IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED COMMUNICATIONS AND NETWORKING TECHNIQUES FOR WIRELESS CONNECTED INTELLIGENT ROBOT SWARMS

Received November 12, 2019, accepted November 29, 2019, date of publication December 9, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958328

Anti-Jamming Communications in UAV Swarms: A Reinforcement Learning Approach

JINLIN PENG[1], ZIXUAN ZHANG[2], QINHAO WU[3], AND BO ZHANG[1]

[1]Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing 100010, China
[2]College of Computer, National University of Defense Technology, Changsha, China
[3]College of Electronic Science, National University of Defense Technology, Changsha, China

Corresponding author: Bo Zhang (bo.zhang.airc@outlook.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 91648204 and Grant 61601486, in part by the Research Programs of National University of Defense Technology under Grant ZDYYJCYJ140601, and in part by the State Key Laboratory of High Performance Computing Project Fund under Grant 1502-02.

ABSTRACT Intelligent unmanned aerial vehicle (UAV) swarm may accomplish complex tasks through cooperation, relying on inter-UAV communications. This paper aims to improve the communication performance of intelligent UAV swarm system in the presence of jamming, by multi-parameter programming and reinforcement learning. This paper considers a communication system, where the communication between a UAV swarm and the base station is jammed by multiple interferers. Compared with the existing work, the UAVs in the system can exploit degree-of-freedom in frequency, motion and antenna spatial domain to optimize the communication quality in the receiving area. This paper proposes a modified Q-Learning algorithm based on multi-parameter programming, where a cost is introduced to strike a balance between the motion and communication performance of the UAVs. The simulation results show the effectiveness of the algorithm.

INDEX TERMS Intelligent UAV swarm, anti-jamming communication, multi-parameter joint programming, antenna pattern, motion cost.

I. INTRODUCTION

With the rapid development of artificial intelligence, intelligent unmanned aerial vehicle (UAV) is widely used in daily life [1], [2]. Intelligent UAVs have been used in disaster relief [3], environmental monitoring [4], marine search, rescue [5] and other fields. However, the energy constraint of the UAVs [6], [7], along with the scarcity of spectrum resources and interference, the communication design within the intelligent UAV swarm become an important constraint on its practical application [8], [9].

The communication system of an intelligent UAV swarm is shown in Figure 1. Firstly, the UAVs are communication nodes of high mobility, and each UAV may serve as transmitters, relays and receivers. Secondly, UAV swarm demands multi-channel access and networking in the presence of jamming, but the omni-directional antenna generally used for inter-UAV coordination may reduce spectrum efficiency.

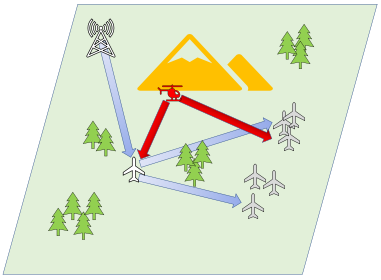


FIGURE 1. Diagram of communication scenario.

Thirdly, a UAV is usually energy-constrained, which should be taken into account in UAV communication design.

At present, the UAV systems mainly rely on anti-jamming techniques such as frequency hopping and spread spectrum to maintain reliable point-to-point control links between the ground station and the UAVs, which provide limited data transmission capabilities due to channel utilization penalties [10], [11]. These solutions may not support the high-speed data transmission required in UAV swarms, such as inter-UAV data consensus and multi-UAV coordination [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Hui Cheng.

Against the background, this paper considered the joint communication-motion-antenna programming in UAV swarms to combat jamming [13]. Specifically, autonomous frequency selection, beam synthesis and motion control in intelligent UAV swarm are jointly optimized to improve the anti-jamming capabilities. However, two challenges need to be addressed. Firstly, it is difficult to capture and model the effects of complex electro-magnetic environment on swarm UAV communication system. Secondly, it is challenging to solve the problem of joint communication-motion-antenna programming in UAV swarms. Therefore, this paper is devoted to the design of effective algorithm which may generate reasonable action policy for intelligent UAV swarm.

In order to address the above challenges, the reinforcement learning approach is adopted. Firstly, reinforcement learning does not need to model the complex environment as a whole, and it only needs to evaluate all candidate actions taken by the UAVs. Secondly, reinforcement learning is unsupervised learning so that it directly models and analyzes data for generating policies. Specifically, a multi-dimensional anti-jamming reinforcement learning (MDAJRL) algorithm is proposed, which effectively solves the problems brought by the environment and its effectiveness is verified by simulation experiments.

In this paper, a swarm UAV communication system model is constructed. Specifically, A multi-dimensional ''frequency-motion-antenna'' parameter space is constructed to support the decision-making process of the UAV communication system. Then, a multi-dimensional anti-jamming reinforcement learning algorithm based on energy constraints is proposed. Considering the limited energy of the UAVs, an energy constraints module is added, which effectively improves the decision-making effect. Under the condition of multiple UAV receivers, the algorithm tunes the antenna beam for improving the overall communication quality of the receiving UAVs.

This paper is organized as follows: In Section 2, we review the recent work related to networked UAV communication and anti-jamming based on reinforcement learning. Section 3 gives the system modelling, including the modelling of antenna pattern and spectrum. In Section 4, we propose MDAJRL algorithm based on joint programming of multi-dimensional parameters, and introduce a cost module to restrict the motion of the relay UAV. Section 5 illustrates the simulation results, which prove the effectiveness of the proposed algorithm. We conclude our work in Section 6.

II. RELATED WORKS

UAV communication network is widely applied in military and civil fields, such as ground surveillance, anti-jamming, data transmission and other fields. In the research of UAV anti-jamming communication network, [14] proposed a power control and channel selection method based on correlation vector regression, which gave the minimum power of data transmission link. Reference [15] proposed a power allocation scheme with cooperative anti-jamming policy to improve the anti-jamming ability of ad hoc network communication under limited resources. In [16], by optimizing the transmitting power of communication UAV and jammer, it could improve the confidentiality to the greatest extent. A comprehensive tutorial of aeronautical ad-hoc communication is given by [17], summarizing the research in mobility model, network scheduling and routing. Reference [18] proposed an adaptive coding and modulation for high-rate data transmission in aviation communications.

At present, there are some anti-jamming communication researches based on reinforcement learning. Reference [19] studied the anti-jamming communication policy under unknown environment. They used the spectral waterfall as the basis for establishing the Markov decision process (MDP) and achieved spectrum programming under jamming conditions. Reference [20] extended the parameters of communication system to joint frequency-motion programming based on a hotbooting deep Q-network. It accelerated the iteration process and improved the ability of the agent to resist jamming. Reference [21] designed a deep reinforcement learning method for heterogeneous information fusion and realizes intelligent anti-jamming in high frequency band. Reference [22] proposed a multi-agent cooperative anti-jamming algorithm. It could not only effectively avoid external malicious jamming, but could also deal with the interference between users. Reference [23] used the broadband spectrum sensing capability of cognitive radio to accelerate the learning process of reinforcement learning, and achieved a better policy set than traditional Q-Learning. Reference [24] proposed a real-time reinforcement learning algorithm based on Q-Learning, which achieved better real-time policy by using broadband spectrum sensing and greedy policy.

Most research of UAV ad-hoc network considered power allocation, adaptive coding and modulation according to mitigate jamming from the environments, which achieve better communication quality and data transmission rate. In the field of anti-jamming research based on reinforcement learning, most researchers built Markov decision process based on feedback of electromagnetic environment to facilitate programming in the spectrum domain. Some studied the optimization of learning speed, which improved the efficiency of iteration [25]. Some considered the programming of geographic parameters and used the motion of UAVs to optimize the model [20], [26].

In summary, the joint programming of multi-parameters in UAV communication systems is not fully studied in the literature. Many of the above literatures only regard spectrum or power allocation as the optimization objective, while the location of the agent and other parameters are not fully used, hence limiting the decision-making and anti-jamming performance.

III. SYSTEM MODEL

A. MODELING BACKGROUND

This paper considers the intelligent programming of communication network composed of UAV swarms and ground

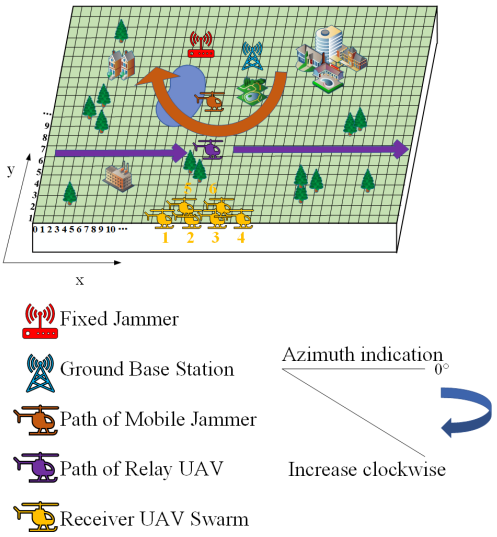


FIGURE 2. Grid modeling of the system.

base stations under jamming conditions. In this scenario, the communication system is composed of a ground base station, a relay UAV and a UAV swarm. The communication system is jammed maliciously by adversarial jammer UAVs.

The above scenarios are modeled as shown in Figure 2. In order to quantify the parameters of each agent conveniently, grid modeling is introduced, as shown in Figure 2. The position of the agent is represented by two-dimensional coordinates $[x, y]$, and the beam pointing is represented by the azimuth angle shown in the Figure 2. The ground base station, the relay UAV and receiver UAV swarm constitute the communication system. Its task is to transmit the data from the ground base station to the relay UAV (communication link 1), and then extract the spectrum from the relay UAV and forward it to the receiver UAV swarm (communication link 2). It is assumed that the direct link is negligible due to high pathloss. The frequency bands used in the two transmission processes are different, so there is no problem of mutual interference.

The ground base station is fixed and its transmitting frequency is tunable. The relay UAV moves in a certain area. Its transmitting frequency, beam pointing, mainlobe width and position are tunable. After receiving signal from the ground base station, the relay UAV filters and forwards the signal in the transmitting signal band, and transmits the signal to the receiver UAV swarm in the new band. The receiver UAV swarm is composed of multiple UAVs, and the swarm distributes into a receiving area of certain shape.

There are two types of jammers. The fixed jammer is the long-distance ground-based jamming equipment. Its position and antenna mainlobe width remain unchanged, and its frequency and beam pointing are tunable. The fixed jammer can track and jam the relay UAV. The mobile jammer is deployed over an UAV, which may adapt the frequency and position, while keeping the beam pointing and mainlobe width. The mobile jammer only jams the receiver UAV swarm. The tunable parameters of the above agents are given in Table 1.

TABLE 1. Tunable parameter list of agents.

Agent	Frequency	Beam pointing	Mainlobe	Position
Ground Base Station	√			
Relay UAV	√	√	√	√
Fixed Jammer	√	√		
Mobile Jammer	√			√

In the case of malicious jamming, the communication system enables multi-parameter decision-making, so that each receiver UAV may receive signal with better quality. To ensure the communication quality for every UAVs in the swarm, the UAV formation is essentially different from the common channel power allocation model. In traditional channel power allocation research, a channel gain is often allocated to each channels and its policy is to adjust channel gain parameters under different jamming modes. In this paper, the aim is to improve the signal-to-interference power (SIR) of the whole communication system by combining the three dimensions of "frequency-motion-antenna". Since the shape of the UAV swarm formations may be irregular, a certain number of receivers are arranged in this scenario, and their average SIR level represents the communication status of the swarm.

B. ANTENNA PATTERN MODELING

Antenna pattern refers to the pattern of antenna radiation intensity distribution with spatial angle. Considering that the relay UAV mainly carries small phased array antenna (uniform linear array), according to the general theory of phased array antenna, the pattern of uniform linear array is as follows:

$$E(\theta) = \left| \frac{1}{N} \frac{\sin[\frac{N}{2}(\frac{2\pi}{\lambda} d_1 \sin\theta - \varphi)]}{\sin[\frac{1}{2}(\frac{2\pi}{\lambda} d_1 \sin\theta - \varphi)]} \right| \tag{1}$$

where N is the number of elements, d_1 is the spacing of elements, φ is the phase difference of elements, λ is the operating wavelength, and the corresponding carrier frequency is f. Since the radiation energy of phased array antenna pattern is generally concentrated in the mainlobe and the first sidelobe, in order to simplify the modeling process, the pattern is approximated by the sinc function of $y = \sin(ux)/ux$. Figure 3 shows the comparison of the true pattern and the approximate pattern under the conditions of $N = 70$, $d_1 = 1\text{cm}, f = 2.4\text{GHz}, u = 1.8\pi$.

It can be seen that the sinc function has good approximation in the range of mainlobe and first sidelobe. Therefore, the sinc function is used to model the antenna pattern. Because the pattern with wider mainlobe radiates more energy, it is necessary to restrict the pattern energy of antenna. All antenna patterns are normalized according to the antenna

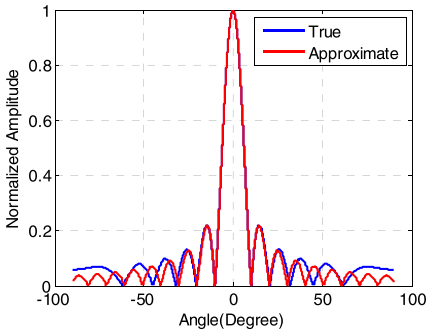


FIGURE 3. Comparison of true and approximate patterns.

pattern energy when the zero power mainlobe width is 30°. That is, when the antenna mainlobe is wider than 30°, the maximum value of the pattern is less than 1.0, otherwise it is greater than 1.0.

IV. ALGORITHM DESIGN

Based on the system model of Section 3, this section designs a reinforcement learning algorithm to solve the multi-parameter optimization in UAV communication system.

A. PROBLEM FORMULATION

At present, reinforcement learning is an important anti-jamming method, which does not need to model the environment. It is to produce correct policies through the interaction between agents and environment where the reward suggests the feedback from environment to agents. The purpose of reinforcement learning algorithm in our communication system is to generate the adaptive policies given the jamming form, so as to maximize the SIR of the whole receiving area. Therefore, it is necessary to set up a reasonable reward standard according to the system topology. Thus, the calculation method of SIR is given first.

Since the continuous pulse in time domain can be regarded as a continuous rectangular window function, its spectrum can also be regarded as a sinc function. When designing the spectrum, we consider that the width of the mainlobe of the spectrum is the zero-power bandwidth of the signal, and the central angle of the mainlobe is the position of the current carrier frequency. The signal transmission process is that the relay UAV receives the spectrum transmitted by the ground base station and the fixed jammer at the same time. The relay UAV extracts and forwards the two spectrums according to the range of the transmitting signal band. The receiver UAV swarm receives the converted signal forward by the relay UAV and the signal from mobile jammer together. Then the receiver UAV swarm calculates the SIR of the signal within the transmitting band of relay UAV. The parameters of each agent are shown in Table 2.

The instantaneous spectrum of communication signals received by a receiver UAV can be expressed by:

$$P_{rt} = S_t P_t L_t F_1 R_{g1} R_{g2} L_r \tag{2}$$

TABLE 2. The parameters of each agent.

Parameters	Variable
Transmitting power of ground base station	P_t
Spectrum sequence of ground base station	S_t
Forwarding gain of relay UAV	R_{g1}
Antenna gain of relay UAV	R_{g2}
Filter function of relay UAV	F_1
Transmitting power of fixed jammer	P_{j1}
Spectrum sequence of fixed jammer	S_{j1}
Antenna gain of fixed jammer	J_{g1}
Transmitting power of mobile jammer	P_{j2}
Spectrum sequence of mobile jammer	S_{j2}
Antenna gain of mobile jammer	J_{g2}
Filter function of receiver UAV swarm	F_2
Communication link 1 path loss	L_t
Communication link 2 path loss	L_r
Jamming link 1 path loss	L_{j1}
Jamming link 2 path loss	L_{j2}

The instantaneous spectrum of jamming signals received by a receiver UAV can be expressed by:

$$P_{rj} = S_{j1} P_{j1} J_{g1} L_{j1} F_1 R_{g1} R_{g2} L_r + S_{j2} P_{j2} J_{g2} L_{j2} \tag{3}$$

Formula for calculating SIR in passband of a receiver UAV can be expressed by:

$$\mathrm{SIR} = 10 \log 10(F_2 P_{rt} / P_{rj}) \tag{4}$$

The sum of SIR of the receiver UAVs will be used as reward in reinforcement learning, which reflects the overall level of communication quality in receiving area. In order to show the signal transmission process more clearly, we give the spectrums of the relay UAV and a receiver UAV at a certain time. In Figure 4, the spectrum transfer diagrams are plotted, given the carrier frequency range of communication link 1 is 2.35-2.4 GHz and that of communication link 2 is 2.4-2.45 GHz. Figure 4 (a) is the spectrum received by the relay UAV, and Figure 4 (b) is the spectrum received by the receiver UAV.

State and action are important parameters in reinforcement learning. Usually an agent takes an action and reaches a certain state, so that the action is usually a process quantity. In this problem, the cost of changing parameter in communication systems is largely unrelated to process volume (e.g., frequency). Therefore, the combination of jamming parameters at each time is regarded as a state, and the combination of communication parameters at each time is regarded as an action.

In this problem, the decision-making process of communication system has obvious Markov property, that is, no after-effect. Therefore, MDP can be used to construct system decision-making model to assist agents in making optimal decisions.

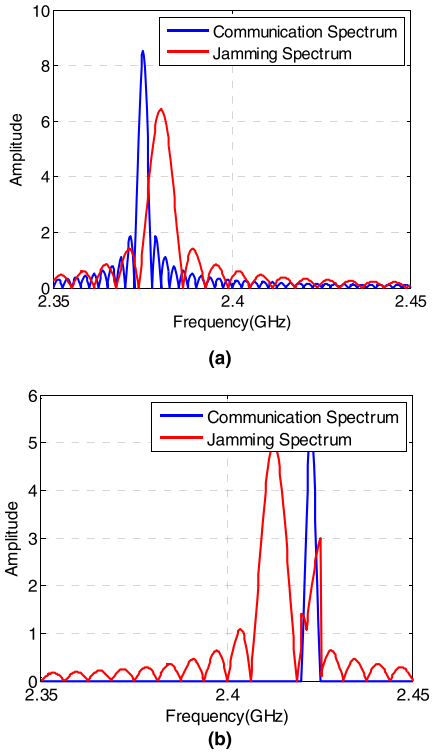


FIGURE 4. Spectrum transfer diagram: (a) spectrum received by relay UAV; (b) spectrum received by a receiver UAV.

In the decision-making process of an agent, a standard Markov decision-making process can be described by a quaternion shown in (5).

$$M = (S, A, P_{sa}, R) \tag{5}$$

S : represents a set of states

A : represents a set of actions.

P_{sa} : represents the state transition probability matrix of the system.

R : In the decision-making process, there is $S \times A \rightarrow R$. In this case, R is used to represent the reward function.

B. MULTI-DIMENSIONAL ANTI-JAMMING REINFORCEMENT LEARNING ALGORITHM

This paper considers Q-learning algorithm as the basic framework to design the algorithm. The features of Q-Learning are as follows: Firstly, Q-Learning is to solve the optimal Q value by directly solving the bellman optimal equation, rather than to choose the Q value of the optimal policy among the infinite policies π. Secondly, by incremental improvement, Q-Learning can achieve policy improvement. Thus, learning from any state can be realized, which can converge to the optimal value function.

In our UAV communication system, the combination of jammer parameters constitutes the environment. The agent here refers to the communication system. S represents the overall state of the communication system. A represents the combined action of the communication system in the three dimensions of "frequency-position-antenna". P_{sa} matrix is a full-1 matrix. R value represents the reward of the algorithm after taking an action in a certain state. The communication policy and jamming policy of all combinations are used to form the R table.

R-Table Computing Module: For each combination, the SIR is calculated according to (4) and then recorded in the R table. Usually, R table is a combination of different states and actions. In this scenario, each time step (or the corresponding jamming policy) is regarded as a state, and the communication policy to be adopted at the current moment is regarded as an action to construct the R table.

Policy Iteration Module: After the R table is generated and energy cost added, the basic Q-Learning algorithm is used to learn the jamming scenarios within 10 steps. The iterative refresh formula is as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \lambda \max_a Q(s_{t+1}, a_t) - Q(s_t, a_t)] \tag{6}$$

Among them, α is the learning rate and λ is the discount factor, which is used to describe the expected sum of multi-step rewards. The policy generated by the algorithm at each time step is regarded as the current action, and R table is accessed according to the combination value of (s, a).

Cost Constraint Module: This module is included in Policy Iteration Module. Because the current R table does not consider the motion range of the relay UAV in the two consecutive time steps, so the relay UAV may move in a wide range. For example, in order to avoid jamming better, the relay UAV may move from the leftmost to the rightmost in the adjacent time. However, this is not realistic as the motion speed and energy carried by UAV is limited. Therefore, we introduce an energy cost constraint module, whose main function is to measure the cost allocation problem in multi-dimensional parameter programming. Its core idea is to use the following cost factor c_{total} to restrict reward when Q-Learning accesses R value.

Specifically, the modified Q-Learning algorithm calculates the reward multiplied by the current cost factor c_{total} at each step of each iteration. Parameters are specified as follows: c_1 represents the transmitting frequency cost of ground station. c_2 represents the transmitting frequency cost of relay UAV. c_3 represents beam pointing cost of relay UAV. c_4 represents mainlobe cost of relay UAV. c_5 represents motion cost of relay UAV. c_{total} is determined by the following equation:

$$c_{total} = 0.025c_1 + 0.025c_2 + 0.025c_3 + 0.025c_4 + 0.9c_5 \tag{7}$$

For c_1, c_2, c_3 and c_4, they are 0.95 if the current action changes from the previous time step. They are 1 if the current action remains unchanged. For c_5, it is 0.95 if the motion range of relay UAV is less than 3. It is 0.01 if the motion range of relay UAV is not less than 3. In practice, the values of the cost factors may be tuned according to the specific UAV platforms. In this paper, the cost of relay UAV motion is set to

TABLE 3. The Pseudocode of MDAJRL algorithm.

```
Build transmission policy set p_t and jamming policy set p_j
For i = 1,2,...,M :
    For j = 1,2,...,N :
        r(i,j) = R-Table_Computing_Module[p_t(i), p_j(j)]
    end
end
Initialize Q(s,a) arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        Choose a from s using policy derived from Q
        Call the action generated by the last iteration a_1
        c_total = Cost_Constraint_Module(a, a_1)
        Take action a, observe r, s'
        Q(s,a) ←
        Q(s,a) + α[r(s',a') * c_total + γ max_a' Q(s',a') − Q(s,a)]
        s ← s';
        until s is terminal
    end
end
```

90%, and the other four parameters are set to 2.5%. These five parameters are only for simulation experiment. In the actual situation, these need to be measured based on the real agent, and the proportion is consistent since the energy consumption of UAV motion far exceeds that of communication circuit reconfiguration.

In summary, the design overview of the MDAJRL algorithm is given in Figure 5. First, the algorithm calculates the R value according to the underlying environment of the communication system, and stores the R value in the R table. Second, the algorithm iterates policy according to R table, in which cost constraint module is included. Finally, the algorithm generates the policy according to the final Q table.

The pseudocode of the algorithm is as in Table 3:

V. SIMULATION RESULTS

We give the range of optional parameters for each agent in the experimental group first. In order to describe the topology between agents in the experiment more clearly, we give the schematic diagram of the system in Figure 6. On the basis of grid modelling in Figure 2, the position and beam pointing of each agent are depicted.

The range of carrier frequency of ground base station is from 2.35 to 2.39 GHz. The moving area of relay UAV is discretized into 9 coordinates. The range of beam pointing of relay UAV is from 48° to 132°. The range of carrier frequency of relay UAV is from 2.41 to 2.45 GHz. The range of mainlobe of relay UAV is from 40° to 60°.

A. EXPERIMENTAL GROUP 1 (EFFECTS OF SWARM TOPOLOGY)

The purpose of this experiment is to give the results under symmetrical and asymmetrical topology of receiver UAV swarm. When the topology of receiver UAV swarm is symmetrical in topology (group A1), the beam of relay UAV points to the geometric center of the UAV swarm by default. Because there are fewer combinations of relay UAV policy at this time, the algorithm will generate policies in a relatively short time. When the topology of receiver UAV swarm is asymmetric (group A2), the beam pointing of relay UAV will be determined by the policy generated from the algorithm. More computational time is needed to generate the correct policy.

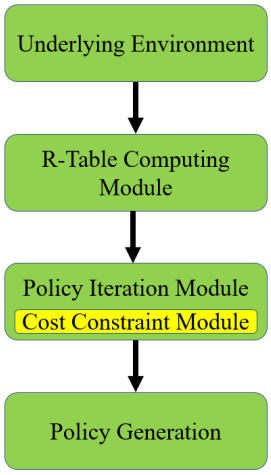


FIGURE 5. Structure diagram of MDAJRL algorithm.

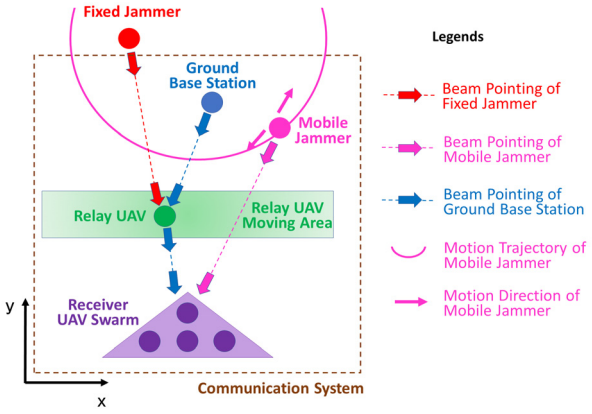


FIGURE 6. Schematic diagram of the system.

1) SYMMETRICAL RECEIVER UAV SWARM

In this experiment, a symmetrical receiving region is constructed, which is shown in Figure 7. The beam pointing of the relay is always the geometric center of the receiving area. In this way, the beam pointing of the relay UAV is related to the location of relay UAV, thus greatly reducing the iteration time required by the algorithm.

The time sequence diagram is Figure 8. The dots in the figure represent the positions of each agent. Lines represent the 3dB width range of the agent antenna. Frequency chart,

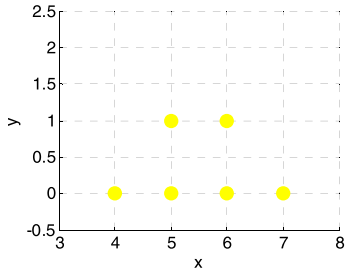


FIGURE 7. The distribution of receiver UAV swarm in group A1.

cost curve chart and receiver SIRs are shown in Figure 9, where we also generate random policy and default policy to compare with the experimental one.

Random policy refers to the selection of 10 groups randomly from the combination of communication policies with SIR greater than -30 dB. The default policy refers to the combination of policies adopted by the ground base station and the relay UAV without knowing any prior information of the receiver UAV swarm.

In this experiment, the default policies are set as follows: carrier frequency of ground base maintains at 2.37 GHz. Location of relay UAV is fixed and its beam swings back and forth within the range from 63° to 117°(azimuth). The carrier frequency of relay UAV maintains at 2.43 GHz. Mainlobe of relay UAV maintains at 50°.

The time sequence diagram in Figure 8 displays the mainlobe policy and motion policy of each agent. It can be seen that the trend of relay UAV motion is affected by the change of the beam scanning of the fixed jammer. It shows the trend of ''driving'' by the beam of fixed jammer, moving from the left to the right. For antenna policy of relay UAV, the accuracy of beam pointing plays a decisive role in the overall SIR level of the receiver UAV swarm, and the energy efficiency depends on whether the mainlobe width is well attached to the receiver UAV swarm. Because of the energy conservation of antenna pattern, the radiation of mainlobe will be weakened when wider mainlobe is used. If the mainlobe is too narrow, it cannot cover all receiver UAVs. From Figure 8, the antenna of relay UAV may fit the receiving area by selecting a proper mainlobe width, where the effectiveness of UAV antenna programming is illustrated.

The frequency chart shown in Figure 9 (a) proves the effectiveness of the frequency policy. It can be seen that the communication frequency and the jamming frequency keep a good isolation at each time. From the 7th-8th frames of the sequence diagram, it can be seen that the correct policy cannot be generated by relay UAV motion at these two moments, but the jamming can still be avoided by frequency selection. It shows that in the multi-dimensional programming model and algorithm established in this paper, each agent takes advantage of the degree of freedom.

Although Figure 9 (a) shows that the policies in the frequency domain can avoid the jamming frequency well, according to the spectrum modeling shown in Figure 4, the agent cannot completely avoid the jamming spectrum

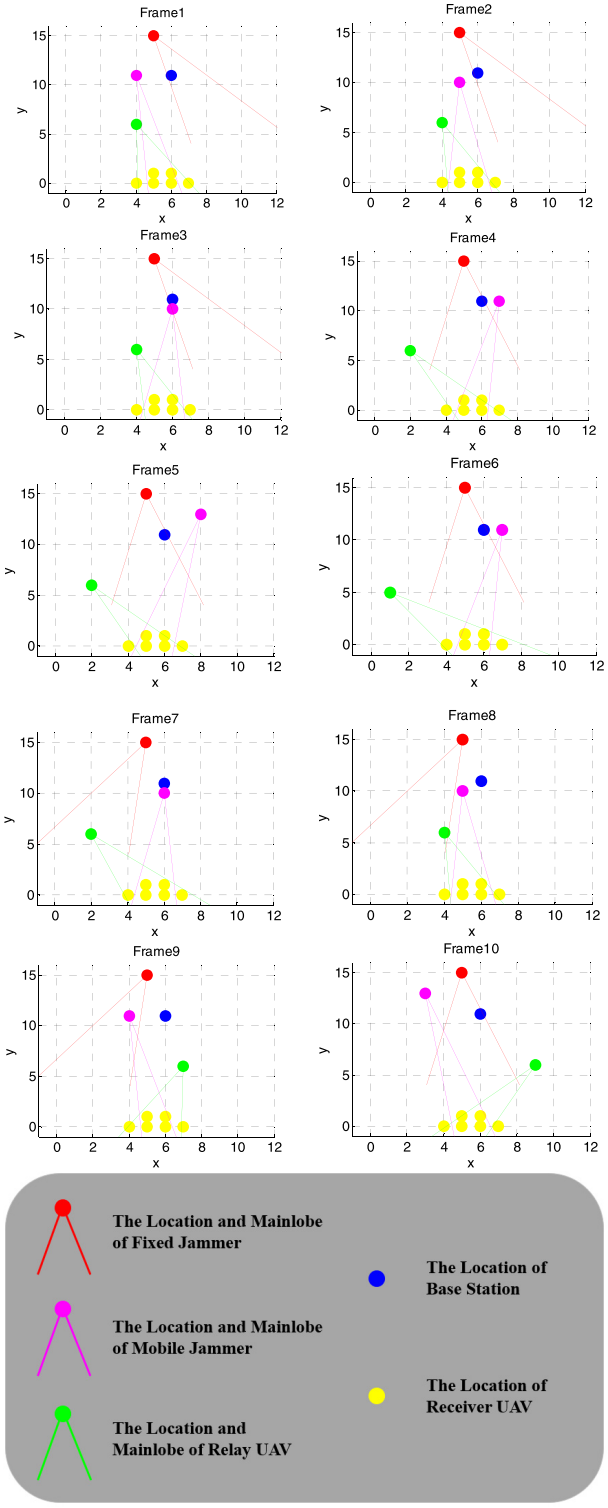


FIGURE 8. The time sequence diagram of experimental group A1.

in the frequency domain. This is because in the real scene, the signal is usually finite in time domain. Based on the theory of signal processing and linear system, the signal must be infinite in frequency domain. Thus, the spectrum components may also be outside the passband. After introducing the motion dimension, the agent may achieve a higher SIR by

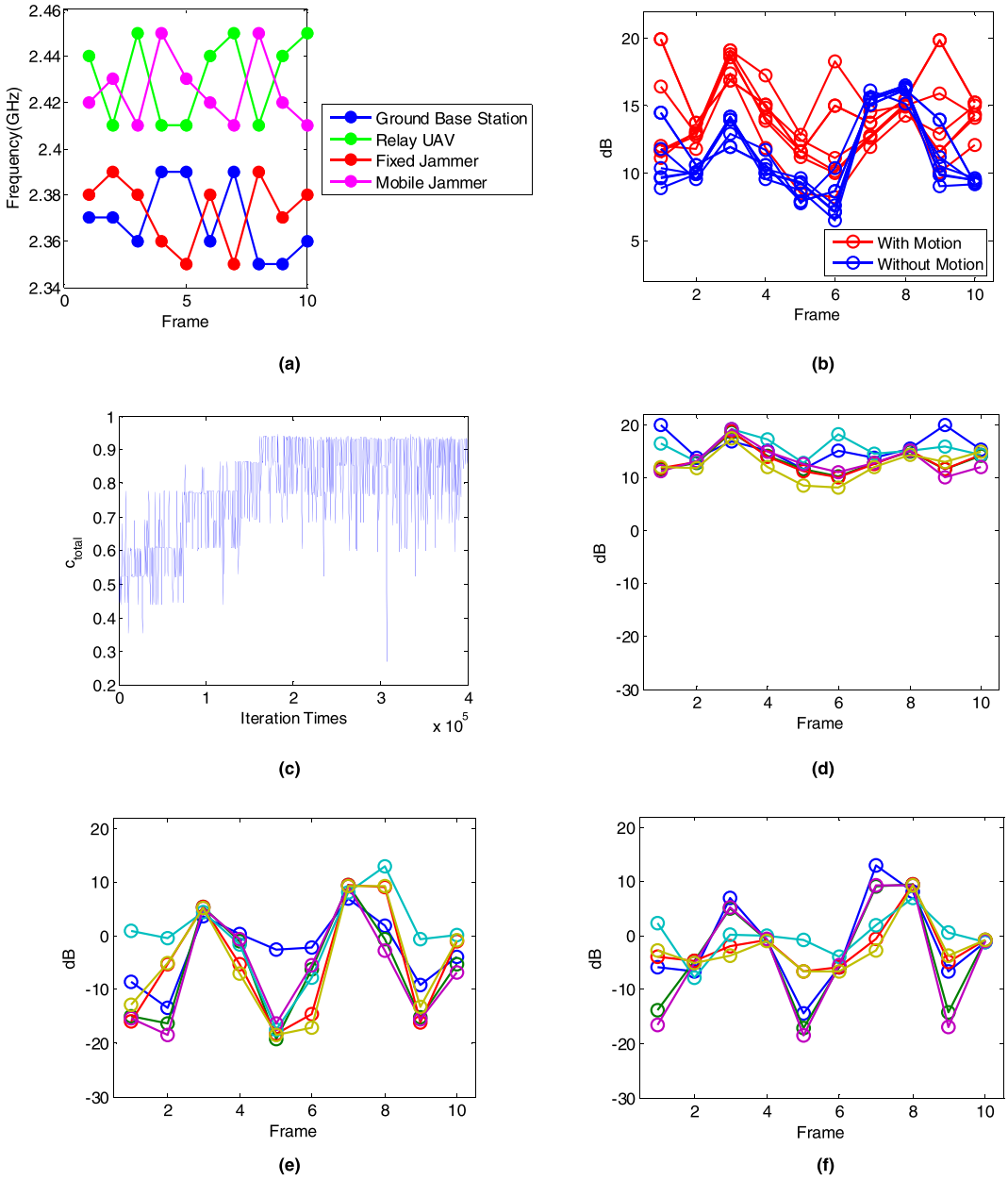


FIGURE 9. Simulation results of group A1: (a)Frequency chart (b) Receiver SIR contrast chart (policy with/without relay UAV motion) (c) Cost curve (d) Receiver SIR contrast chart (policy of MDAJRL) (e) Receiver SIR contrast chart (random Policy) (f) Receiver SIR contrast chart (default Policy).

selecting the location where the antenna radiation of jammer is weak. This result is shown in Figure 9 (b), which suggests the UAV motion may contribute to a SIR improvement of 4-5 dB.

Cost curve shown in Figure 9 (c) proves the effectiveness of the designed algorithm, and also reflects the minimum number of iterations required for convergence. It can be seen that in the initial stage of iteration, the value of c_{total} in the initial stage is small, because the UAV policy is uncertain and the energy consumption is large when it adopts large moving pace. This cost is computed into the Q value table by R value access link in MDAJRL algorithm. In the process of continuous learning, agents grasp the influence of spatial scale on relay motion, and gradually restrain their own motion until the balance between motion cost and SIR is achieved. As the number of iterations increases, the value of cost factor c_{total} becomes higher, and converges when it reaches about 0.95. There are six curves in Figure 9 (d) (e) (f) respectively, which represent the trend of SIR of six UAV receivers in the swarm over time. Compared with the random policy and default policy shown in Figure 9 (e) and (f), the receiver SIR contrast chart shown in Figure 9 (d) suggests that the current communication policy improves the overall SIR level in the receiver UAV swarm zone.

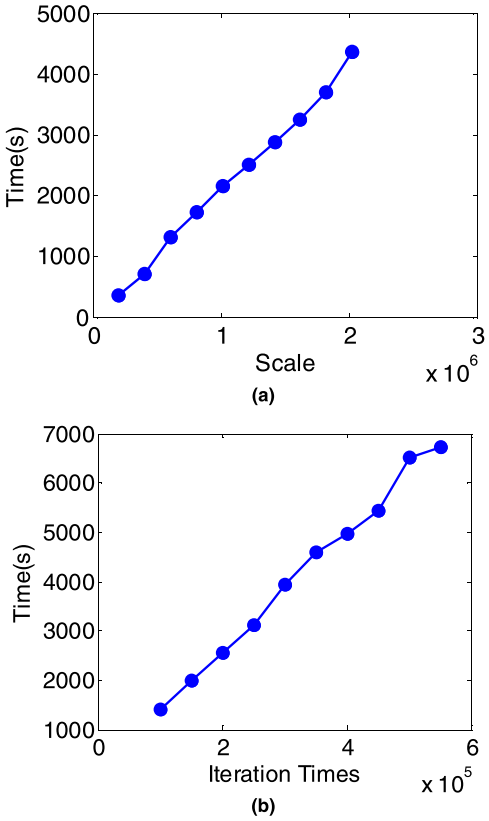


FIGURE 10. The calculation time of different scales and iteration times.

In order to quantify the complexity of R table calculation and policy iteration, we give the computational time required by different scales and iterations times under the current experimental group conditions. R table calculation is realized by MATLAB (version of 2014a), and policy iteration is realized by Pycharm (version of 2018.3.5). We use Intel (R) core (TM) i5-8300h CPU @ 2.30ghz processor, and the operating system is Windows 10. The results are shown in Figure 10. The scale in Figure 10 (a) refers to the number of policy combinations jointly determined by transmission policies and jamming policies, which reflects the problem complexity. The results show that the computational time grows linearly with the problem scale and number of iterations.

2) ASYMMETRICAL RECEIVER UAV SWARM

Compare with experimental group A1, the area of receiver UAV swarm is set to irregular arrangement, as shown in Figure 11. In this experiment, the beam pointing of the relay UAV is no longer determined solely by the location of relay UAV, but is generated according to the MDAJRL algorithm. Other setting remains unchanged. Thus, the number of communication policy combinations is greatly increased. The algorithm needs longer operation time. For simplicity, we no longer give time sequence diagram here.

The result of experimental group A2 is shown in Figure 12. The frequency chart is given in Figure 12 (a), which proves the effectiveness of frequency programming.

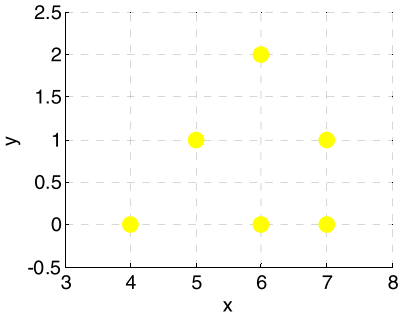


FIGURE 11. The distribution of receiver UAV swarm in group A2.

In Figure 12 (b), the convergence of the algorithm is proved by the cost curve. Since the number of communication policy combinations has changed from 1125 to 10125, the number of iterations has greatly increased. The number of iterations required for algorithm convergence is about 9 times as much as that in experimental group A1, and the running time has also increased to 9 times as much as that in experimental group A1. It consumes more computing time and requires higher computational power. In practical use, the programming method can be flexibly selected according to the actual situation of the task.

The setting of default policy is the same as that in experimental group A1. Compared with the random policy and default policy shown in Figure 12 (d) and (e), the SIR at the receiver UAVs are shown in Figure 12 (c) proves that MDAJRL algorithm can still generate correct policy of relay UAV when the programming of relay UAV beam pointing is independent.

B. EXPERIMENTAL GROUP 2 (EFFECTS OF ANTENNA PATTERN)

In this group, we set the antenna of relay UAV to omni-directional to examine the effects of antenna pattern. The conditions are the same as those in experimental group A1. The result of experimental group B is shown in Figure 13. Compared with Figure 9 (d), the SIR of the receiver in this group decreases by 6-8 dB on average, due to the waste of energy dispersed on non-intentional directions. From the results, it can be seen that a well-tuned directional antenna used in the relay UAV may concentrate energy and improve the communication efficiency.

C. EXPERIMENTAL GROUP 3 (EFFECTS OF MOTION COST)

In this experimental group, we remove the cost constraint in MDAJRL algorithm to illustrate the effectiveness of motion constraints on relay UAV. The other conditions are the same as those in experimental group A1.

The cost curves of experimental group A1 and C are shown in Figure 14. Compared with the result of A1, it can be seen that without cost constraints, the algorithm can make the relay UAV move in a wide range. Thus, the cost curve cannot be as stable as that in group A1, which is almost above 0.9 after 1.6×10^5 iteration times.

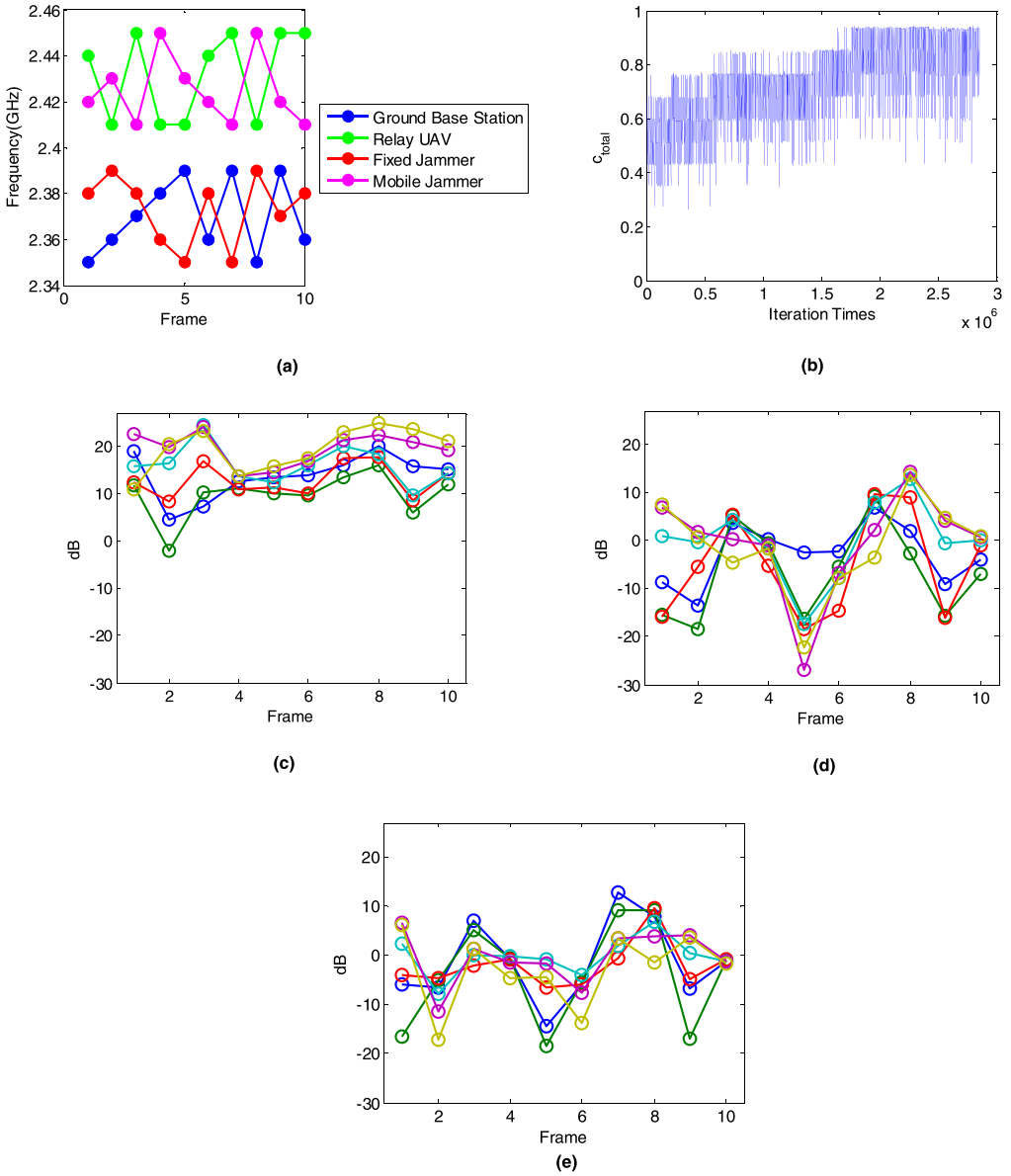


FIGURE 12. Simulation results of experimental group A2: (a)Frequency chart (b) Cost curve (c) Receiver SIR contrast chart (policy of MDAJRL) (d) Receiver SIR contrast chart (random policy) (e) Receiver SIR contrast chart (default policy).

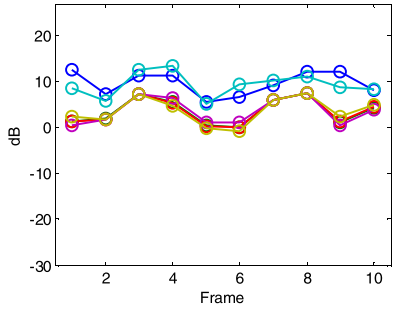


FIGURE 13. SIR at each receiver UAV using omnidirectional antennas.

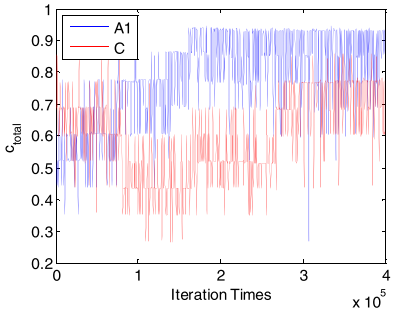


FIGURE 14. Cost curves for MDAJRL with motion costs (Group A1) and without motion costs (Group C).

In order to show the motion of the relay UAV under the condition of no cost constraint, we show the moving range of each time frame (compared with last frame) in Figure 15.

We give the coordinates of relay UAV in 10 frames in Table 4. It can be seen that at frame 4 and 10, the relay UAV adopts

TABLE 4. Trajectory of relay UAV.

Frame	Coordinate $[x, y]$
1	[4,6]
2	[4,6]
3	[4,6]
4	[9,6]
5	[7,6]
6	[9,6]
7	[8,5]
8	[8,5]
9	[7,6]
10	[1,5]

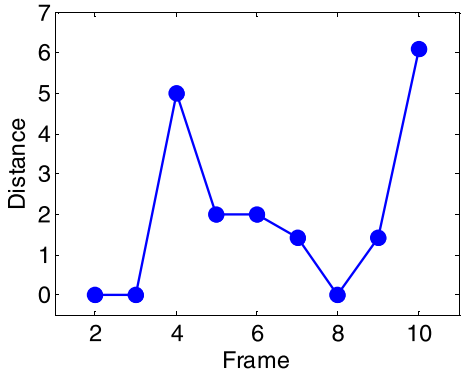


FIGURE 15. Moving range of relay UAV in each time frame.

the policy of large-range motion rather than the incremental motion in group A1 or A2.

VI. CONCLUSION

This paper focuses on the multi-dimensional programming of complex UAV communication networking. A framework of UAV intelligent communication system based on "frequency-motion-antenna" is established, and various parameters are set for each agent to programme. We regionalize the receiver UAV swarm to reflect the average SIR level of a particular shape area. Based on the limited energy of UAV, the reward is constrained by 3 dimensions, thus the existing Q-Learning algorithm is improved. The simulation results show that the policy generated by the MDAJRL algorithm can make the antenna of relay UAV fit well on the two parameters of beam pointing and mainlobe. The SIR of the receiving area can also be improved by the "frequency-motion-antenna" joint programming. It provides a new idea for UAV anti-jamming communication in complex scenarios.

REFERENCES

[1] J. Kim, C. Park, J. Ahn, Y. Ko, J. Park, and J. C. Gallagher, "Real-time UAV sound detection and analysis system," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Mar. 2017, pp. 1–5.

[2] H. Kim, J. Ben-Othman, L. Mokdad, S. Cho, and P. Bellavista, "On collision-free reinforced barriers for multi domain IoT with heterogeneous UAVs," in *Proc. IEEE 8th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2018, pp. 466–471.

[3] T. Ahmed, D. Feil-Seifer, T. Jiang, S. Jose, S. Liu, and S. Louis, "Development of a swarm UAV simulator integrating realistic motion control models for disaster operations," in *Proc. ASME Dyn. Syst. Controls Conf. (DSCC)*, 2017, pp. 1–10.

[4] M. R. Brust and B. M. Strimbu, "A networked swarm model for UAV deployment in the assessment of forest environments," in *Proc. IEEE 10th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process.*, Apr. 2016, pp. 1–6.

[5] C. Sampedro, H. Bavle, J. L. Sanchez-Lopez, R. A. S. Fernández, A. Rodríguez-Ramos, M. Molina, and P. Campoy, "A flexible and dynamic mission planning architecture for UAV swarm coordination," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, 2016, pp. 355–363.

[6] H. Sallouha, M. M. Azari, and S. Pollin, "Energy-constrained UAV trajectory design for ground node localization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[7] W. Bentz, T. Hoang, E. Bayasgalan, and D. Panagou, "Complete 3-D dynamic coverage in energy-constrained multi-UAV sensor networks," *Auto. Robots*, vol. 42, no. 4, pp. 825–851, 2018.

[8] G. Varela, P. Caamaño, F. Orjales, Á. Deibe, F. López-Peña, and R. J. Duro, "Swarm intelligence based approach for real time UAV team coordination in search operations," in *Proc. 3rd World Congr. Nature Biologically Inspired Comput.*, 2011, pp. 365–370.

[9] L. Weng, Q. Liu, M. Xia, and Y. D. Song, "Immune network-based swarm intelligence and its application to unmanned aerial vehicle (UAV) swarm coordination," *Neurocomputing*, vol. 125, pp. 134–141, Feb. 2014.

[10] A. G. Holubnychyi and G. F. Konakhovych, "Spread-spectrum control channels for UAV based on the generalized binary Barker sequences," in *Proc. IEEE 2nd Int. Conf. Actual Problems Unmanned Air Vehicles Develop. (APUAVD)*, Oct. 2013, pp. 99–103.

[11] E. Venosa, B. Vermeire, C. Alakija, F. Harris, D. Strobel, C. J. Sheehe, and A. Krunz, "Non-maximally decimated filter banks enable adaptive frequency hopping for unmanned aircraft vehicles," in *Proc. Integr. Commun. Navigat. Surveill. (ICNS)*, 2016, pp. 8E1-1–8E1-12.

[12] Y. Zhang, B. Zhang, and X. Yi, "Adaptive data sharing algorithm for aerial swarm coordination in heterogeneous network environments?" in *Proc. Int. Conf. Collaborative Comput., Netw., Appl. Worksharing*, 2018, pp. 202–210.

[13] B. Zhang, Y. Wu, X. Yi, and X. Yang, "Joint communication-motion planning in wireless-connected robotic networks: Overview and design guidelines," 2015, *arXiv:1511.02299*. [Online]. Available: https://arxiv.org/abs/1511.02299

[14] W. Zhang, W. Ding, and C. Liu, "A channel selection and power control method of UAV data link," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 43, no. 3, pp. 583–591, 2017.

[15] X. Wang, M. Lei, M. Zhao, and M. Li, "Cooperative anti-jamming strategy and outage probability optimization for multi-hop ad-hoc networks," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[16] A. Li and W. Zhang, "Mobile jammer-aided secure UAV communications via trajectory design and power control," *China Commun.*, vol. 15, no. 8, pp. 151–161, 2018.

[17] J. Zhang, T. Chen, S. Zhong, J. Wang, W. Zhang, X. Zuo, R. G. Maunder, and L. Hanzo, "Aeronautical *Ad Hoc* networking for the Internet-above-the-clouds," *Proc. IEEE*, vol. 107, no. 5, pp. 868–911, May 2019.

[18] J. Zhang, S. Chen, R. G. Maunder, R. Zhang, and L. Hanzo, "Adaptive coding and modulation for large-scale antenna array-based aeronautical communications in the presence of co-channel interference," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1343–1357, Feb. 2018.

[19] X. Liu, Y. Xu, L. Jia, Q. Wu, and A. Anpalagan, "Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 998–1001, May 2018.

[20] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.

[21] X. Liu, Y. Xu, Y. Cheng, Y. Li, L. Zhao, and X. Zhang, "A heterogeneous information fusion deep reinforcement learning for intelligent frequency selection of HF communication," *China Commun.*, vol. 15, no. 9, pp. 73–84, 2018.

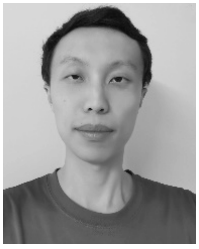
[22] F. Yao and L. Jia, "A collaborative multi-agent reinforcement learning anti-jamming algorithm in wireless networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1024–1027, Aug. 2019.

[23] F. Slimeni, B. Scheers, Z. Chtourou, and V. Le Nir, "Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, 2015, pp. 1–7.

[24] F. Slimeni, Z. Chtourou, B. Scheers, V. Le Nir, and R. Attia, "Cooperative Q-learning based channel selection for cognitive radio networks," *Wireless Netw.*, vol. 25, pp. 4161–4171, Oct. 2018.
[25] F. Slimeni, B. Scheers, Z. Chtourou, V. Le Nir, and R. Attia, "Cognitive radio jamming mitigation using Markov decision process and reinforcement learning," *Procedia Comput. Sci.*, vol. 73, pp. 199–208, Jan. 2015.
[26] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2087–2091.

JINLIN PENG received the B.S. degree in electromechanical engineering from the Beijing Institute of Technology, Beijing, China, and the Ph.D. degree in electrical engineering from the University of Leeds, U.K. He was a Postdoctoral Researcher with the Department of Electronic Engineering, Tsinghua University, Beijing. His main research interests include wireless network protocols, ad hoc networks, signal processing in wireless communications, and machine learning.

ZIXUAN ZHANG received the B.S. degree from the College of Electronic Science, in 2016, and the M.S. degree from the College of Computer, National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is currently pursuing the Ph.D. degree. His current research interests include MIMO, satellite communication, and joint motion programming for UAV communication.



QINHAO WU received the B.S. and M.S. degrees from the College of Electronic Science, National University of Defense Technology (NUDT), Changsha, China, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include radar coincidence imaging, metasurface antenna, and intelligent radar networking.

BO ZHANG received the Ph.D. degree from the Southampton Wireless Group, University of Southampton, U.K., in 2015. He is currently an Associate Professor with the Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, China. He has published more than 40 journal articles and conference papers. He is a principal investigator for three national foundation granted projects. His research interests include cognitive communication and networking for robot systems.