applied sciences

MDPI

Article

A Novel Searching Method Using Reinforcement Learning Scheme for Multi-UAVs in Unknown Environments

Wei Yue [1,2], Xianhe Guan [1] and Liyuan Wang [2,*]

[1] School of Marine Electrical Engineering, Dalian Maritime University, Dalian 116000, China; weiy@dlmu.edu.cn (W.Y.); xhg@dlum.edu.cn (X.G.)

[2] Key Laboratory of Intelligent Perception and Advanced Control of State Ethnic Affairs Commission, Dalian Minzu University, Dalian 116600, China

* Correspondence: wangliyuandmu@163.com; Tel.: +86-138-8968-2365

Received: 28 July 2019; Accepted: 14 November 2019; Published: 18 November 2019

Abstract: In this paper, the important topic of cooperative searches for multi-dynamic targets in unknown sea areas by unmanned aerial vehicles (UAVs) is studied based on a reinforcement learning (RL) algorithm. A novel multi-UAV sea area search map is established, in which models of the environment, UAV dynamics, target dynamics, and sensor detection are involved. Then, the search map is updated and extended using the concept of the territory awareness information map. Finally, according to the search efficiency function, a reward and punishment function is designed, and an RL method is used to generate a multi-UAV cooperative search path online. The simulation results show that the proposed algorithm could effectively perform the search task in the sea area with no prior information.

Keywords: multi-UAV; cooperative search; reinforcement learning; dynamic target

1. Introduction

With the rapid development of sensors, wireless communication, intelligent control, and other technologies, the functions and the application fields of unmanned group systems are increasing day by day. Because of their expansibility, strong cooperation, and low loss, the cooperative theory and applied research on unmanned group systems received increasing attention in the fields of academia, industry, and national defense [1]. Multi-unmanned aerial vehicle (UAV) cooperative search systems can effectively improve the search efficiency, especially for the search of dynamic targets under complex sea conditions such as uncertainty, strong interference, and so on. Therefore, multi-UAV cooperative sea area searching is one of the important research directions of unmanned group systems [2].

The multi-UAV area target search problem was widely studied by research groups around the world. Generally, the target search problem is divided into two categories: static target search and moving target search. For static targets, the traditional search method is a covering search (e.g., echo search, traversing search, etc.) [3–6]. This search method generally maximizes the coverage of the task area to find as many targets as possible. In References [6,7], a search map model was established according to the existence probability of target, and distributed model predictive control was used to solve the problem, which effectively reduced the solution scale of the search decision problem. In Reference [8], considering the limitations of sensing and communication capabilities, coverage and topology control algorithms were designed for the path planning of mobile agents. For dynamic targets, a Bayesian method was used to calculate the average detection time and average detection probability [9,10], but it was only suitable for searching a single target. In Reference [11], a target motion prediction model based on Markov chain was proposed, and a greedy iterative decision method

based on distributed model predictive control was designed to solve the problem. For a mobile target in a closed and bounded region, a receding-horizon cooperative search algorithm was presented [12]. With the goal of minimizing search time, a strategy of "the reward of discount time" was proposed, and a cross-entropy optimization algorithm was used to find the location where the target has the highest probability as soon as possible [13]. In Reference [14], the problem of scheduling multiple UAVs to search for missing tourists was addressed, and a method for estimating tourist location probabilities which change with topographic features, weather conditions, and time was proposed. In addition, some intelligent algorithms were also applied to such problems. In Reference [15], a novel objective function that naturally and coherently integrates the conflicting objectives of target detection, target tracking, and vehicle survivability into a single scalar index was designed, and a modified particle swarm optimization algorithm was used to determine which trajectory is the best on average at detecting and tracking targets. Combined with the traditional Maximum-Q-learning (MAX-Q) algorithm, a multi-UAV cooperative search strategy was proposed, which effectively completes the search tasks of independent targets and clustering targets but is limited to static targets [16,17]. In References [18,19], a reinforcement learning (RL) algorithm was used for path planning without colliding with obstacles in unknown environments.

In order to solve the above problems, this paper proposes a multi-UAV cooperative search method based on RL, which fully considers the characteristics of an unknown sea area. It synthesizes a target probability map and certainty value map to establish a multi-UAV sea area search map, and the concept of the territory awareness information map is proposed to coordinate the tracks between multiple UAVs. The search efficiency function is obtained according to the extended search map. A new reward and punishment function is designed by using the search efficiency function. The multi-UAV search track can be planned online, according to the efficiency of the reinforcement learning method, and the search map can be updated with the search results. Finally, the effectiveness of the algorithm is verified by simulation experiments.

2. Materials and Methods

In this paper, for a specialized sea area E which is built using two-dimensional coordinates, there are unknown targets $\{Target_i, i = 1, 2, \ldots, N_t\}$, unknown no-fly zones $\{Mence_i, i = 1, 2, \ldots, N_m\}$, and known homogenous UAVs $\{V_i, i = 1, 2, \ldots, N_v\}$, where N_t, N_m, and N_v are the numbers of targets, no-fly zones, and UAVs, respectively. Airborne UAVs enter the sea area where the mission needs to be carried out. After that, each UAV uses its own onboard sensor to search the unknown targets independently. It is expected that multiple UAVs can find as many targets as possible through cooperative search in the shortest time with the least cost. In this research, we assumed that all UAVs could communicate through a relay station to ensure that the communication between the UAVs was normal. Below, we provide detailed descriptions of our environmental model, our collision avoidance strategy, the aerial vehicle model, and our objective.

2.1. Environmental Model

The environment E is represented as an $L_x \times L_y$ sea area, and the information for the search map can be defined as $P_{mn}(k) = [\ p_{mn}(k) \quad \chi_{mn}(k)\ ]$, where $p_{mn}(k) \in [0, 1]$ is the existence probability of the targets on the grid (m, n) with $(m \in \{1, 2, \ldots, L_x\}, n \in \{1, 2, \ldots, L_y\})$ at instant k. $\chi_{mn}(k) \in [0, 1]$ is the certainty value, where $\chi_{mn}(k) = 1$ represents that the UAVs fully understand the target information, and $\chi_{mn}(k) = 0$ represents that the UAVs have no information on the target.

Target probability map: Before the search task begins, the search map is given a certain initial value, which reflects the prior information of the target (obtained by external intelligence reconnaissance). With the continuous search, the search map information mastered by the UAVs is constantly updated.

Considering the influence of sensor uncertainty, the updating equation of the target probability was designed as follows:

$$p_{mn}(k+1) = \begin{cases} \tau p_{mn}(k) + \Delta p_{mn}(k) & no\ access \\ \frac{p_D p_{mn}(k)}{p_F + (p_D - p_F) p_{mn}(k)} & access \cap b_k = 1 \\ \frac{(1-p_D) p_{mn}(k)}{1 - p_F + (p_F - p_D) p_{mn}(k)} & access \cap b_k = 0 \end{cases}, \quad (1)$$

where p_D and p_F are the sensor detection rate and false alarm rate, respectively. $\tau \in [0, 1]$ denotes a discount factor, which represents the forgetting factor of the probability map. Since only a few grids are accessed at the same time, and these grids' probability change affect other grids, the changing probability $\Delta p_{mn}(k)$ in grid (m, n) is defined as

$$\Delta p_{mn}(k) = \sum_{(i,j) \in D(k)} [p_{ij}(k) - p_{ij}(k+1)] / (L_x \times L_y - N_v), \quad (2)$$

where $D(k)$ is a set of all accessed grids at time k. When UAVs access (m, n), the update of $p_{mn}(k)$ is related to the detection variable b_k of the platform sensor. $b_k = 1$ represents that the airborne sensor detected the target, and $b_k = 0$ represents that the sensor did not detect the target.

Certainty value map: As the search task proceeds, UAVs have a constant understanding of the search area. The certainty value map reflects the degree of understanding of the whole map. Due to the probability of sensor detection and false alarm, $\chi_{mn}(k)$ is used to represent the degree of determination of the information at the grid (m, n) by UAVs at time k. When the grid is not accessed, the certainty of the grid decreases. As the number of times the grid is detected increases, the degree of determination of the information at the grid is increased by the UAVs. The update equation is as follows:

$$\chi_{mn}(k+1) = \begin{cases} \tau_c \chi_{mn}(k), & no \quad access \\ \chi + (1 - \chi) \chi_{mn}(k), & access \end{cases}, \quad (3)$$

where τ_c is the information factor of certainty; $\chi \in [0, 1]$ is a constant, and its value is related to sensor performance. $\chi = 1$ represents that the UAV can fully grasp the information when it carries out information testing; that is, $p_D = 1$ and $p_F = 0$. $\chi = 0$ represents that the information is not available, which is equivalent to the complete failure of the sensor.

2.2. Effect of Collision Avoidance on Search Map

When multiple UAVs perform tasks together, security is a top priority. This paper draws lessons from the idea of hormone information dissemination and diffusion, and the concept of the territory awareness information map [20] is used to establish a search map. When a UAV moves to the grid (m, n) at time k, the pheromone information is generated at the corresponding location of the search map, which affects the generation and updating of other UAVs' pheromones by diffusion. The diffusion of existing pheromones inhibits the generation of other kinds of pheromones, which is the role of the territory awareness information map. A new environmental search information map can be constructed, which can be used as the basis to solve the collision avoidance problem in multi-UAV cooperative control.

$H_{mn}(k)$ is defined as the total pheromone concentration at the grid (m, n). The concentration is a function of the grid position and time; thus, the environment search map is rewritten as follows: $P^*_{mn}(k) = [\ p_{mn}(k) \quad \chi_{mn}(k) \quad H_{mn}\ (k)]$.

When V_i searches grid (m, n), it produces pheromone $H_{i(mn)}(k)$, which can diffuse to other grids in the search map. Taking the grid (a, b) as an example, the diffusion propagation function is

$$H_{i(ab)}(k) = \frac{\beta}{\rho^2} \times e^{-\frac{(a-m)^2 + (b-n)^2}{2\rho^2}}, \quad (4)$$

where ρ and β are constants.

When N_v UAVs perform search tasks, N_v kinds of pheromones are constantly generated and diffused. Taking the grid (c, d) as an example, the concentration of a pheromone at the current time is the sum of the concentration of pheromones left by volatilization at the previous time and the newly produced pheromone diffusing to the grid. The updating equation is as follows:

$$H_{cd}(k) = \tau_H H_{cd}(k-1) + \sum_{i=1}^{N_v} H_{i(cd)}(k), \tag{5}$$

where $\tau_H \in [0, 1]$ is a volatility factor.

When V_i detects a high concentration of other kinds of pheromones at grid (m, n), it means that other UAVs' activities are frequent at grid (m, n). That is, flying V_i into the grid not only reduces the search efficiency, but also has a high probability of collision. The concentration of other kinds of pheromones detected by V_i is as follows:

$$\overline{H}_{imn}(k) = H_{mn}(k) - \sum_{j=0}^{k} H_{imn}(j). \tag{6}$$

2.3. Unmanned Aerial Vehicle Model

In this research, the UAVs used in the search task have the same performance and keep flying at a specific altitude. Under the inertial reference coordinates, the motion model can be described as follows (see References [4,21] for details):

$$\begin{cases} \dot{x}_i = v_i \cos\phi_i \\ \dot{y}_i = v_i \sin\phi_i \\ \dot{\phi}_i = u_i \eta_{\max i} \end{cases}, \tag{7}$$

where $p_i = (x_i, y_i) \in R^2$ is the position state of the ith UAV in the search plane, and φ_i and v_i are respectively yaw angle and speed, which needs to satisfy $v_i \in [v_{\min}, v_{\max}]$. $u_i \in [-1, 1]$ is a decision variable. $\eta_{\max i}$ is the maximum turn angular velocity, and is constrained by the performance of the ith UAV.

At each decision time k, the UAV can take actions such as left deviation $\eta_{\max i}$, direct flight, or right deviation $\eta_{\max i}$. For simplicity, the control decision variable and state variable can be written as $\mathbf{u}(k) = [u_i(k), i = 1, 2, \ldots, N_v]^{\mathrm{T}}$ and $\mathbf{s}(k) = (\mathbf{s}_i(k), i = 1, 2, \ldots, N_v)$, with $u_i(k) \in \{-1, 0, 1\}$ and $\mathbf{s}_i(k) = [x_i(k), y_i(k), \varphi_i(k)]^{\mathrm{T}}$. In the process of multi-UAV cooperative search, in order to avoid collision between UAVs, D is defined as the minimum safe distance, and the requirement is as follows:

$$d = \sqrt{(x_i(k) - x_j(k))^2 + (y_i(k) - y_j(k))^2} \geq D, \tag{8}$$

where d is the actual distance between UAVs.

Due to no-fly zones in the mission sea area, the position of UAVs should satisfy

$$(x_i(k), y_i(k)) \notin Mence_j, (i = 1, 2, \ldots, N_v, j = 1, 2, \ldots, N_m). \tag{9}$$

In this research, the no-fly zones were set to be circular, and the position constraint of UAVs can be described as

$$\sqrt{(x_i(k) - X_j)^2 + (y_i(k) - Y_j)^2} > D^*, \tag{10}$$

where (X_j, Y_j) is the center coordinate of $Mence_j$, and D^* is the radius of the no-fly zone.

2.4. Sensor Model

The range of the airborne sensors is an important factor to measure the instantaneous search area for the UAV. As shown in Figure 1, a visible light sensor was installed at a fixed angle. Then, in the relative coordinate, the detection width can be obtained as follows:

$$d_u = 2 \cdot h_u \cdot \tan \gamma_u / \sin \alpha_u, \tag{11}$$

where h_u is the UAV flight altitude, α_u is the installation angle of the sensor, and γ_u is the sensor's horizontal field of view.

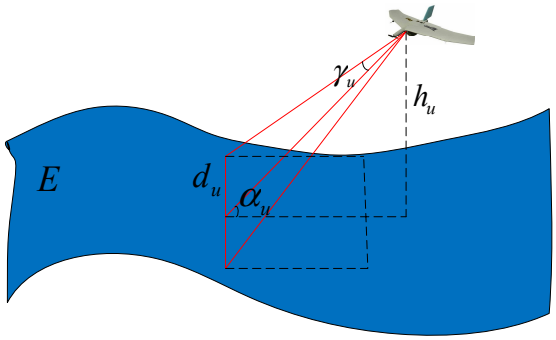

Figure 1. Airborne sensor detection model.

2.5. Multi-Objective Function Establishment

The main goal was to find as many targets as possible, on the premise of ensuring safety. Therefore, the optimization function $J(s(k), u(k))$ is composed of the benefit of the target found J_p, the benefit of searching environment J_χ, the cost of execution C, and the cost of collision I, which are described below.

The benefit of discovering targets: J_p is the possibility of discovering the target by UAVs in flight. It can be described as the sum of the target probabilities at the area R_i^n occupied by the trajectories. It is defined as

$$J_p(k) = \sum_{i=1}^{N_v} \sum_{(m,n) \in R_i^n} [(p_D - p_F) p_{mn}(k) + p_F]. \tag{12}$$

The benefits of searching the environment: as the task proceeds, the UAVs can obtain information on the search area, which means the entropy [6] search map is gradually reduced. Thus, J_χ is defined as the decrement of the entropy of information.

$$J_\chi(k) = H(k) - H(k+1), \tag{13}$$

where $H(k) = -\sum_{(m,n) \in E} (1 - \chi_{mn}(k)) \ln(1 - \chi_{mn}(k))$ is the entropy of information at instant k.

The cost of execution: The cost C is the consumption of time and fuel during the task, which can be estimated as

$$C(k) = \sum_{i=1}^{N_v} \|p_i(k) - p_i(k+1)\| / v_i(k). \tag{14}$$

The cost of collision: Collision avoidance is the primary consideration of cooperative search for multiple UAVs in the same plane. Combined with the characteristics of the territory awareness information map, the concentration of pheromones is high at the locations which are current grid locations of each UAV or where flight tracks are dense. Therefore, when the ith UAV takes a control

decision, the direction of flight is where the pheromone is lower. This decision reduces the possibility of collision with other UAVs, and I is defined as

$$I(k) = \sum_{i=1}^{N_v} \overline{H}_{imn}(k). \tag{15}$$

According to the analysis, the overall efficiency function of the UAVs' cooperative search problem is as follows:

$$J(s(k), u(k)) = w_1 J_p(k) + w_2 J_\chi(k) - w_3 C(k) - w_4 I(k), \tag{16}$$

where $0 \le w_i \le 1 (i = 1, 2, 3, 4)$ are weights. Note that the above functions have different dimensions; thus, it is necessary to normalize them separately before summation.

3. Design of Cooperative Search Strategy

In this section, the Q-Learning method is used to design the yaw angle decision $\mathbf{u}(k)$ of UAVs. When UAV i is at the state $\mathbf{s}_i(k) = [x_i(k), y_i(k), \varphi_i(k)]^{\mathrm{T}}$, the corresponding row of the state in the Q-table is set as the $\mathbf{s}_i(k)$ row, in which each value represents the effect of a control decision. The decision corresponding to the maximum value is the optimal decision $u_i(k)$, and the optimal decision set $\mathbf{u}(k)$ is acquired.

3.1. Establishment of Q-Value Table

Since the table method is vulnerable to the problem of "dimension disaster", here, the state and control input of UAVs are simplified as much as possible when the table is designed. The possible location of a UAV is determined by the total number of grids. There are n yaw angles at each grid; thus, the number of rows in the Q-table is $L_x \times L_y \times n$, which is the number of UAV states. There are m optional control inputs for each UAV; thus, the number of columns in the Q table is m, which is the number of decisions contained in decision set A. Variable $Q(\mathbf{s}_i(k), u_i(k))$ is defined as the value that V_i selects the decision $u_i(k)$ in the state $\mathbf{s}_i(k)$.

In the initial stage of learning, because there is no prior information, the decision in the Q-value table is made randomly; thus, Q-learning should have more opportunities to explore the unknown decision space. If only the maximum Q-value is used, the algorithm converges quickly to a poor $\mathbf{u}(k)$. If the random selection strategy is used, although the environmental state information can be fully explored and the optimal strategy can be found, the algorithm converges too slowly. Therefore, designing a reasonable decision selection mechanism to achieve balance between exploring information and Q-value can ensure the fast convergence of the algorithm to a better strategy.

In this paper, the Boltzmann distribution mechanism is used to select the decision in the Q-learning process. That is, the probability that the policy set $\mathbf{u}(k)$ is selected in the state $\mathbf{s}(k)$ is determined as follows:

$$P(\mathbf{u}(k)) = \frac{e^{Q(\mathbf{s}(k), \mathbf{u}(k))/T}}{\sum\limits_{\mathbf{u} \in A} e^{Q(\mathbf{s}(k), \mathbf{u})/T}}, \tag{17}$$

where $\mathbf{u} \in A$ represents that strategy $\mathbf{u}$ is an enforceable strategy in decision set A. The value of T determines the ability of learning to explore unknown spaces, and as T increases, the ability to explore new decision spaces is improved (if T is infinite, it is a random decision because $P(\mathbf{u}(k)) = 1/m$). T is defined as

$$T = T_0 \cdot M^{-1/\lambda}, \tag{18}$$

where $\lambda > 1$, $T_0 > 0$, and M is the number of iterations of the algorithm.

In the initial stage of learning, T is set to be large in order to explore more decision space $(T = T_0)$. Upon increasing the number of learning times, T gradually decreases in Equation (18), increasing the empirical effect of the Q-value and speeding up the convergence of the algorithm.

3.2. Q-Value Update Process

When the ith UAV is in state $\mathbf{s}_i(k) = [x_i(k), y_i(k), \phi_i(k)]^{\mathrm{T}}$, the policy $u_i(k)$ is selected according to the largest Q-value in the $\mathbf{s}_i(k)$ row in the Q-table. After executing it and arriving at state $\mathbf{s}_i(k+1)$, then the immediate reward or penalty value is used to update $Q(\mathbf{s}_i(k), u_i(k))$ in the Q-table. If the UAV gets a reward, $Q(\mathbf{s}_i(k), u_i(k))$ increases; that is, when the ith UAV is next in the state $\mathbf{s}_i(k) = [x_i(k), y_i(k), \phi_i(k)]^{\mathrm{T}}$, the maximum Q-value corresponding to the decision is selected. On the contrary, when the penalty value is obtained, the Q-value corresponding to the decision $u_i(k)$ becomes smaller until it is not the maximum Q-value, then the values of other decisions are selected. When the Q-table finally converges, the optimal decision $u_i(k)$ is obtained, from which the updating rule of the Q-value function can be obtained.

$$Q(\mathbf{s}_i(k), u_i(k)) = (1-\alpha)Q(\mathbf{s}_i(k), u_i(k)) + \alpha[r(k) + \gamma \max_{u \in A} Q(\mathbf{s}_i(k+1), u)], \tag{19}$$

where $\mathbf{s}_i(k)$ is the current state of the ith UAV, $u_i(k)$ is the decision of the current choice, which is the variation of yaw angle, $r(k)$ is the immediate reward value or penalty value, $\max_{u \in A} Q(s_i(k+1), u)$ represents the maximum Q-value obtained by the policy u in the state $\mathbf{s}_i(k+1)$, and $\alpha \in [0, 1]$ is the learning rate. When $\alpha = 1$, the original Q-value has no effect on the new Q-value, and all knowledge learned is new, but it is easy to cause Q-value instability; when $\alpha = 0$, the Q-value remains unchanged, and the learning stops. Therefore, α determines the learning ability of the algorithm. γ is the discount factor, and $\gamma = 1$ means that there is no discount on the delay return; that is, it attaches great importance to the influence of the current decision on the future. $\gamma = 0$ means that the Q-value does not calculate the delay return (it only calculates the immediate return); thus, γ determines the importance that the learning algorithm attaches to the delay return. Below, we describe the specific design of $r(k)$ in Equation (19) and the strategy selection in the learning process.

3.3. Design of Reword and Punishment Function

In this paper, we consider that multiple UAVs perform search tasks in the framework of reinforcement learning and obtain a higher overall efficiency $J(\mathbf{s}(k), \mathbf{u}(k))$. If a UAV gets a higher efficiency when it executes a search task, it is rewarded immediately. If it gets a lower efficiency, it is punished immediately. Therefore, the reward and punishment functions are designed as follows:

$$r(k) = \begin{cases} R & , d \geq D \,\&\, discover \quad target \\ a \times J(\mathbf{s}(k), \mathbf{u}(k)), & d \geq D \,\&\, no \quad target \\ -R & , d < D \end{cases}, \tag{20}$$

where a is a constant which influences the generalization ability of the learning process, and $a \times J(\mathbf{s}(k), \mathbf{u}(k)) \in (-R, R)$, R, and $-R$ are the maximum reward and punishment, respectively. $J(\mathbf{s}(k), \mathbf{u}(k))$ is determined by Equation (16). In order to ensure the safe flight of each UAV, $d \geq D$ needs to be satisfied. If the no-fly zone is considered, B should be greater than the radius of the no-fly zone D^* where B is the distance between the UAV and the center of the no-fly zone. At this time, the reward and punishment functions are rewritten as follows:

$$r(k) = \begin{cases} R, & d \geq D \& B \geq D^* \,\&\, discover \quad target \\ a \times J(\mathbf{s}(k), \mathbf{u}(k)), & d \geq D \& B \geq D^* \,\&\, no \quad target \\ -R, & d < D \| B < D^* \end{cases}. \tag{21}$$

That is, if a UAV may collide or fly into the no-fly zone, the maximum punishment is employed.

Algorithm 1 Q-Learning of cooperative search

Input:
Initialize unknown $L_x \times L_y$ search areas: E
Initialize the Q-table that represents the state-decision model and corresponding parameters n λ.
Initialize the state of multiple UAVs $\mathbf{s}(k) = (\mathbf{s}_i(k), i = 1, 2, \ldots, N_v)$.
Start:
For episode = 1 to M do
Initialize the state of multiple UAVs, Get the initial state $\mathbf{s}(k)$.
For k = 1 to T do
A decision is randomly selected with
$P(\mathbf{u}(k)) = \frac{e^{Q(\mathbf{s}(k),\mathbf{u}(k))/T}}{\sum_{\mathbf{u}\in A} e^{Q(\mathbf{s}(k),\mathbf{u})/T}}$ as probability. After the decision is executed, the multiple UAVs reach a new round of state $\mathbf{s}(k+1)$, and the return $r(k)$ is calculated according to the reward and punishment function:

$$r(k) = \begin{cases} R & , \ d \geq D \& \ \ discover \quad target \\ a \times J(\mathbf{s}(k), \mathbf{u}(k)), & d \geq D \& \ \ no \quad target \\ -R & , \ d < D \end{cases}.$$

Update the status of the multiple UAVs $\mathbf{s}(k) = \mathbf{s}(k+1)$, and replace $r(k)$ with the update formula of the Q-value.

$$Q(\mathbf{s}_i(k), u_i(k)) = (1-\alpha)Q(\mathbf{s}_i(k), u_i(k)) + \alpha[r(k) + \gamma \max_{u\in A} Q(\mathbf{s}_i(k+1), u)].$$

Update the Q-value.
End for
End for
Output: Q-table

The complexity of an iterative Algorithm 1 is about $O(\|L\|) + O(N_v \times m \times k)$, where $\|L\|$ is the scale of the problem space. Thus, the time complexity of the completion of all the iterations is $O(M \times \|L\|) + O(M \times N_v \times m \times k)$, where the number of iterations of the algorithm is M. The space complexity of the algorithm is $O(\|L\|) + O(N_v \times m)$. For a given multi-UAV, the overall computational cost is determined by M, m, and k.

4. Simulation Results

In order to verify the effectiveness of the algorithm, a multi-UAV cooperative search simulation environment was established in MATLAB. The information in the search area was completely unknown, and the purpose of the search was to identify all targets and possible trends in the sea area. The effectiveness of the algorithm was verified by comparative simulation, which was aimed at the targets of the independence and the formation, respectively. The relevant parameters in the simulation were set as follows: the weighted parameters were $w_1 = 0.25$, $w_2 = 0.15$, $w_3 = 0.1$, $w_4 = 0.5$, the decision-making period was 20 s, and the speed of the UAV was 30 m/s. The speed of the warships was 9 kn. The sensor parameters were $p_D = 0.9$ and $p_F = 0.1$, and the search map parameters were $\tau = 0.98$, $\tau_c = 0.9$, and $\tau_H = 0.9$.

4.1. Independent Random Distribution of Targets

The nine warship targets were randomly distributed in an unknown sea area with a range of 10 nm × 10 nm, which excludes the already controlled sea area and shallow sea area. Each warship carried out its mission independently, and four UAVs searched the warships. The initial positions of UAVs were located at the four corners of the sea area. In the initial stage (k = 1 to 120) of RL, due to the lack of prior target information, the whole sea area was searched. As shown in Figure 2, the four curves with different colors are the trajectories of the four UAVs. Because the UAVs did not know the specific location, shape, or size of the shallow area, they would slip into the shallow area for a search. However, the targets were unable to enter these areas. With continuous learning, the information

mastered by the UAVs increased, and the shallow area (represented by ×), the sea area controlled by the red side (circular sea area), and the position of warships (pentagonal stars and diamonds in the Figure 2) were gradually identified.

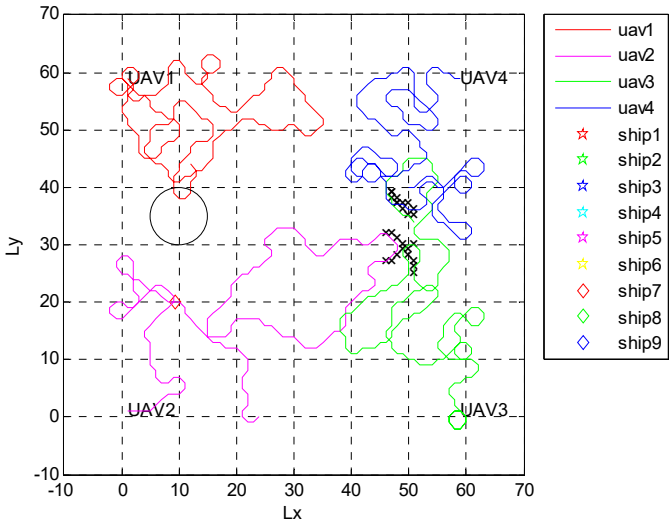

Figure 2. Reinforcement learning (RL) search initial stage.

Figure 3 shows the search trajectories of UAVs over a period of time with the increase in learning time. UAVs detected shallow water with a size of 1.00 nm × 2.50 nm, and the horizontal and longitudinal coordinate ranges were respectively 46 to 51 and 25 to 39. Meanwhile, the circular sea area controlled by the red side with a center coordinate of (10, 35) and a radius of 0.83 nm was also detected. It can be seen that, after learning and understanding the sea area information, UAVs no longer searched the sea area where it was impossible for the target ship to appear. Figure 4 shows the sea area information mastered by the UAVs after the search mission, including the sea area covered by the shallow area, and the nine warships cruising along a straight line.

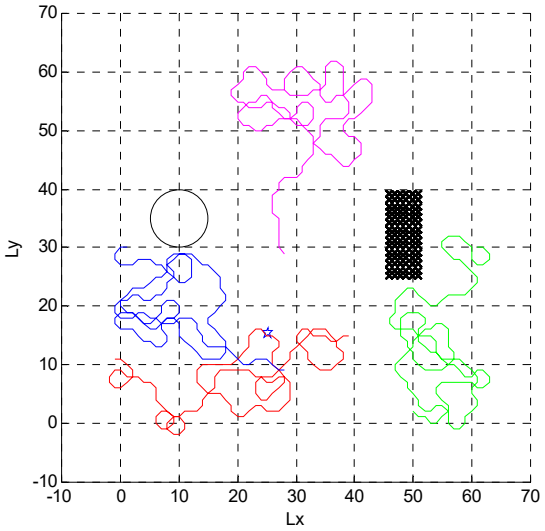

Figure 3. Sea area information learning search.

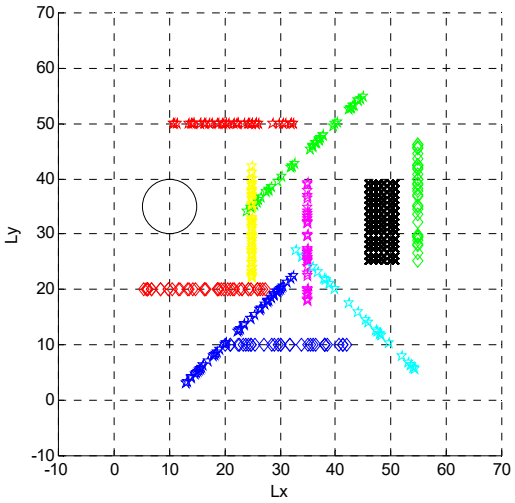

Figure 4. Unmanned aerial vehicle (UAV) search result under proposed method.

Case 1—Dynamic targets: Under this condition, the search effect based on the Q-learning algorithm was compared with random search and traversing search, and the Monte Carlo method was used for 500 experiments. Random search is a non-cooperative search method. When the UAV makes a decision, it randomly selects a direction to search. The traversing search is a fixed search mode. Full coverage of the area can be achieved as long as the duration of flight is allowed. The simulation diagrams of random search and traversing search are shown in Figures 5 and 6. The comparison results are shown in Table 1. The numbers of targets found by the three search methods were compared, and it is clear that the efficiency of the RL search was the highest; at each statistical moment, is the value was about one more than that of the random search. Over time, traversing searches may also find all targets, but the efficiency is extremely low because the targets are moving. This result is due to the fact that the random search and traversing search do not update the search map and, thus, search for invalid areas. At this point, the RL search shows obvious advantages.

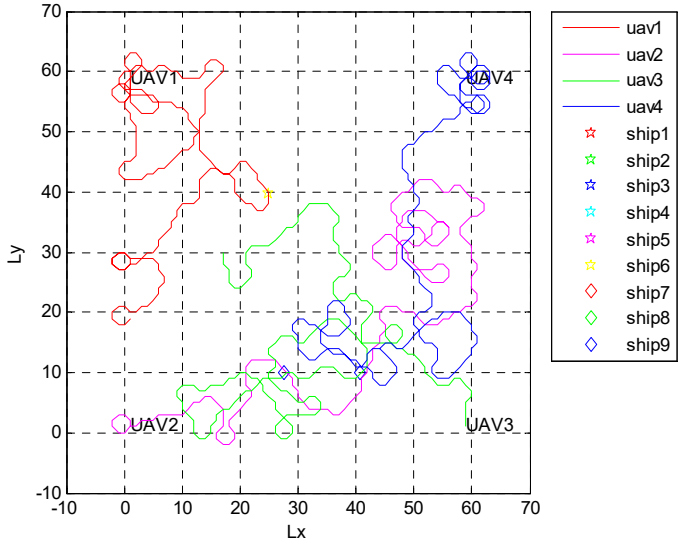

Figure 5. Random search result.

Table 1. Comparison of the number of targets found using dynamic targets.

Method \ Time (s)	1000	2000	3000	4000	5000	6000	7000
Random Method	0.396	0.794	1.210	1.702	2.262	2.652	3.050
Traversing Method	0.336	0.638	0.894	1.262	1.626	2.196	2.532
Proposed Method	1.188	1.734	2.154	2.672	3.068	3.660	4.166

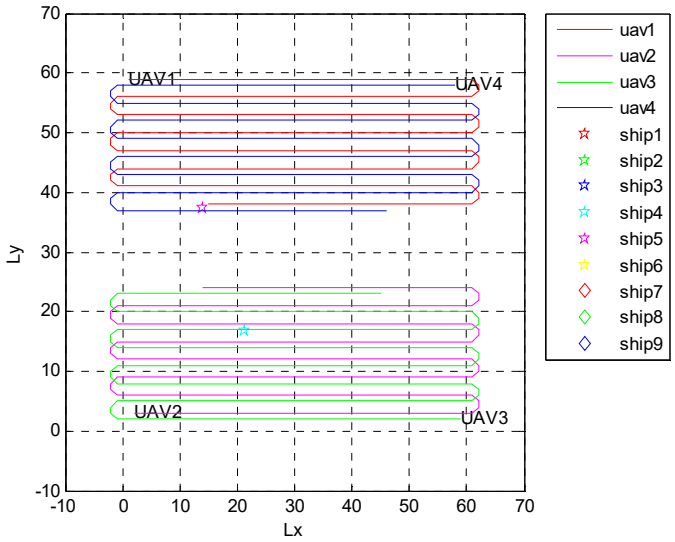

Figure 6. Traversing search result.

Case 2—Static targets: An experiment with the initial probability of targets according to a uniform distribution was designed to verify the search efficiency of the algorithm for different numbers of targets.

In total, 500 simulations were carried out to obtain the average number of targets found in various cases, and the running time of each simulation was 2000 s. The search results of the three methods for different numbers of targets are compared in Table 2.

In this scene, as long as the search track of the UAVs can completely cover the whole mission area E, most of the targets can be found in the case of reliable detection by airborne sensors. The random search is a blind search method, which leads to repeated searches of the already searched sea area, and reduces the coverage of the sea area during the task time. Therefore, for the random search, the fewest targets were found. For the probability of static target discovery, the traversing search was 6.333% better than when using dynamic targets. If the whole area of the search is traversed and the sensor meets the accuracy requirements, all the targets are found. The method proposed in this paper can better cover the region where the target has a high probability; thus, the number of targets found was more than that of the random search and traversing search.

Table 2. Comparison of the numbers of targets found using static targets.

Method \ Total Number	5	9	14	20	25	30
Random Method	0.442	0.950	1.836	2.534	3.224	3.822
Traversing Method	0.626	1.204	2.128	2.862	3.598	4.074
Proposed Method	1.028	2.122	3.376	5.252	7.326	8.544

4.2. Formation Targets

The warships traveled in the sea area in a v-shaped formation in this paper. The v-shaped formation consisted of a main warship at the front end, with four vessels arranged on either side. Four UAVs were used to search for the blue warships in the unknown sea area. In the initial stage (k = 1 to 120), UAVs did not have any prior information, meaning that there was still a need for a full map search. Figures 7 and 8 are simulation diagrams of the four red UAVs searching for the blue v-shaped warship configuration. It can be seen that, with the progress of learning, the multiple UAVs gradually reduced the search frequency of undiscovered ships in the sea area, and focused on searching the active area of the warships until all targets were found. Because of the large difference between the speed of warships and UAVs, warships were found in the same position for a certain period of time, which continuously increased the reward value and formed a closed loop. As shown in Figure 9, the red UAVs' flight path formed a ring.

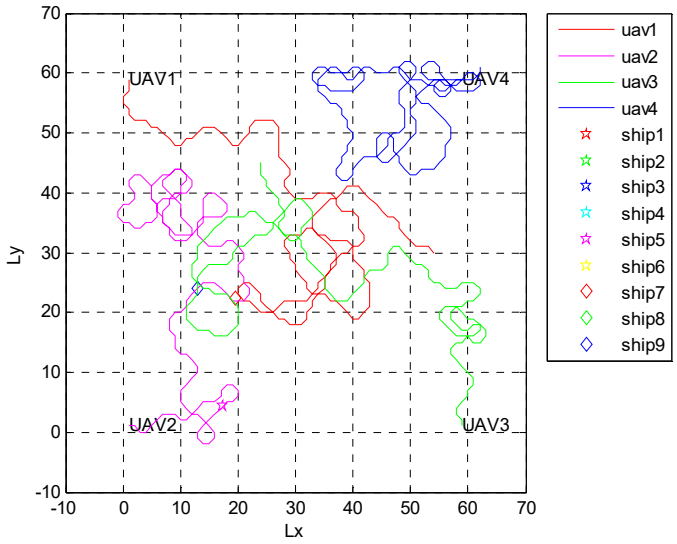

Figure 7. The initial stage of the v-shaped formation search.

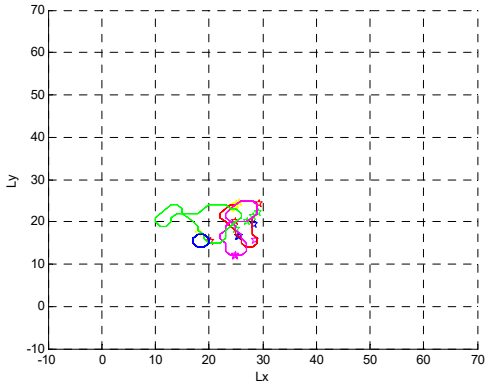

Figure 8. RL search of the v-shaped formation.

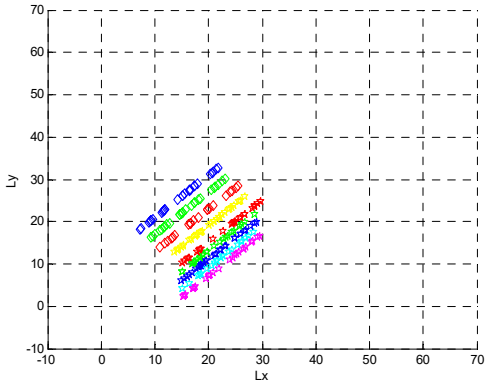

Figure 9. Search result under RL.

Figure 9 shows the red UAVs' search results when nine blue warships cruised in a straight line in a v-shaped formation.

4.3. Algorithm Parameter Analysis

At the beginning of the algorithm iteration, if the reward value is over utilized, the algorithm converges quickly, which leads to missing the optimal solution. When different values of λ are taken, the change curve of T is as shown in Figure 10. The initial temperature T_0 was 500. The corresponding policy selection probability is shown in Figure 11. Each λ corresponded to three curves, which represented the three strategies. The rise of the curve (short dotted line) was due to the strategy having the maximum reward. On the contrary, the fastest falling curve (full curve) was due to the minimum reward. When $\lambda = 1.8$, the algorithm converged after about 200 iterations. Under these conditions, if the decision space is particularly large, several better solutions would be missed. When $\lambda = 5$, the algorithm was too slow to converge, and the efficiency was low.

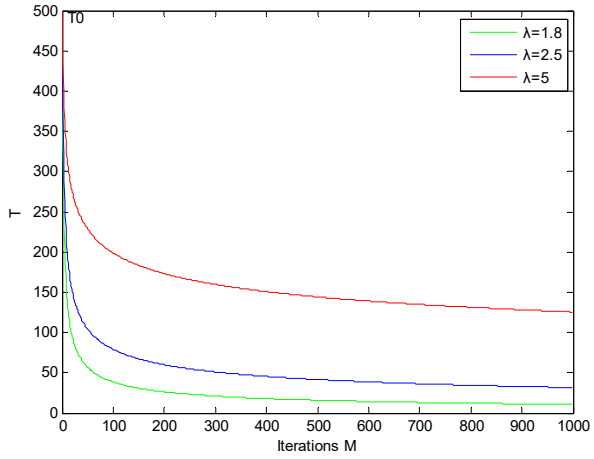

Figure 10. Profile of T variation.

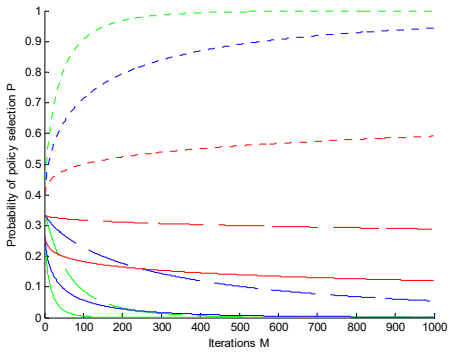

Figure 11. Relation between P and M.

5. Discussion

In this paper, a multi-UAV cooperative search algorithm based on RL was proposed to solve the problem of multi-UAV cooperative dynamic target searching in unknown sea areas. According to the comprehensive efficiency function, a reward and punishment function was designed. At the same time, the target probability map and certainty value map were mixed to describe the unknown sea area, and the territory awareness information map was introduced to coordinate the cooperation between multiple UAVs. The extended search map can be updated online according to the search situation of multiple UAVs. The simulation results showed that the algorithm was effective, and the multi-UAV cooperative dynamic target search was verified by comparative analysis, which was more effective than the original search method.

This paper did not consider issues induced by the communication network shared by the UAVs. In practice, the communication networks may have constraints such as transmission delays, packet dropouts, and bandwidth issues, which may degrade the performance. Exploiting the effect of the communication constraints to derive effective algorithms for multi-UAV cooperative searches is an interesting topic worthy of investigation in the next step of research work.

Author Contributions: Conceptualization, W.Y.; methodology, W.Y. and X.G.; validation, W.Y. and L.W.; investigation, Y.W.; writing—original draft preparation, W.Y. and X.G.; writing—review and editing, L.W.; supervision, W.Y.; project administration, W.Y.; funding acquisition, W.Y.

Funding: This research was funded by the Natural Science Foundation of China (grant number 61703072), the Dalian Science and Technology Innovation Fund (grant number 2019J12GX040), the Fundamental Research Funds for the Central Universities (grant number 3132019355), the Liaoning Natural Science Foundation Project (grant number 20180551118), and the Key Laboratory of Intelligent Perception and Advanced Control of State Ethnic Affairs Commission (grant number MD-IPAC-201901).

Acknowledgments: We would like to thank Cunming Zou from Dalian University of Science and Technology for his help with UAV tests.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chung, T.H.; Hollinger, G.A.; Isler, V. Search and pursuit-evasion in mobile robotics. *Auton. Robot.* **2011**, *31*, 299–316. [CrossRef]
2. Cui, X.T.; Yang, R.J.; He, Y. Modeling and simulation of multi-ship coordinated spiral call-searching. *Ship Sci. Technol.* **2010**, *32*, 95–98.
3. Yu, Y.D.; Zhang, D. Modeling and Simulation of Alpine Leaf search and potential search with Magnetometer. *Ship Electron. Eng.* **2017**, *37*, 88–92.
4. Peng, H.; Shen, L.C.; Huo, X.H. A Study of Multi-UAVs Cooperative Area Coverage Search. *J. Syst. Simul.* **2007**, *19*, 2472–2476.
5. Yao, P.; Xie, Z.; Ren, P. Optimal UAV Route Planning for Coverage Search of Stationary Target in River. *IEEE Trans. Control Syst. Technol.* **2017**, *27*, 822–829. [CrossRef]

6. Angley, D.; Ristic, B.; Moran, W.; Himed, B. Search for targets in a risky environment using multi-objective optimisation. *IET Radar Sonar Navig.* **2019**, *13*, 123–127. [CrossRef]
7. Huang, L.; Qu, H.; Ji, P.; Liu, X.; Fan, Z. A novel coordinated path planning method using k-degree smoothing for multi-UAVs. *Appl. Soft Comput.* **2009**, *48*, 182–192. [CrossRef]
8. Hu, J.; Xie, L.; Lum, K.Y.; Xu, J. Multiagent Information Fusion and Cooperative Control in Target Search. *IEEE Trans. Control Syst. Technol.* **2013**, *21*, 1223–1235. [CrossRef]
9. Bourgault, F.; Ktogan, A.; Furukawa, T. Coordinated search for a lost target in a Bayesian world. *Adv. Robot.* **2004**, *18*, 979–1000. [CrossRef]
10. Kassem, M.A.; El-Hadidy, M.A. Optimal multiplicative Bayesian search for a lost target. *Appl. Math. Comput.* **2014**, *247*, 795–802.
11. Shem, A.G.; Mazzuchi, T.A.; Sarkani, S. Addressing Uncertainty in UAV Navigation Decision-Making. *IEEE Trans. Aerosp. Electron. Syst.* **2008**, *44*, 295–313. [CrossRef]
12. Riehl, J.R.; Collins, G.E.; Hespanha, J.P. Cooperative Search by UAV Teams: A Model Predictive Approach using Dynamic Graphs. *IEEE Trans. Aerosp. Electron. Syst.* **2011**, *47*, 2637–2656. [CrossRef]
13. Pablo, L.P. *Minimum Time Search of Moving Targets in Uncertain Environments*; Universidad Complutense de Madrid: Madrid, Spain, 2013.
14. Du, Y.C.; Zhang, M.X.; Ling, H.F. Evolutionary Planning of Multi-UAV Search for Missing Tourists. *IEEE Access* **2019**, *7*, 480–492. [CrossRef]
15. Pitre, R.R. An Information Value Approach to Route Planning for UAV Search and Track Missions. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 2551–2565. [CrossRef]
16. Matsuda, A.; Misawa, H.; Horio, K. Decision making based on reinforcement learning and emotion learning for social behavior. In Proceedings of the IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 27–30 June 2011.
17. Cai, Y.; Yang, S.X.; Xu, X. A combined hierarchical reinforcement learning based approach for multi-robot cooperative target searching in complex unknown environments. In Proceedings of the 2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Singapore, 16–19 April 2013; pp. 52–59.
18. Pham, H.X.; La, H.M.; Feil-Seifer, D.; Van Nguyen, L. Reinforcement Learning for Autonomous UAV Navigation Using Function Approximation. In Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Philadelphia, PA, USA, 6–8 August 2018; pp. 1–6.
19. Hung, S.M.; Givigi, S.N. A Q-Learning Approach to Flocking with UAVs in a Stochastic Environment. *IEEE Trans. Cybern.* **2016**, *47*, 186–197. [CrossRef] [PubMed]
20. Poole, W.E. Field enclosure experiments on the technique of poisoning the rabbit, Oryctolagus cuniculus (L.). II. A study of territorial behaviour and the use of bait stations. *Csiro Wildl. Res.* **1963**, *8*, 28–35. [CrossRef]
21. Jin, Y.; Liao, Y.; Minai, A.A.; Polycarpou, M.M. Balancing search and target response in cooperative unmanned aerial vehicle (UAV) teams. *IEEE Trans. Cybern.* **2006**, *36*, 571–587. [CrossRef] [PubMed]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).