IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024EAL2075

Publicized:2024/11/25

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Integrated Path Planning, Spectrum and Power Allocation for Multi-UAV via Deep Reinforcement Learning*

Hongbin ZHANG†, Ao ZHAN†a), Jing HAN††, Chengyu WU†, *Nonmembers*, *and* Zhengqiang WANG†††, *Member*

SUMMARY The application of deep reinforcement learning(DRL) has become a hot research topic in unmanned aerial vehicle (UAV) path planning and resource allocation. However, current DRL methods do not consider coordination among spectrum, path and power, leading to a waste of spectrum resources. A coordinated routing and resource allocation Q network(CRRQN) algorithm with low computing complexity in multiple UAVs scenarios is proposed, and a co-optimization module is proposed to enhance coordination among path planning, spectrum and power allocation in CRRQN by designing their reward functions. Moreover, double deep Q network(DDQN) is employed to guarantee its stability. The simulation shows that the CRRQN algorithm reduces the flight time by about 4% and improves the channel capacity by about 15% compared to the existing algorithms. The running time per test epoch of CRRQN reduces by about 35%.

key words: *Multi-UAVs, path planning, resource allocation, deep reinforcement learning, joint optimization*

1. Introduction

With the rapid development of wireless communication technology, multi-unmanned aerial vehicles (multi-UAVs) systems have shown great potential in both military and civil sectors[1]. Path planning and resource allocation are key factors to ensure effective communication and mission execution in these systems. Traditional methods have been widely used to solve these issues[2][3]. However, these methods typically depend on precise mathematical models and ideal assumptions, making them inefficient in large-scale or dynamically changing environments.

With the rise of Deep Reinforcement Learning (DRL) techniques, novel perspectives are proposed for solving these complex problems. The issue of deep reinforcement learning-based UAV-assisted communication trajectory design and resource allocation in a single UAV scenario is explored in [4]. Some DRL based works consider multi-UAV scenarios, optimizing paths and resource allocation, respectively. In [5], the communication capability of UAVs is enhanced by using a simulated annealing algorithm for path planning and then using a double Deep Q Network (DDQN) for resource allocation. In [6], a real-time dynamic solution is proposed that utilizes DRL for UAV path planning and task allocation respectively, to reduce latency and improve decision quality in resource-constrained environments. However, these methods are typically inefficient in large-scale or dynamically changing environments.

In [7], a DRL-based dynamic trajectory control algorithm is proposed to minimize user equipment (UE) energy consumption by optimizing user associations, resource allocation, and UAV trajectories. In [8], a DRL-based trajectory design and resource allocation algorithm is proposed to maximize the overall system utility of multi-UAV networks. None of these works consider spectrum allocation. However, spectrum allocation is crucial for maximizing wireless efficiency and minimizing interference.

This letter presents a novel coordinated routing and resource allocation Q-network (CRRQN) algorithm with low computing complexity in multi-UAVs scenarios. CRRQN employs DDQN to ensure stability by reducing overestimation bias in Q-value updates. We also propose a co-optimization module, and its joint optimization reward functions consider both path distance and spectrum interference, improving the coordination among path planning, spectrum and power allocation for multi-UAVs performing tasks. Simulation shows that the CRRQN reduces the flight time by about 4% and improves the channel capacity by about 15% compared with the separated optimization algorithm in some scenarios with many UAVs. The running time per test epoch of CRRQN reduces by about 35%.

2. System model

2.1 Flight model

The flight scenario of the UAVs is shown in Fig.1, where an $\{M * M\}$ square area is randomly distributed with obstacles such as buildings and trees. A base station is set up at the center of the area for perform the proposed training learning process. A number L of UAVs start from the same starting point and fly to their respective random destinations. These UAVs are required to fly in one of the eight predefined discrete flight directions while satisfying specific cornering constraints [9]. The UAVs perform their tasks in consecutive time slots denoted as t, $t \in \{1, 2, ..., T\}$, and the total time slot is denoted as T. Each UAV forms a communication link with three neighboring UAVs, and a total of U communication links.

†The author is with School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou, 310018, P.R.China

††The author is with Hangzhou Synway Information Engineering Co.,Ltd, Hangzhou, 310053, P.R.China

†††The author is with School of Communications and Information Engineering, Chongqing University of Posts and Telecommunication, Chongqing, 400065, P.R.China

*This work was supported by National Key Laboratory of Science and Technology on Space Micrwave, No. HTKJ2022KL504016.

a) E-mail: zhanao1983@zstu.edu.cn(Corresponding author)

Fig. 1 UAV flight scenario

2.2 Path loss model

To simulate the communication environment more realistically, we consider the effect of path loss links, and the relationship between received power and transmitted power can be expressed as follows

$$P_{R,u} = P_{T,u} - PL_u \tag{1}$$

where $P_{R,u}$ denotes received power, $P_{T,u}$ denotes transmit power and PL_u denotes path loss.

The communication link between UAVs is dominated by line-of-sight(LOS) propagation and determined by distance[10]. It can be expressed as

$$PL_u = 20\log_{10}\left(\frac{d}{d_0}\right) + 10 \cdot \beta \cdot \log_{10}(d) + X_0 \tag{2}$$

where β is the path loss exponent, the distance d represents the straight line propagation distance between two UAVs, the reference distance d_0 is used for normalization, and X_0 adjusts for link-specific loss.

2.3 Communication model

In this letter, we divide the total bandwidth W equally into L orthogonal sub-bands. These sub-bands are occupied by the communication links between L UAVs and the base station. To improve the efficiency of using limited spectrum resources, each inter-UAV communication link is allowed to reuse one uplink spectrum. Combined with the Shannon-Hartley theorem, the communication capacity of the u-th link at time t can be expressed as

$$C_t^u = W\log_2\left(1 + \frac{P_{T,u}g_{U,u}}{\sigma^2 + I_t^u}\right) \tag{3}$$

where W is the bandwidth, and σ^2 represents the noise power. The transmit power is denoted by $P_{T,u}$. The channel gain $g_{U,u}$ depends on factors such as distance and the propagation environment. The interference power I_t^u includes contributions from other links and environmental factors. It can be expressed as follows

$$I_t^u = \sum_{i \neq u} P_{T,i}\widetilde{g}_{U,u} \tag{4}$$

where $\widetilde{g}_{U,u}$ denotes the interference power gain of links sharing the same spectrum resource.

The optimization goal is to minimize the flight time and maximize the sum capacity of multi-UAVs.The optimization goal is limited by a series of constraints, including path planning for UAVs to effectively avoid obstacles while adhering to fixed flight altitude and speed, as well as maximizing system capacity through precise spectrum and power allocation, while also considering the impact of mutual interference between drones on communication links.We propose a reinforcement learning framework to address this optimization issue.

3. Reinforcement Learning Framework

This section first introduces the three fundamental elements of DRL, including state, action, and reward, and then describes the CRRQN framework.

3.1 RL elements

State The state is defined as a collection of parameters at time step t, which includes flight path position information and communication link state information. The state space vector can be expressed as

$$s_t = \{p_t^i, D_t^i, d^i, D_{obs,t}^i, I_t^u, C_t^u, Z_t^u\} \tag{5}$$

where p_t^i represents the position of UAV i at time t, d^i denotes the position of the destination corresponding to UAV i, $D_{obs,t}^i$ denotes the distance between UAV i and the nearest obstacle, I_t^u denotes the interference power, C_t^u denotes the channel capacity and Z_t^u represents the remaining transmission time.

Action The action space A defines the possible actions for agents. In this letter, these actions include the selection of discretized flight direction, spectrum, and power. The flight directions are divided into eight basic directions, and the agent selects the next action based on the current Q-value.

Spectrum selection involves the agent choosing the reused spectrum sub-band to send its data. In addition, the agent needs to select a discretized power level P that is discretized into four levels to balance the transmission efficiency and interference limitations, while maintaining policy diversity and adaptability.

Reward The reward functions include the efficiency of flight and communication, we consider the following. R_{path} denotes the reward of path planning, defined as

$$R_{path} = \sum_{i=1}^{L}\left(\frac{\kappa_1}{\|p_t^i - d^i\|} - \frac{\kappa_2}{D_{obs,t}^i}\right) \tag{6}$$

where $\|p_t^i - d^i\|$ denotes from UAV i to its destination, κ_1 and κ_2 is weighting parameters. The R_{path} goal is to minimize the distance of the UAV to reach the destination and to consider the distance to the obstacles when planning the path.

R_{power} is a reward for power allocation and can be defined as

$$R_{power} = \sum_{j=1}^{U}\left(\eta_1 \cdot C_t^j - \eta_2 \cdot I_t^j\right) \tag{7}$$

where η_1, η_2 are weighting parameters, and the power allocation reward goal is to maximize channel capacity and minimize interference.

R_{spec} represents the reward for spectrum allocation, only when the selected channel is idle and data is sent within the maximum allowable transmission delay Z_0^u, is it considered successful,

$$R_{spec} = -(Z_0^u - Z_t^u) \tag{8}$$

Considering high interference situations may have an impact on the safety of the UAV, we propose the interactive impact of path planning and interference power as R_{p_int}.

$$R_{p\_int} = -\sum_{i=1}^{L}\sum_{j=1}^{U}\left(\epsilon \cdot \|p_t^i - d^i\| \cdot I_t^j\right) \tag{9}$$

where ϵ is a weight parameter. The goal is to minimize the product of the distance from the UAV to its destination and the interference. In this way, the interference of the communication link is taken into account in path planning to avoid high interference paths.

Therefore, the joint reward function at time slot t is

$$R_t = \delta_1 \cdot R_{path} + \delta_2 \cdot R_{power} + \delta_3 \cdot R_{spec} + \delta_4 \cdot R_{p\_int} \tag{10}$$

where δ_1, δ_2 , δ_3 and δ_4 are weighting parameters.

3.2 CRRQN framework

In this letter, we propose a CRRQN algorithm for complex multi-UAV environments as shown in Fig. 2.

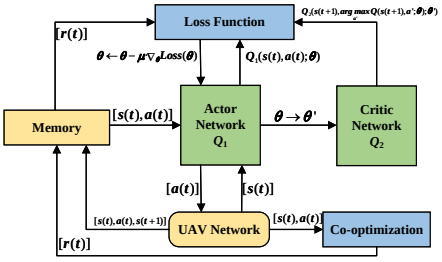

Fig. 2 CRRQN Framework

The UAV network interacts with the environment to obtain the state $s(t)$. Using the actor network Q_1, it selects the optimal action $a(t)$ for the current state to guide UAV operations. The obtained $s(t)$ and $a(t)$ are then input into the co-optimization module, which includes four specific reward functions. Among these, the joint optimization reward function correlates the path deviation distance of the target point with the interference level. This minimizes the product of the UAV's distance to the destination and the interference, thus obtaining the reward $r(t)$ and transitioning the system to a new state $s(t + 1)$.

The experiences of these interactions $[s(t), a(t), r(t), s(t+1)]$ are stored in a memory buffer, and the experiences accumulated in this buffer will be randomly sampled to form small batches of data for training. At each training step, a mini-batch of experiences $[s(t), a(t)]$ is randomly selected from memory. This small batch is crucial to breaking the correlation between successive experiences and stabilizing the learning process. Next, the actor network predicts the Q_1-value $Q(s(t), a(t); \theta)$ for each action in a given state, and the critic network evaluates these actions by computing the Q_2-values.The loss function is calculated based on the Q_1-value and the Q_2-value,

$$Loss(\theta) = \left(r(t) + \gamma Q\left(s(t+1), \max_{a'} Q'\left(s(t+1), a'(t); \theta\right); \theta'\right) - Q\left(s(t), a(t); \theta\right)\right)^2 \tag{11}$$

where γ is the discount factor. θ and θ' are the weight parameters of the actor network and critic network, respectively. The gradient is computed based on the loss and back-propagated through the actor network. The weights of the actor and critic networks are updated accordingly to minimize the loss. The updated actor network generates improved strategies to guide the UAV's movements. This network learns continuously to maximize the expected cumulative reward, leading to more efficient path planning and resource allocation over time.

4. Simulation

In this letter, a 20 × 20 grid simulation environment is constructed with each grid cell representing 0.25 km to simulate a 5 km × 5 km UAV operation space. Twenty obstacles are randomly arranged in the area, and each UAV departs from the map coordinates (1,1) with the destination randomly generated on the map. The UAVs fly at a fixed altitude and speed. We simulate using RTX 3090 and AMD R9 5950X, the training cycle is divided into 100 epochs, and each epoch has 2000 time steps.The detailed parameter settings are shown in Table 1.

Table 1 Parameters setting

Parameters	Value
L	{8, 16, 24, 32, 40}
X_0	5dB
β	3.5
$\delta_1, \delta_2, \delta_3, \delta_4$	0.2,0.2,0.2,0.4
η_1, η_2	1.0,0.5
ϵ	0.2
d_0	1m
$P_{T,u}$	{10,15,20,25}dBm
σ^2	-100dB
γ	0.6

4.1 Simulation results

This letter focuses on the average total capacity and average flight duration of different algorithms for different numbers of UAVs. This letter compares the proposed CRRQN with three baseline methods derived from reference [5]: The separated optimization algorithm performs sequential optimization, first applying simulated annealing for path planning and then using reinforcement learning for resource allocation, allowing for separated optimization of route and resource allocation tasks. The PSO-based method refines a candidate allocation strategy by assessing its quality and adjusting based on the best-known solutions from neighboring particles to converge efficiently. In the random allocation strategy, each agent randomly selects a communication subband and transmits power at each time step.

In Fig.3, the flight time is slightly longer than the separated optimization algorithm when the number of UAVs is small is because CRRQN requires more computational resources in the initial phase to coordinate and optimize the

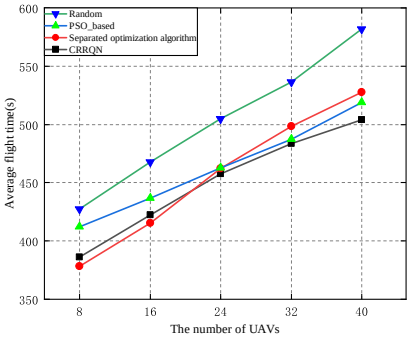

Fig. 3 Average flight time of different algorithms

global task allocation. As the number of UAVs increases, CRRQN can utilize resources more efficiently and reduce redundant flights, reducing the flight time by about 4%.

In Fig.4, the CRRQN also shows its superiority, especially when dealing with a large number of UAVs, it can allocate tasks more reasonably, improving the channel capacity by about 15%.

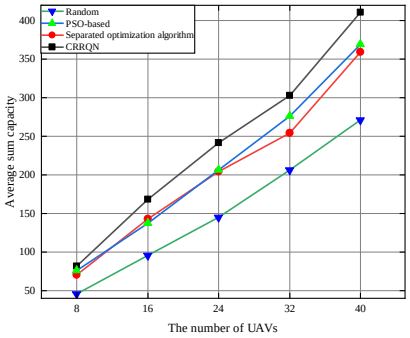

Fig. 4 Average sum capacity of different algorithms

To visualize the results of the path planning more, Fig.5 shows the flight trajectories of four UAVs guided by the CRRQN, where the black squares represent obstacles and the different colored lines indicate the path of each UAV. The interference power significantly affects the flight trajectories, this adjustment may cause the UAVs to choose not the shortest path, but the path that maximizes the overall reward according to the joint reward function, even if this requires bypassing obstacles or increasing the flight distance.

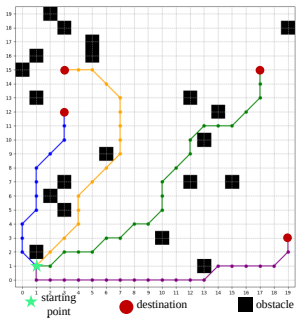

Fig. 5 UAV flight paths

When the number of UAVs is 32, the total training time takes about 16 hours, and the specific test time complexity for the four methods is shown in Table 2. The CRRQN simultaneously considers path planning, spectrum and power allocation, resulting in a shorter testing time. On the other hand, the separated optimization algorithm decomposes the problem into several sub-problems, increasing the complexity and thus slightly extending the testing time. PSO-based methods have a relatively long testing time due to evaluating and updating a large number of particles. Random algorithms have the shortest testing time, but the optimization results may be poor.

Table 2 The test time complexity

Algorithms	Test time
Random	0.047s
PSO_based	0.706s
Separated optimization algorithm	0.459s
CRRQN	0.296s

4.2 Parameter analysis and ablation experiment

To achieve optimal performance of the model, parameter settings are crucial, so we conduct detailed analyses on the parameters.The detailed analyses of η_1 and η_2 are shown in Fig.6 (a). By adjusting η_1 (0.5, 1.0, 1.5, 2.0) and η_2 (0.3, 0.5, 1.0), we observe that the combination of $\eta_1 = 1.0$ and $\eta_2 = 0.5$ provides the best balance. As shown in Fig.6 (b), for ε, we compare the effect of different parameter values on the system performance. Taken together when $\varepsilon = 0.2$, the results are optimal.

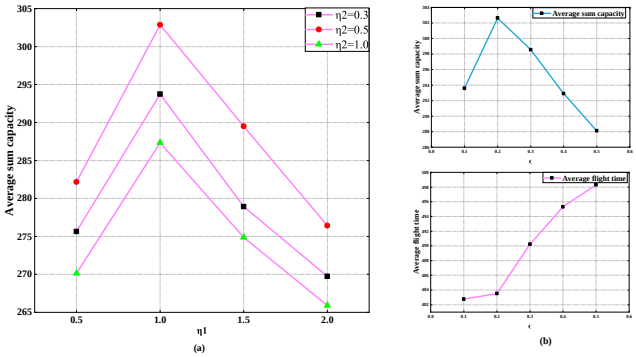

Fig. 6 η_1 η_2 and ε detailed analysis

To further investigate the effectiveness of each reward function and the effect of the weighting parameters on the performance, we perform detailed analysis by varying the weighting parameters δ_1, δ_2, δ_3 and δ_4. We conduct experiments by varying each parameter between [0,0.1,0.2,0.3,0.4,0.5,0.6], while evenly distributing the remaining three. The results, shown in Fig.7 (a) and Fig.7 (b), indicate that different parameter values significantly affect performance. However, overemphasizing any single parameter leads to a decline in the complementary metric.This analysis identifies $\delta_1 = 0.2$, $\delta_2 = 0.2$, $\delta_3 = 0.2$, $\delta_4 = 0.4$ as the optimal configuration, balancing flight time and channel capacity.

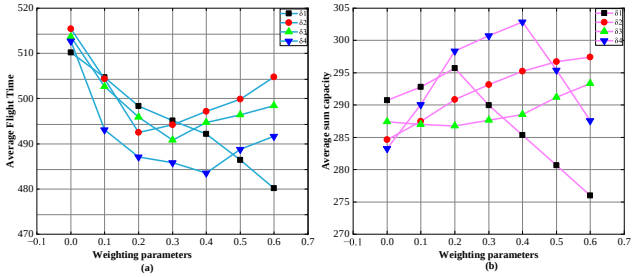

Fig. 7 Reward weighting parameters

To verify the validity of the four reward functions, we conduct an ablation experiment in the above experiment by setting the weighting parameters of each reward to zero. As shown in Fig. 7, removing any reward significantly leads to longer average flight time and lower average sum capacity. This confirms that each reward function plays a crucial role in improving flight efficiency and communication quality.

5. Conslusion

This letter proposes the CRRQN based on DDQN with low computing complexity. It improves the coordination of multi-UAVs through co-optimization modules, significantly reducing flight time and increasing channel capacity. Experimental results show that the CRRQN reduces the flight time by 4% and increases the channel capacity by 15% compared to the separated optimization algorithm.

References

[1] S. Javaid, N. Saeed, Z. Qadir, H. Fahim, B. He, H. Song, and M. Bilal, "Communication and control in collaborative uavs: Recent advances and future trends," IEEE Transactions on Intelligent Transportation Systems, vol.24, no.6, pp.5719–5739, 2023.

[2] P. Guo, W. Xu, Y. Zhu, Y. Chen, S. Zhang, and C. Wei, "Multi-uav collaborative path planning based on improved genetic algorithm," International Conference on Autonomous Unmanned Systems, pp.2648–2657, Springer, 2021.

[3] L. Huang, H. Qu, and L. Zuo, "Multi-type uavs cooperative task allocation under resource constraints," IEEE Access, vol.6, pp.17841–17850, 2018.

[4] C. Zhang, Z. Li, C. He, K. Wang, and C. Pan, "Deep reinforcement learning based trajectory design and resource allocation for uav-assisted communications," IEEE Communications Letters, vol.27, pp.2398 – 2402, 2023.

[5] C. Liu, L. Huang, and Z. Dong, "A two-stage approach of joint route planning and resource allocation for multiple uavs in unmanned logistics distribution," IEEE Access, vol.10, pp.113888–113901, 2022.

[6] M.A. Dhuheir, E. Baccour, A. Erbad, S.S. Al-Obaidi, and M. Hamdi, "Deep reinforcement learning for trajectory path planning and distributed inference in resource-constrained uav swarms," IEEE Internet of Things Journal, vol.10, no.9, pp.8185–8201, 2022.

[7] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and A. Nallanathan, "Deep reinforcement learning based dynamic trajectory control for uav-assisted mobile edge computing," IEEE Transactions on Mobile Computing, vol.21, no.10, pp.3536–3550, 2021.

[8] Z. Chang, H. Deng, L. You, G. Min, S. Garg, and G. Kaddoum, "Trajectory design and resource allocation for multi-uav networks: Deep reinforcement learning approaches," IEEE Transactions on Network Science and Engineering, vol.10, no.5, pp.2940–2951, 2022.

[9] Z. Cui and Y. Wang, "Uav path planning based on multi-layer reinforcement learning technique," IEEE Access, vol.9, pp.59486–59497, 2021.

[10] Q. Zhu, Y. Wang, K. Jiang, X. Chen, W. Zhong, and N. Ahmed, "3d non-stationary geometry-based multi-input multi-output channel model for uav-ground communication systems," IET Microwaves, Antennas & Propagation, vol.13, no.8, pp.1104–1112, 2019.