Multi-UAV Collaborative Search and Strike based on Reinforcement Learning

Bo Liu1, Xiaoping Wang1a* and Wen Zhou1

1Aeronautics Engineering College, Air Force Engineering University, Xi’an, Shaanxi, 710038, China

ashijingping@nwpu.edu.cn  *Corresponding author’s e-mail: 81875303@qq.com

Abstract. The problem of multi-UAV collaboration is one of the important research contents of multi-UAV systems. Aiming at the optimal strategy of multi-UAV collaborative search and strike against learning targets for moving targets in an unknown environment, a multi-UAV collaborative search and strike algorithm based on multi-team reinforcement learning is proposed. First, we establish a search probability map for the environment to search for moving targets. After finding the target, the moving target is allocated by auction to form an independent investigation team for different goals. Then each inspection team learns its optimal strategy in cycles at the same time. At the same time, by changing the learning speed and supervising the learning results during the learning process, the convergence and effectiveness of the learning results are guaranteed. The simulation results show that the UAV can better adapt to the dynamic environment through training. Under the condition of no prior information, it can effectively execute the search and strike task against the moving target with learning ability. And compared with the learning method of Markov decision process, the convergence speed and learning speed are faster.

1. Introduction

With the continuous development of computer technology, artificial intelligence, control technology and other high-tech, unmanned aerial vehicles (UAV) can self-perceive and make decisions during the execution of tasks, and are widely used in target search, reconnaissance and strike [1]. Therefore, how to understand the dynamic changes of the environment in real time, coordinate cooperation and make optimal decisions in the UAV group is one of the hot issues that have been widely concerned in the field of UAVs in recent years [2].

Aiming at the problem of multi-UAV collaborative search and strike. In [3], the author proposed a multi-UAV side-by-side search cooperative search model to improve the search ability of high-speed moving targets. But this article assumes that there is only one moving target in the environment. In [4], the author established a honeycomb environment model and proposed improved pigeon group optimization and Markov chain search algorithms, which improved the problems of static targets and low search efficiency. In [5], the author realized the effective search for random targets by establishing the model of centroid V graph division. In [6], the author calculated the coordinated strike position and pitch angle of the UAV by establishing the UAV capability function, realized the assignment of dynamic tasks, and improved the coordinated strike capability of multiple UAVs. The author in [7] has proposed an improved distributed ant colony algorithm for the UAV with integrated detection and strike, which can achieve the avoidance of threats and the investigation of missions.
Through research on the cooperative search and strike algorithm of multiple UAVs, there are still some limitations in the following two aspects: (1) The moving target does not have the ability to perceive and make decisions; (2) It is impossible to effectively search and strike multiple moving targets. Therefore, this paper is based on multi-team reinforcement learning to coordinate the behavior of each UAV. Through self-learning to improve the decision-making ability of coordinated detection of moving targets in unknown environments, so that it can better adapt to dynamic environments.

2. Problem Description
Assuming that the task area to be searched is a finite rectangle, some obstacles with different shapes, sizes and positions are set in the area to be searched. In this paper, the grid-based model [8] representation method is used to divide the task area (including obstacles) into \( L_x \times L_y \) cells, then the task area can be expressed as \( E = \{(i, j) | i = 1, 2, \cdots, L_x; j = 1, 2, \cdots, L_y\} \), \((i, j)\) represents the unit cell located in the \( i\)-th row and \( j\)-th column, as shown in figure 1. Among them, the black irregular graphics represent obstacles. The UAV moves one cell at a time. It is assumed that the effective attack radius of the UAV is 1 cell area, that is, the area surrounded by the green line. And the effective detection radius of the sensor is 2. The area of the cell is surrounded by the yellow line. The moving target also has the perception ability, and its effective detection radius is the same as that of the UAV, that is, the area surrounded by the red line.

![Figure 1. Rasterized discrete area model](image)

Suppose that there are \( n \) UAVs and \( m \) moving targets in the environment to be searched. \( U = \{U_1, \cdots, U_n\} \) represents the set of UAVs. \( A = \{A_1, \cdots, A_m\} \) represents the set of moving targets. Both UAVs and moving targets are learning agents with self-aware decision-making capabilities. At time \( t \in T = \{1, 2, 3, \cdots\} \), the position of the UAV is \( X_{U(t)} = (x_{u(i1)}, \cdots, x_{u(it)}) \), and the position of the moving target is \( X_{A(t)} = (x_{a(i1)}, \cdots, x_{a(it)}) \). It is stipulated that any two agents at the same time cannot be in the same position, and each agent has 8 possible directions of movement but can only perform one action.

There are four types of moving targets in the environment \( R_\alpha = \{\alpha \mid \alpha = \{I, II, III, IV\}\} \). Reflects the difficulty of different surveillance missions and the cost of completing the task. The UAV carries sensors with limited perception capabilities, which can recognize the type and value of moving targets and coordinate the fight against specific types of targets. Information on how much UAV cooperation is required for a moving target. Assuming that the moving target is within the effective attack range of the required multiple UAVs at the same time, the UAV will strike the moving target. That means to complete the search and strike task for this moving target.

3. Cooperative search and strike based on multi-team reinforcement learning
According to the characteristics of the multi-UAVs collaborative search and strike problem in an unknown environment. First search the environment. After the moving target is found, according to
the number of UAVs required for different types of moving targets, the investigation team is dynamically generated with other UAVs. Finally, the optimal surveillance strategy is solved according to the surveillance environment, the state of the UAV and the moving target.

3.1. Moving target search
For moving targets, search by establishing a search probability map [9]. At time \( t \), each cell \((i, j)\) of the grid map has a probability value \( p(i, j, t) \), which represents the search value of this cell. The set of probability values for the entire raster map is \( P \). Since the probability map information cannot be exchanged in real time between different UAVs. The probability map information of different UAVs is represented as \( P_k \), and \( k \) represents the number of the UAV.

When the UAV is searching. First, the information of each cell in the probability map is initialized to \( E \). At each moment of the search process, the target existence probability of the grid where the UAV is located is recorded as the minimum value \( E_{\min} \). With time, the \( p(i, j, t) \) of the grid gradually increases at a certain rate \( v \), but does not exceed the maximum value \( E_2 \). In order to make the UAV search the unknown field as much as possible. The initial value \( E_1 \) of the design cell is greater than the maximum value \( E_2 \) of the grid information that has been searched, that is \( E_1 > E_2 \). The update formula of \( P \) in the probability graph is shown in equation (1):

\[
p(i, j, t) = \begin{cases} E_1, & \text{Unsearched grid} \\ E_{\min}, & \text{Grid searched at a certain moment} \\ \max \times p(i, j, t), & \text{After searching the probability value starts to increase} \\ E_2, & \text{The maximum value that the searched grid probability value can reach} \end{cases}
\]

3.2. Form a collaborative search and strike team
This paper will create a merger between the search and strike team and the UAV collaborative search and strike target. And implement task allocation based on auction [10]. Reduce the scale of the multi-UAV search and strike system. The specific auction algorithm is as follows:

3.2.1. Basic concept description
The price of the sports target \( A \) at the time of auction is \( d \). The benefit that the UAV can obtain when it completes this target is \( b \). The UAV bids based on the benefit and price of the target. Then the value of the target to the UAV \( U \) is \( c = b - d \). Denoting \( U \) represents the set with the largest \( l \) before bidding on the target \( A \) in the UAV. It is expected that each target will be assigned to the set with the highest bid. As shown in equation (2):

\[
\max \sum_{i \in U} c_i = \sum_{i \in U} (b_i - d_i)
\]

When formula (2) is satisfied, it shows that the UAV combination scheme for this goal is satisfactory. If the number of UAVs cannot meet the requirements of the type of moving target, the auction will be suspended. Track the target and wait for more idle UAVs to be auctioned again.

The benefits that a UAV can detect when hitting a moving target can be calculated by matching the attributes of the target and the UAV. Here we take the attraction of the moving target from the UAV, the repulsive force of obstacles, the speed of movement, type and value as the task attribute standard.

Let the number of attribute requirements of the moving target \( A \) be \( R \), and the value of the \( K \)-th attribute requirement value is \( H_k \). The \( K \)-th attribute of the UAV \( U \) corresponding to the moving target attribute requirement value is \( H_a \). Then the benefit value \( b \) is calculated as (3) shown:

\[
b = \sum_{k=1}^{K} \lambda_k \times \frac{H_k}{H_a}
\]
Where, \( \lambda_i \) is the weight of each attribute, and \( \sum_{i=1}^{n} \lambda_i = 1 \).

3.2.2. Auction algorithm steps

Step 1: The UAV searches in an unknown environment according to the search algorithm;

Step 2: After the UAV finds the target, the UAV is equivalent to a manager. Obtain and publish information about sports goals, and the benefits that can be achieved by completing the investigation of the goals \( b_i \);

Step 3: Other UAVs accept the target's information. Bid according to the benefits and attributes required, and release the bid information to the manager;

Step 4: The manager waits for other UAVs to submit bid information. If the number of bidding UAVs does not meet the requirements of the type of moving target, go to step 7;

Step 5: The manager sets the price of the sports target to \( d_j \), and ranks the bids of other UAVs from high to low;

Step 6: The manager chooses the first \( l \) UAVs to form a team with himself. Collaborate to inspect the sport target, and release the winning bid information to other UAVs. It is not allowed to bid for other targets, go to step 8;

Step 7: Suspend the auction. The manager tracks the un-auctioned targets and waits for the auction. If the bidding UAV meets the quantity requirements, go to step 5;

Step 8: End the auction.

3.3. UAV collaborative search and strike based on multi-team reinforcement learning

After the UAV confirmed its investigation team, it began to siege the moving target. In this paper, through the introduction of reinforcement learning methods [11], the optimal decision-making under unknown environment is achieved, and the global benefits are maximized while accomplishing the goal of combating movement. The key to UAV learning in the multi-UAV surveillance system is the processing of other UAVs, that is, the estimation of state values must consider the joint actions of all UAVs. In this paper, through the multi-team reinforcement learning method (MTRL), the joint state behavior space divided based on the search probability graph makes the multi-cooperative team learn the optimal decision at the same time.

3.3.1. Multi-team reinforcement learning

UAVs form multiple surveillance teams for different sports goals. Each team member coordinates their behaviors. According to the game learning idea, in each learning cycle, according to the state of the \( s \), the optimal behavior strategy selection is based on the action behavior to be taken by other UAVs in this state as the best response action. After repeated countermeasures many times, all UAVs are satisfied. The optimal equilibrium state of the UAS is the Nash equilibrium of each UAV in the current state. This paper uses Bayesian formulas and probability statistics to estimate the strategic knowledge of other UAVs to update the Q value.

Denote the probability that UAV \( U_i \) believes that UAV \( U_j \) may take action \( a_k \) in state \( s \) as \( p'(s, j, a_k) \), UAV \( U_i \) and \( U_j \) belong to the same inspection team. If UAV \( U_j \) performs different actions in state \( s \) is \( N_i^j \). The number of times to perform action \( a_k \) is \( N_i^j(k) \). Then the calculation of \( p'(s, j, a_k) \) is as shown in equation (4):

\[
p'(s, j, a_k) = \frac{N_i^j(k)}{N_i^j} \tag{4}
\]

Before learning starts, set the initial value of \( p'(s, j, a_k) \) to \( 1/N_i^j \), where \( N_i^j \) is the number of actions the UAV can choose in the initial state. Each UAV of the same inspection team takes its own action in state \( s \). After UAV \( U_i \) observes the new state \( s' \) after the joint action and other UAV
actions. According to the Bayesian formula, update the probability that the UAV $U_i$ will take its own actions to other UAVs in the state $s$. The update formula is shown in equation (5):

$$p(u' | u', s') = \frac{p(s' | u', u') \times p(u')}{p(s' | u')}(5)$$

Where, $p(s' | u', u')$ is the transition probability of the UAV $U_i$ and $U_j$ to reach the new state $s'$ after taking a joint action. $p(s' | u')$ is the transition probability of the UAV $U_i$, taking the action to reach the new state $s'$ alone. The estimate of UAVs $U_i$ to $p(u')$ can be obtained by $p'(s, j, a_k)$ defined above. Then the update formula can be transformed as shown in equation (6):

$$p(u' | u', s') = \frac{p(s' | u', u') \times p'(s, j, a_k)}{p(s' | u')}(6)$$

All action strategies of UAV $U_i$ are denoted as $a_k$, and all joint action strategies of other UAVs are denoted as $a'_k$. The corresponding joint action probability is calculated as shown in equation (7):

$$p(a'_k | a_k, s') = \prod_{i,j} p(u' | u', s')(7)$$

Then the Q value of UAV $U_i$ is updated as shown in equation (8):

$$Q_{i,t}(s, a_k, a'_k) = (1 - \alpha_t)Q_{i,t}(s, a_k, a'_k) + \alpha_t[R_i + \beta Q_{j,t}(s')]\prod_{i,j} p(u' | u', s')(8)$$

The original Q value of UAV $U_i$ is replaced by the maximum Q value of the action. So that other UAVs choose the value obtained by the best joint action corresponding to their own action, and choose the optimal balance.

### 3.3.2. Learning result revision

Each UAV updates a Q value table to represent the Q function. Since different surveillance teams cannot be completely independent during the learning process, mutual influence may cause the learning results to oscillate or not converge. Therefore, by accelerating the learning speed and learning Results supervision, to ensure the convergence of learning.

(1) Correct the learning speed

The update of the Q value is mainly affected by $p'(s, j, a_k)$. The relative square error method can be used to evaluate the confidence of $p'(s, j, a_k)$ to exclude actions that do not meet the confidence level conditions. Thereby speeding up the learning speed. The calculation of the relative square error is shown in equation (9):

$$E = \left(\frac{\sum p'(s, j, a_k) - 1}{n^j}\right)^{1/2}(9)$$

Where, $n^j$ is the number of actions that the UAV $U_j$ can choose in any state, then the confidence calculation of $p'(s, j, a_k)$ is shown in formula (10):

$$\eta_{p'(s, j, a_k)} = \sum_{k=1}^{n^j} p'(s, j, a_k) / n^j \cdot E(10)$$

In order to evaluate the confidence of $p'(s, j, a_k)$, the threshold should be set to satisfy $\eta_{p'(s, j, a_k)} \in [a, b]$, in general $a, b$ meets $a, b \geq 0.5$.

(2) Supervised learning results

In the learning process, the action selection of different teams is inspected. When the Q value shows a tendency to oscillate or does not converge. The learning rate is controlled by changing the learning rate. The convergence of the learning results is guaranteed. Use $\Delta Q^t_r$ to represent the continuous learning period The absolute instantaneous differential Q value. As shown in equation (11):
\[
\Delta Q_T = \sum |Q_T(s,a_t,a_t') - Q_{T-1}(s,a_t,a_t')|, T \geq 1
\] (11)

Where, \( n_{r}(s,a_t,a_t') \) represents the number of triples \((s,a_t,a_t')\) explored in a learning cycle, and \( Q_{T}(s,a_t,a_t') \neq Q_{T-1}(s,a_t,a_t') \). Use equation (11) to judge the learning result. If \( \Delta Q_T \) is decreasing, it indicates that the learning result is convergent. Otherwise, use equation (12) to reduce the learning rate.

\[
\alpha_T = (n_{r}(s,a_t,a_t'))^{-1} \cdot \alpha_{T-1}
\] (12)

4. Computer simulation
In order to verify the feasibility and effectiveness of the proposed method, several simulation experiments were carried out. Set the unknown environment to a rectangular area of 300*300. Discretize into a grid area of 60*60 according to the size of 5*5. There are 10 UAVs, 4 different types of moving targets and a certain number of static obstacles in the environment. UAVs and targets are agents with the ability to perceive decisions and learn. The visual sensor and distance sensor it carries can perceive all the information of the grid where it is located. It is assumed that the sensor is not affected by obstacles. When a moving target is within the range of multiple UAVs attacked by its type at the same time, it represents the completion of the search and strike task for this target.

During the learning process, the UAV observes the current state and selects actions to obtain instant rewards \( R \). In the event of a collision, the system reward is -10. Help to check the sport goal reward is 30. The UAV coordinated to the target to get 300 rewards each. Remuneration in other circumstances-2. Set the parameters of the multi-team reinforcement learning algorithm, when \( t = 0, \gamma = 0.95, Q_{T}(s,a_t,a_t') = 1, \alpha_T = 1 \). 20 experiments were carried out, and 3000 trials were carried out for each experiment. After the learning is completed, the algorithm is tested. Figures 2(a)-(d) are the simulation results corresponding to the different observation stages. Figure 2(a) is the initial time distribution map of the UAV, moving targets and obstacles. Figure 2(b) is the trajectory of the UAV searching for the target. Figure 2(c) is the formation of a reconnaissance team to coordinate the detection of the moving target. Figure 2(d) is the final result of the reconnaissance. In the initial state, the red solid small circle represents the UAV, and the blue solid small circle represents the moving target. When the target is found, the UAV inspects the team and is represented by the purple solid circle.

![Multi-UAV coordinated detection of multiple moving targets in an unknown environment](image)

Figure 2. Multi-UAV coordinated detection of multiple moving targets in an unknown environment

In order to verify the convergence and learning speed of the proposed algorithm, it is compared with the learning method of Markov decision process. The average results of the two learning methods output after 20 experiments are shown in Figure 3.
Figure 3. MDPL and MTRL algorithm comparison

It can be seen from Figure 3. (1) In terms of convergence. At the beginning of learning, MTRL and MDPL algorithms have relatively close curves because there are few pairs of state actions to be explored. As the learning progresses, the MTRL algorithm converges faster, and the MDPL algorithm converges to the same Q value; (2) Learning speed. If there are \( m \) UAVs in the environment, the state action log of the MDPL algorithm is \( P^m \cdot 8^n \). The MTRL algorithm proposed in this paper is that multiple independent investigation teams learn at the same time, and the number of UAVs in the largest number of teams is \( n \), and \( n < m \). Then the state action pair explored by the MTRL algorithm is \( P^n \cdot 8^n \). Therefore, the reduction of the state action pair makes the MTRL learning speed better than the MDPL algorithm.

5. Conclusion

Through the study of reinforcement learning. Based on the multi-team reinforcement learning algorithm, a method of collaboratively searching for moving targets with perception and learning capabilities under the unknown environment is proposed. The simulation results show that:

(1) Discretization of the area to be searched using a grid model and the establishment of a search probability map can effectively achieve the search for moving targets. (2) Based on the auction algorithm to achieve task allocation, the merger of the reconnaissance team and the coordinated surveillance target of unmanned aerial vehicles will be combined to reduce the scale of the coordinated surveillance system of multiple unmanned aerial vehicles. (3) The proposed multi-team reinforcement learning method enables multi-collaborative teams to learn optimal decisions at the same time, and corrects the learning results to ensure the convergence of the algorithm and the speed of learning.

References

[1] Barkaoui M., Berger J., Boukhtouta A. (2019) An evolutionary approach for the target search problem in uncertain environment. J. Journal of Combinatorial Optimization, 38(1–2).
[2] Ha, I.-K.; Cho, Y.-Z. (2018) A Probabilistic Target Search Algorithm Based on Hierarchical Collaboration for Improving Rapidity of UAVs. J. Sensors, 18, 2535.
[3] Zeng G. Q., Bai Y., Lin W., et al. (2018) Multi-UAV cooperative search method for ground moving targets. J. Systems Engineering and Electronics, v.40; No.466(07):87-94.
[4] Wang R., Xiao B. S. (2019) Cooperative search for multi-UAVs via an improved pigeon-inspired optimization and Markov chain approach. J. Chinese Journal of Engineering,41(10):1342-1350.
[5] Xuan Y. B., Huang C. Q., Wu W. C., et al. (2013) Cooperative search strategies of multi-UAVs for random targets. J. Control and Decision, 28(05):711-715.
[6] Chen X., Qiao Y. Z. (2018) Task Allocation Method for Multi-UAV Cooperative Combat. J. Electronics Optics & Control, 25(06):20-24.
[7] Gao C. (2016) Autonomous Mission Planning Method for Multiple Unmanned Aerial Vehicles. D. Nanjing University of Aeronautics and Astronautics.

[8] Wang X. Q. (2018) Optimal Search Strategy for UAVS based on Markov Mobile Targets. D. Henan University.

[9] Li P., Duan H. (2017) A potential game approach to multiple UAV cooperative search and surveillance. J. Aerospace Science & Technology, 68(sep.):403-415.

[10] Zhao M. M., Li B., Wang M. L. (2015) Auction Algorithm Based Multi-UAV Arriving Simultaneously to Attack Multiple Targets with Uncertain Information. J. Electronics Optics & Control, 22(2).

[11] Kong X., Xin B., Wang Y., et al. (2017) Collaborative Deep Reinforcement Learning for Joint Object Search. C. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.