applied to the air combat scene without model or with incomplete environment information, because these methods need to accurately model and describe the strategy model. In short, rule-based strategies are monotonous and rigid, which can not adapt to the complex and highly dynamic air combat scenarios, and can not meet the requirements of intelligent operations.

Learning-based strategies optimize the model and structural parameters by means of online learning or training data, which mainly includes artificial immune system, genetic algorithm, heuristic learning algorithm, neural network method, deep reinforcement learning (DRL), etc. Artificial immune [17] uses training data to make UAV's maneuver system can deal with different air combat situations, but the convergence is slow. The fuzzy tree proposed in [18] is too complex to be used in beyond visual range air combat. According to the characteristics of beyond visual range air combat, the University of Cincinnati has built the "alpha" intelligent air combat system [19], which has used more than 150 dimensional input data in a two-to-four air combat scene. The data dimension will be further increased if the control of airborne sensors is considered in the future. Therefore, the processing and application of high-dimensional massive data are the main difficulty in multi-UAVs autonomous air combat.

Deep learning (DL) has a good application effect in data processing [20–23], and reinforcement learning (RL) is widely used in autonomous maneuver [24–26]. With the successful application of deep reinforcement learning (DRL) in complex sequential decision making problems such as Alpha-Go [27] and Alpha-Go Zero [28], it is possible to solve air combat maneuver strategy problems by using reinforcement learning. Reinforcement learning is a learning method that uses "trial and error" to interact with the environment [29], which is a feasible method for autonomous decision making of UAV intelligent air combat maneuver strategy. The application of reinforcement learning in air combat is mainly based on value function search [30, 31] and strategy search [32, 33]. Deep Q network (DQN) algorithm is improved in [32] and realizes the UAV close range one-to-one air combat, but the algorithm uses discrete state and motion spaces, which makes the results of air combat quite different from reality. The Actor-Critic (A-C) framework is used in [34] to realize the continuous expression of UAV maneuver strategy in state space, but the algorithm is only effective in two-dimensional space. In [35], the deep deterministic policy gradient (DDPG) is applied to air combat decision

making. However, the design of reward and punishment function is relatively simple, and it is difficult for agents to learn complex tactics. In general, the above studies improve the effect and ability of air combat maneuver strategy algorithm to a certain extent, but ignores the generalization ability of the algorithm. That is, the existing research on UAV air combat based on reinforcement learning algorithm mainly focuses on specific scenarios. Whether it is one-to-one air combat or multi-UAVs cooperation, the number of UAVs on both sides must be fixed, has poor compatibility.

To sum up, the strategy of swarm autonomous maneuver in dynamic environment is an urgent problem to be solved. That is, the maneuver module should have good compatibility, which is suitable for swarms of different sizes. Because the number of UAVs is constantly changing in the process of combat, the amount of information of both sides is dynamically changing.

In this paper, an autonomous maneuver strategy of swarm combat in beyond visual range air combat based on reinforcement learning is proposed. In this strategy, the autonomous maneuver decision and cooperative operation of UAV in swarm are realized, and the scalability of the algorithm is improved. The main contributions are as follows: First, the overall framework of autonomous maneuver strategy methods of UAV swarm is designed. Second, the swarm beyond visual range air combat model is established, the target entry angle and target azimuth are defined, and the situation assessment function is designed. Third, a two-stage maneuver decision strategy is proposed to solve dimension explosion. Two maneuver actions are designed in the first stage to realize the rapid collision avoidance and good communication in swarm. Multi-to-one air combat is transformed into one-to-one air combat in the second stage, avoiding directly dealing with high-dimensional swarm information. Fourth, based on the basic principle of reinforcement learning and the requirements of swarms control, the Actor-Critic network framework is designed, the state space and action strategy are given, and the reward function is designed based on the distance to realize the rapid convergence of the algorithm. Fifth, based on memory bank and target network, an algorithm is designed and A-C network is trained to obtain an autonomous cooperative maneuvering strategy method for UAV swarm. The simulation experiments are carried out to verify that the algorithm can be applied to UAV swarms of different scales.

The following parts are arranged as follows: Section 2 gives the algorithm design framework, establishes the model of air combat. Section 3 designs the swarm combat maneuver decision, gives the algorithm flow of swarm air combat based on DDPG. Section 4 conducts simulation analysis. Section 5 summarizes the full paper.

Fig. 1 Process of swarms fight

2 Swarm air combat model

2.1 Overall framework

The process of UAV swarm combat includes three modules [36]: environment awareness, situation assessment and maneuver decision as shown in Fig. 1. Each UAV obtains the current battlefield information through environment awareness module via data links or airborne sensors, which mainly includes the states of the enemy target and other UAVs. Through the situation assessment module, each UAV assesses its current situation such as “advantageous” or “disadvantageous” situations, based on which the decision on how to maneuver at next time step is made. The three modules form a closed loop, which ultimately achieve the effect of inter-aircraft cooperation and air combat tasks.

2.2 Air combat model

In this paper, a multi-to-one air combat is studied as shown in Fig. 2, where multiple UAVs are deployed to monitor and attack one enemy target. In the ground coordinate system $Oxyz$ as shown in Fig. 3, Ox axis takes the east, Oy axis the north, and Oz axis the vertical. Denote the i -th UAV by UAV_i and its position at time t

by $P_{i,t}$, where t is the discrete time index under fixed sampling period T . The way-point motion model of the UAVs is given by

$$P_{i,t+1} = P_{i,t} + v_{U_{i,t}}, \quad (1)$$

where $v_{U_{i,t}}$ is the vector from $P_{i,t}$ to $P_{i,t+1}$, which is seen as the control input to be designed in the follow parts. The position of target is denoted by $P_{T,t}$.

Remark 1 In general, the actual position $P_{i,t+1}$ at time $t + 1$ may not be exactly equal to $P_{i,t} + v_{U_{i,t}}$ due to the flight interferences or the model uncertainties, even if $v_{U_{i,t}}$ is within the physical control limitations. It is a classical control problem to minimize the error $\|P_{i,t+1}^* - P_{i,t+1}\|$ given an approximated UAV flight model which can be solved by conventional PID or robust control methods, where $P_{i,t+1}^* = P_{i,t} + v_{U_{i,t}}$ is the desired next-time position. This problem is not the focus of our research, and in this paper it is assumed that $P_{i,t} = P_{i,t}^*$ holds all the time. The control limitations will be considered in practice to make $v_{U_{i,t}}$ applicable for a specific type of UAV.

Fig. 2 Schematic Diagram of Multi-to-One Air Combat

Fig. 3 Motion model of UAV

In the same coordinate system, a multi-to-one air combat model is established as shown in Fig. 4. Denote by $\varphi_{U_i,t}$ the angle between the vectors $P_{T,t} - P_{i,t}$ and $v_{U_i,t}$, which is named as the target azimuth, and similarly, $\varphi_{T,t}$ for the angle between the vectors $P_{i,t} - P_{T,t}$ and $v_{T,t}$, which is named as the target entry angle. They can be computed in real time as follows, where “ $\top$ ” denotes the transpose operation,

$$\varphi_{U,i,t} = \arccos \left(\frac{v_{U,i,t}^\top (P_{T,t} - P_{i,t})}{\|v_{U,i,t}\| \|P_{T,t} - P_{i,t}\|} \right), 0 \leq \varphi_{U,i,t} \leq \pi, \quad (2)$$

$$\varphi_{T,t} = \arccos \left(\frac{\mathbf{v}_{T,t}^\top (P_{i,t} - P_{T,t})}{\|\mathbf{v}_{T,t}\| \|P_{i,t} - P_{T,t}\|} \right), 0 \leq \varphi_{T,t} \leq \pi. \quad (3)$$

In the situation assessment module, we use the situation assessment function to evaluate the real-time situation of UAV. Based on the attack model [34] and evaluation function [37], the effective attack range of a UAV in the air combat is a cone with an axis in the direction of $v_{U_i,t}$ and angle of φ_m , which is truncated by a ball of radius R_W as shown in Fig. 4, where R_W represents the attack range of weapons. Similarly, we can define the cone-shape attack range for the target. Heuristically, if UAV_{*i*} is in the attack range of the target, UAV_{*i*} is said to be in the “disadvantageous” situation; if the target is in the attack range of UAV_{*i*} while UAV_{*i*} is not in the attack range of the target, UAV_{*i*} is said to be in the “advantageous” situation; otherwise, UAV_{*i*} is said to be in the “balance” situation. The three different situation assessment results can be denoted by “-1”,

“1” and “0” respectively, which are defined as follows,

$$K_{i,t} = \begin{cases} -1, & \text{if } R_{i\Gamma,t} \leq R_W & \& \quad \varphi_{\Gamma,t} \leq \frac{1}{2}\varphi_m; \\ 1, & \text{if } R_{i\Gamma,t} \leq R_W & \& \quad \varphi_{\cup_{i,t}} \leq \frac{1}{2}\varphi_m & \& \quad \varphi_{\Gamma} > \frac{1}{2}\varphi_m; \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where $R_{iT,t} = \|P_{i,t} - P_{T,t}\|$.

3 Maneuver strategy optimization based on deep reinforcement learning

3.1 Maneuver strategy for air combat

A two-stage maneuver decision strategy is proposed in this paper. In the first stage, each UAV only considers the interactions between vehicles in the swarm by checking the stable communication and safety constraints. If all the constraints are satisfied, then the UAV moves to the second stage of maneuver decision, in which the UAV only considers the situation between the target and itself. In the following part, we will discuss the two stages separately.

In the first stage, we consider two basic principles [37]: 1) UAVs should keep a safe maneuver distance to avoid inter-vehicle collisions; 2) UAVs should stay in a closed ball to ensure stable communication. Thus, we define two maneuver actions: separating and gathering.

Separating If the distance between two UAVs is less than a predefined range $d_{\min}$, there might be a potential threat of collision. That is, if $\|P_{j,t} - P_{i,t}\| < d_{\min}$ we set

$$v_{i,t} = \lambda_1 (P_{i,t} - P_{j,t}) + \lambda_2 v_{i,t-1}, \quad (5)$$

where λ_1 and λ_2 are positive weight numbers, which are selected such that $(\lambda_1 \|P_{i,t} - P_{j,t}\|) / (\lambda_2 \|v_{i,t-1}\|)$ is small and $\|v_{i,t}\|$ satisfies the speed limitations.

Fig. 4 Multi-to-one air combat model

Fig. 5 A-C Network Architecture

Gathering The swarm should be constrained in a ball of a predefined radius R_m . For each UAV_{*i*}, if the condition of separating is not satisfied, then it computes the center of the ball by the average of all UAV positions excluding itself, i.e.,

$$P_{ci,t} = \frac{1}{n-1} \sum_{j=1 \& j \neq i}^n P_{j,t}, \quad (6)$$

where n is the total number of UAVs. If the distance between UAV_{*i*} and P_{ci} is less than R_m , it needs to approach the center by making a small maneuver. Thus, we set

$$v_{i,t} = \lambda_3 (P_{ci,t} - P_{i,t}) + \lambda_4 v_{i,t-1}, \quad (7)$$

where λ_3 and λ_4 are positive weight numbers, and their selection rules are similar to those for λ_1 and λ_2 .

If the UAV does not need to execute the first stage maneuver actions, it will enter the second stage. In the second stage, we consider two principles for combat against target : 1) UAV should escape from the attack range of target if the situation is disadvantageous; 2) UAV should approach the target if the situation is advantageous. Swarm air combat is converted into one-to-one air combat in this stage, we design DDPG algorithm to realize it and explain it in detail in the following section.

3.2 One-to-one Maneuver decision algorithm design

Reinforcement learning algorithm obtains the optimal action strategy π_* by finding the optimal action value function [38].

$$\begin{aligned} q_\pi(s, a) &= E_\pi [G_t | s_t = s, a_t = a] \\ &= E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a \right], \end{aligned} \quad (8)$$

$$\pi_*(a|s) = \underset{\pi}{\operatorname{argmax}} Q_\pi(s, a), \quad (9)$$

where γ is the discount factor, it can control the proportion of future rewards in cumulative rewards. (9) is to look up the table to find the optimal strategy in different states, and it is only suitable for discrete space. In order to solve the autonomous decision making problem in continuous air combat environment, we add neural network and use DDPG algorithm to realize the continuous motion control of UAV.

DDPG uses Actor-Critic network to fit the action strategy π and the action value Q_π , and the parameters are θ^μ and θ^Q respectively. Actor network is used to generate maneuver strategy $\pi(s, a, \theta^\mu)$, Critic network outputs action value $Q_\pi(s, a, \theta^Q)$ to evaluate $\pi(s, a, \theta^\mu)$, and $\pi(s, a, \theta^\mu)$ is optimized by optimizing $Q_\pi(s, a, \theta^Q)$.

Network Structure The input s_t of the Actor network is a vector from $P_{i,t}$ to $P_{T,t}$, and the output O_a is target speed of UAV $v_{U,i,t}^*$.

$$s_t = P_{T,t} - P_{i,t}, \quad (10)$$

$$O_c = [Q_{O_a^x}, Q_{O_a^y}, Q_{O_a^z}]. \quad (11)$$

O_a and s_t are used as the inputs of Critic network, and O_c is the evaluation of the results of O_a . The DDPG network structure is shown in Fig. 5, and the parameters will be given in the simulation part.

One-to-one Maneuver Design Maneuver a_t realizes swarm air combat by controlling $v_{U,i,t}$, which is generated by A network:

$$a_t = O_a, 0 \leq \|O_a\| \leq \sqrt{3}. \quad (12)$$

In order to realize the agent's exploration of the environment and the optimization of a_t , the noise N generated by OU process [39] is usually added to O_a to synthesize a'_t to participate in network training. The function of N and a'_t are shown in (13) and (14).

$$a'_t = a_t + \xi N, \quad (13)$$

$$N = \theta(\mu - O_a(s_t)) + \sigma W, \quad (14)$$

where θ and σ are weights, μ represents the mean value, W is Gaussian noise. ξ is an exploration coefficient that decreases with the increase of training rounds. OU process can make exploration more efficient.

Reward Function The reward function designed in this paper is as follows:

$$r_t = \|s_t\| - \|s_{t+1}\|. \quad (15)$$

When the UAV approaches the $P_{T,t}$, that is, $\|s_{t+1}\| < \|s_t\|$, a positive reward will be given, otherwise a penalty proportional to the distance will be given. Using (15) as reward function, regardless of the initial position of UAVs and enemy, UAVs will eventually fly to the rear of the target to attack it.

The algorithm flow is shown in Algorithm 1. According to $P_{i,t}$ and $P_{T,t}$, the current state s_t is determined. The online Actor network generates O_a according to the network parameter θ^μ . The OU process generates N and adds it to O_a to synthesize the action a_t . The UAV moves to the next state and gets a reward r_t . Set (s_t, a_t, r_t, s_{t+1}) to the memory bank. UAV repeats the above steps many times to collect a large number of samples. When the number of memory samples meets the requirements, random samples are selected to train A-C network. A batch of samples $\{(s_i, a_i, r_i, s_{i+1})_m | m = 1, 2, \dots, M\}$ are randomly selected from the memory bank to calculate the target value y_i .

Algorithm 1 DDPG model training process combined with optimization algorithm

```

1: Initialize online A-C network with random parameters  $\theta^\mu, \theta^Q$ . Initialize target A-C network with parameters  $\theta^{\mu'} \leftarrow \theta^\mu, \theta^{Q'} \leftarrow \theta^Q$ . Initialize memory buffer  $M$ . Set the target maneuver strategy
2: for episode = 1,  $E$  do
3:   Initialize the initial state of air combat
4:   Receive initial observation state  $s_1$ 
5:   for  $t=1,2,\dots,T$  do
6:     Online A network generates action  $O_a$ , using (14) generates noise  $N$ , get (13)
7:     Execute action  $a'_t$  and observe reward and observe new state  $s_{t+1}$ 
8:     Store transition  $(s_t, a'_t, r_t, s_{t+1})$  in memory bank
9:     Sample a random minibatch  $\{(s_i, a'_i, r_i, s_{i+1})_m | m = 1, 2, \dots, M\}$  from memory bank
10:    Get  $y_i = r_i + \gamma Q'(s_{i+1}, O'_a(s_{i+1} : \theta^{\mu'}); \theta^{Q'})$ 
11:    Update parameters of online C network by minimizing (17)
12:    Update the online Actor policy using (19)
13:    Using (20) update the target Networks
14:  end for
15: end for

```

$$y_i = r_i + \gamma Q'(s_{i+1}, O'_a(s_{i+1} : \theta^{\mu'}); \theta^{Q'}), \quad (16)$$

where γ is the decay rate, the meaning is the same as that in (12), $O'_a(s_{i+1} : \theta^{\mu'})$ is the output of target A network, representing the target action, $Q'(s_{i+1}, O'_a(s_{i+1} : \theta^{\mu'}); \theta^{Q'})$ is the output of target C network, representing the expected value of taking action $O'_a(s_{i+1})$ under the state s_{i+1} , and the parameters of target A and C networks are $\theta^{\mu'}, \theta^{Q'}$. Update target network parameters:

$$L = \frac{1}{n} \sum_i (y_i - Q(s_i, a_i; \theta^Q))^2, \quad (17)$$

$$J(\theta^\mu) = -\frac{1}{n} \sum_{i=1}^n Q(s_i, a_i, \theta^Q), \quad (18)$$

$$\nabla J(\theta^\mu) \approx \frac{1}{n} \sum_i \nabla_a Q(s, a, \theta^Q)|_{s=s_i, a=O_a(s_i)} \nabla_{\theta^\mu} O_a(s; \theta^\mu)|_{s_i}, \quad (19)$$

$$\begin{cases} \theta^{Q'} \leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'}, \\ \theta^{\mu'} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'}. \end{cases} \quad (20)$$

Gradient descent method is used to minimize (17) and optimize the parameters of Critic network. (18) and (19) are loss function and parameter optimization equation of Actor network. The parameters of the neural network are

updated by (20). DDPG uses soft update method to update network parameters. Each time, the parameters are only updated a little, which makes the learning process more stable. With the increase of the number of training rounds, the agent's Maneuver selection in different states tends to be optimal.

3.3 Multi-to-one Maneuver strategy design

When (17) is close to 0 or there is no obvious change, the training should be stopped. After the trained neural network is saved, the one-to-one air combat maneuver decision can be obtained. In the following part, we use the DDPG algorithm and two maneuver actions to realize swarm multi-to-one air combat, and the process is as Algorithm 2.

In Algorithm 2, $R_{ij,t}, P_{T,t}, K_{i,t}$ and $P_{ci,t}$ are put to determine whether the UAV needs to execute "separating" or "gathering". If necessary, get the $v_{U_i,t}$ through (5) and (7); If it is not necessary, enter Algorithm 1 and get $v_{U_i,t}$.

Algorithm 2 Process of swarms combat

```

1: while steps < maxsteps || no attacking do
2:   for  $i=1,2,\dots,n$  do
3:     obtain current situation and battlefield information:  $R_{ij,t}, P_{T,t}, K_{i,t}, P_{ci,t}$ 
4:     if separating then
5:       Using (5) get  $v_{U_i,t+1}^*$ 
6:     else
7:       if gathering then
8:         Using (7) get  $v_{U_i,t+1}^*$ 
9:       else
10:        Calculate  $s_{i,t} = P_{T,t+1} - P_{i,t}$ , feeds the state  $s_{i,t}$  into the Network trained by Algorithm 1 and outputs  $v_{U_i,t+1}^* = O_a$ .
11:      end if
12:    end if
13:    Using (22) get  $v_{U_i,t+1}$ 
14:    Using (1) get  $P_{i,t}$ 
15:  end for
16:  Update the position of all UAVs and target
17: end while

```

Limited by UAV maneuver ability, $v_{U_i,t}$ given by two maneuver actions or neural network makes UAV unable to reach. Assuming that the angle between $v_{U_i,t}$ and $v_{U_i,t-1}$ is ϕ , the maximum turning angle of UAV is α as shown in Fig. 3. The maximum speed of UAV is V_{max} , and the minimum speed is V_{min} . The actual $v_{U_i,t}$ is as shown in (22). Then we use (1) to get $P_{i,t+1}$.

$$\|v_{U_i,t}\| = \begin{cases} \|O_a\| & \text{if } V_{min} < \|O_a\| < V_{max}, \\ V_{max} & \text{if } \|O_a\| \geq V_{max}, \\ V_{min} & \text{if } \|O_a\| \leq V_{min}, \end{cases} \quad (21)$$

Table 1 Simulation parameters

Variables	Value	Variables	Value	Variables	Value
Learning rate of a-network	0.001	τ	0.001	E	1000
Learning rate of c-network	0.001	γ	0.99	C	100000
K_s	5	φ_m	60°	R_m	200
θ	0.2	μ	0	σ	1

$$v_{U_i,t} = \begin{cases} \frac{O_a}{\|O_a\|} \|v_{U_i,t}\| & \text{if } \phi \leq \alpha, \\ \frac{v'_{U_i,t+1}}{\|v'_{U_i,t+1}\|} \|v_{U_i,t}\| & \text{if } \phi > \alpha, \end{cases} \quad (22)$$

$$v'_{U_i,t} = \frac{O_a - v_{U_i,t-1}}{\|O_a - v_{U_i,t-1}\|} \frac{\sin \alpha}{\sin(\pi - \alpha - \beta)} \|v_{U_i,t-1}\| + v_{U_i,t-1}. \quad (23)$$

When $\phi \leq \alpha$, $v_{U_i,t}$ is given by (21) and (22) to meet $V_{\min} \leq v_{U_i,t} \leq V_{\max}$. When $\phi > \alpha$, $v_{U_i,t+1}$ is calculated by (22) and (23), β is the complement of the angle between $v_{U_i,t}$ and $v_{U_i,t} - v_{U_i,t-1}$.

4 Simulation

4.1 Simulation setup

The default parameters of this simulation experiment are shown in Table 1. The velocity of our UAV and target refers to the displacement of UAV in unit step time in the simulation system. The enemy plane moves in a straight line or curve at a constant speed, while our UAV moves in a variable speed. The speed of UAV is determined by the output of the neural network. The position of our UAVs is initialized randomly. The structure of A-C network is shown in Fig. 5. Actor network consists of two layers, and the number of nodes is 400 and 800. In Critic network, the

number of common layer nodes that input information is 800, and the others are 400. After unit conversion, the real combat environment of $50\text{km} \times 50\text{km} \times 50\text{km}$ is simulated. The maximum turning angle of UAV is 60° . The maximum speed is 180km/h . The maximum communication radius is 20km , and the maximum attack distance is 500m .

To test the effectiveness of the proposed algorithm, we define the following three metrics: Compatibility, $\varphi_{U_i,t}$, $R_{i,T,t}$. Compatibility refers to whether the algorithm can be extended from one-to-one air combat to multi-to-one combat, $\varphi_{U_i,t}$ measures the change of $K_{i,T,t}$ and the rationality of maneuvering of UAV in air combat. $R_{i,T,t}$ reflects whether the algorithm converges.

4.2 Simulation results

The UAV fixed-point arrival capability is simulated and verified, as shown in Fig. 6. Given the target point, whether it is a single UAV or a multi aircraft environment, our aircraft can independently plan the path to reach the target point through DDPG algorithm.

Aiming at the dynamic target, the combat effect of a single UAV is verified first, and the enemy plane moves in a uniform straight line or circular motion. As shown in Fig. 7, our UAV can monitor the position of the target in real time and accurately track it. The change of the distance and angle between UAV and target is shown in Fig. 8.

In the early stage of simulation, the angle between the enemy and UAV is obtuse, so UAV is in an inferiority situation. On the one hand, the UAV is far away from the enemy, on the other hand, it quickly adjusts the azimuth, so UAV situation has changed from an inferiority to an advantage. Then the UAV constantly approaches the target and adjusts the azimuth angle in real time to keep it within the maximum attack angle. Finally, the UAV can

Fig. 6 Autonomous Path Planning of UAV with Fixed Target Points

Fig. 7 UAV Dynamic Target Tracking(blue represents target)

lock the target for more than 2s to complete the tracking and attacking of the target.

Next, verify the effect of multi Aircrafts cooperative combat, the target still do uniform linear motion or circular motion. The results of swarms cooperative combat are shown in Figs. 9 and 10.

Compared with Fig. 8, due to factors such as aggregation and separation need to be considered in the swarms, the UAV can not all approach the target point at the same time, so the distance between the UAV and the target is close or far. In the process of tracking, UAV constantly adjusts the azimuth angle, and the azimuth angles

Fig. 8 Result of Azimuth and Distance for one-to-one air combat

Fig. 9 Swarms Dynamic Target Tracking (black represents target)

Fig. 10 Result of Azimuth and Distance for multi-to-one air combat

of several aircraft close to the target are kept in the maximum attack range, showing a dominant situation. It can also be seen from Fig. 10 that all UAVs are kept within the communication range from the swarm center.

In the above simulation process, each simulation step corresponds to a decision making, which is converted to 1s of the real environment. Therefore, the simulation steps of the abscissa in Figs. 8 and 10 can be regarded as time.

5 Conclusion

A maneuver strategy based on DDPG algorithm is proposed to realize UAV swarm combat. Based on swarm framework, the air combat model and behavior set are designed for UAV to realize autonomous decision making. According to the characteristics of DDPG algorithm and task requirements, the distance between UAV and target point and the velocity value of UAV are taken as the input and output of actor network, and the reward function is constructed by relative distance to train the neural network parameters, so that the network converges quickly. A visual simulation environment is built to verify the application effect of the algorithm. The results show that the swarm maneuver strategy based on deep reinforcement learning algorithm can complete the attacking of the target on the premise of clear tasks. No matter the target moves in a straight line or curve, it has good simulation effect. Our UAVs can lock it in the attack range and keep it for a certain period of time. It has a certain practical value.

6 Nomenclature

DDPG: Deep Deterministic Policy Gradient; DL: Deep Learning; DRL: Deep Reinforcement Learning; RL: Reinforcement Learning; UAV: Unmanned Aerial Vehicle

Authors' contributions

The authors read and approved the final manuscript.

Funding

This work is supported by National Natural Science Foundation of China under Grant 61803309, the Key Research and Development Project of Shaanxi Province under Grant 2020ZDLGY06-02, the Aeronautical Science Foundation of China under Grant 2019ZA053008, the Open Foundation of CETC Key Laboratory of Data Link Technology under Grant CLDL-20202101, the China Postdoctoral Science Foundation under Grant 2018M633574.

Availability of data and materials

Not applicable.

Code availability

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Automation, Northwestern Polytechnical University, Xi'an 710129, China. ²School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China.

Received: 1 July 2021 Accepted: 8 October 2021

Published online: 04 December 2021

References

1. Y. Li, X. Qiu, X. Liu, Q. Xia, Deep reinforcement learning and its application in autonomous fitting optimization for attack areas of ucavs. *J. Syst. Eng. Electron.* **31**(4), 734–742 (2020)
2. D. Hu, R. Yang, J. Zuo, Z. Zhang, Y. Wang, Application of deep reinforcement learning in maneuver planning of beyond-visual-range air combat. *IEEE Access*. **PP**(99), 1–1 (2021)
3. A. Xu, X. Chen, Z. W. Li, X. D. Hu, A method of situation assessment for beyond-visual-range air combat based on tactical attack area. *Fire Control Command Control*. **45**(9), 97–102 (2020)
4. Z. H. Hu, Y. Lv, A. Xu, A threat assessment method for beyond-visual-range air combat based on situation prediction. *Electron. Opt. Control*. **27**(3), 8–1226 (2020)
5. W. H. Wu, S. Y. Zhou, L. Gao, J. T. Liu, Improvements of situation assessment for beyond-visual-range air combat based on missile launching envelope analysis. *Syst. Eng. Electron.* **33**(12), 2679–2685 (2011)
6. H. Luo, Target detection method in short coherent integration time for sky wave over-the-horizon radar. *Sadhana*. **45**(1) (2020)
7. T. Liu, R. W. Mei, in *Proceedings of 2019 International Conference on Computer Science, Communications and Multimedia Engineering (CSCME 2019)*, Shanghai, China. Over-the-horizon radar impulsive interference detection with pseudo-music algorithm, (2019). *Computer Science and Engineering* (ISSN 2475-8841)
8. H. Wu, H. Li, R. Xiao, J. Liu, Modeling and simulation of dynamic ant colony's labor division for task allocation of uav swarm. *Phys. A Stat. Mech. Appl.*, 0378437117308166 (2017). <https://doi.org/10.1016/j.physa.2017.08.094>
9. F. Austin, G. Carbone, H. Hinz, M. Lewis, M. Falco, Game theory for automated maneuvering during air-to-air combat. *J. Guid. Control Dyn.* **13**(6), 1143–1149 (1990)
10. J. S. Ha, H. J. Chae, H. L. Choi, A stochastic game-theoretic approach for analysis of multiple cooperative air combat. *Am. Autom. Control Council*, 3728–3733 (2015). <https://doi.org/10.1109/acc.2015.7171909>
11. R. P. Wang, Z. H. Gao, Research on decision system in air combat simulation using maneuver library. *Flight Dyn.* **27**(6), 72–75 (2009)
12. V. Kai, T. Raivio, R. P. Hmlinen, Modeling pilot's sequential maneuvering decisions by a multistage influence diagram. *J. Guidance Control Dyn.* **27**(4), 665–677 (2004)
13. K. Virtanen, J. Kelahti, T. Raivio, Modeling air combat by a moving horizon influence diagram game. *J. Guidance Control Dyn.* **29**(5), 5 (2004)
14. H. Ehtamo, T. Raivio, On applied nonlinear and bilevel programming or pursuit-evasion games. *J. Optim. Theory Appl.* **108**(1), 65–96 (2001)
15. L. Zhong, M. Tong, W. Zhong, Application of multistage influence diagram game theory for multiple cooperative air combat. *J. Beijing Univ. Aeronaut. Astronaut.* **33**(4), 450–453 (2007)
16. Z. Liu, A. Liang, C. Jiang, Q. X. Wu, Application of multistage influence diagram in maneuver decision-making of ucav cooperative combat. *Electron. Opt. Control*. **33**(4), 450–453 (2010)
17. J. Kaneshige, K. Krishnakumar, in *Proceedings of SPIE - The International Society for Optical Engineering*, 6560:656009. Artificial immune system approach for air combat maneuvering, (2007)
18. N. Ernest, D. Carroll, C. Schumacher, M. Clark, G. Lee, Genetic fuzzy based artificial intelligence for unmanned combat aerialvehicle control in simulated air combat missions. *J. Defense Manag.* **06**(1) (2016)
19. N. Ernest, D. Carroll, C. Schumacher, M. Clark, G. Lee, Genetic fuzzy based artificial intelligence for unmanned combat aerialvehicle control in simulated air combat missions. *J. Defense Manag.* **06**(1), 1–7 (2016)
20. L. Fallati, A. Polidori, C. Salvatore, L. Saponari, A. Savini, P. Galli, Anthropogenic marine debris assessment with unmanned aerial vehicle imagery and deep learning: A case study along the beaches of the republic of maldives. *Sci. Total Environ.* **693**, 133581 (2019)
21. B. Neupane, T. Horanont, N. D. Hung, Deep learning based banana plant detection and counting using high-resolution red-green-blue (rgb) images collected from unmanned aerial vehicle (uav). *PLoS ONE*. **14**(10), 0223906 (2019)

22. Z. Jiao, C. G. Jia, C. Y. Cai, A new approach to oil spill detection that combines deep learning with unmanned aerial vehicles. *Comput. Ind. Eng.* **135**, 1300–1311 (2018)
23. X. Zhao, Y. Yuan, M. Song, Y. Ding, F. Lin, D. Liang, D. Zhang, Use of unmanned aerial vehicle imagery and deep learning unet to extract rice lodging. *Sensors (Basel, Switzerland)*. **19**(18) (2019). <https://doi.org/10.3390/s19183859>
24. C. Qu, W. Gai, M. Zhong, J. Zhang, A novel reinforcement learning based grey wolf optimizer algorithm for unmanned aerial vehicles (uavs) path planning. *Appl. Soft Comput. J.* **89**, 106099 (2020)
25. Z. X, Q. Zong, B. Tian, B. Zhang, M. You, Fast task allocation for heterogeneous unmanned aerial vehicles through reinforcement learning. *Aerosp. Sci. Technol.* **92** (2019). <https://doi.org/10.1016/j.ast.2019.06.024>
26. J. Yang, Y. X, G. Wu, M. M. Hassan, A. Almogren, J. Guna, Application of reinforcement learning in uav cluster task scheduling. *Futur. Gener. Comput. Syst.* **95**, 140–148 (2019)
27. S. D, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, Mastering the game of go with deep neural networks and tree search. *Nature*. **529**(7587), 484–489 (2016)
28. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, D. Hassabis, Mastering the game of go without human knowledge. *Nature*. **550**(7676), 354–359 (2017)
29. Y. Ma, W. Zhu, M. G. Benton, J. Romagnoli, Continuous control of a polymerization system with deep reinforcement learning. *J. Process Control.* **75**, 40–47 (2019)
30. Q. Zhang, R. Yang, L. X. Yu, T. Zhang, Z. J, Bvr air combat maneuvering decision by using q-network reinforcement learning. *J. Air Force Eng. Univ. (Nat. Sci. Ed.)* **19**(6), 8–14 (2018)
31. C. U. Chithapuram, A. K. Cherukuri, Y. V. Jeppu, Aerial vehicle guidance based on passive machine learning technique. *Int. J. Intell. Comput. Cybern.* **9**(3), 255–273 (2016)
32. X. Zhang, G. Liu, C. Yang, W. Jiang, Research on air combat maneuver decision-making method based on reinforcement learning. *Electronics*. **7**(11), 279 (2018)
33. B. Kurniawan, P. Vamplew, M. Papasimeon, R. Dazeley, C. Foale, in *AI 2019: Advances in Artificial Intelligence, 32nd Australasian Joint Conference, Adelaide, SA, Australia, December 2–5, 2019, Proceedings*. An empirical study of reward structures for actor-critic reinforcement learning in air combatmanoeuvring simulation (Springer, 2019), pp. 2–5
34. Q. Yang, J. Zhang, G. Shi, J. Hu, Y. Wu, Maneuver decision of uav in short-range air combat based on deep reinforcement learning. *IEEE Access*. **PP**(99), 1–1 (2019)
35. Q. Yang, Y. Zhu, J. Zhang, S. Qiao, J. Liu, in *2019 IEEE 15th International Conference on Control and Automation (ICCA)*. Uav air combat autonomous maneuver decision based on ddpg algorithm, (2019), pp. 16–19. <https://doi.org/10.1109/icca.2019.8899703>
36. H. C. Tien, A. Battad, E. A. Bryce, J. Fuller, A. Simor, Multi-drug resistant acinetobacter infections in critically injured canadian forces soldiers. *BMC Infect. Dis.* **7**(1), 1–6 (2007)
37. R. Z. Xie, J. Y. Li, D. L. Luo, in *2014 11th IEEE International Conference on Control and Automation (ICCA)*. Research on maneuvering decisions for multi-uavs air combat (IEEE, 2014)
38. M. Volodymyr, K. Koray, S. David, A. A. Rusu, V. Joel, M. G. Bellemare, G. Alex, R. Martin, A. K. Fidljeland, O. Georg, Human-level control through deep reinforcement learning. *Nature*. **518**(7540), 529–33 (2019)
39. T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning. *Comput. ence.* **8**(6), 187–200 (2015)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.