KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 9 (2023) 1128–1132

www.elsevier.com/locate/icte

DRL-based Multi-UAV trajectory optimization for ultra-dense small cells

Igbafe Orikumhi[a,*], Jungsook Bae[b], Hyunwoo Park[a], Sunwoo Kim[a,*]

[a] *Department Electronics and Computer Engineering, Hanyang University, Seoul, South Korea*
[b] *Electronics and Telecommunications Research Institute, Republic of Korea*

Received 26 October 2022; received in revised form 6 January 2023; accepted 20 May 2023
Available online 27 May 2023

Abstract

In this paper, we propose a deep reinforcement learning (DRL) based unmanned aerial vehicles (UAV)-assisted trajectory optimization for ultra-dense small cell networks. We assume that each UAV is equipped with a sensing radio to obtain distance information to the UEs and other UAVs in the network which are used to update the UAV's trajectory. The proposed DRL-based system selects the optimal joint control actions for the UAVs that maximizes the system sum-rate. The simulation results show that the proposed DRL-based UAV controller provides fast UAV placement in the network with a high system performance when compared with the benchmark schemes.
© 2023 The Authors. Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: 6G; Multiagent deep reinforcement learning (DRL); Multitarget tracking; Unmanned aerial vehicle

1. Introduction

Unmanned aerial vehicle (UAV) has found it usefulness in wireless communications in areas such as aerial base stations/relays, aerial helpers for providing new services such as data backhaul/offloading, sensor node data collection and cached-content multicasting etc. [1]. While UAV will be key to achieving the goals of future 6G networks, several issues must be addressed in order to exploit its potential benefits. Such issues include the computational load and low energy performance of UAVs, two-way communication with ground base stations, data offloading, real-time control, etc [2].

Although employing UAVs in 6G and beyond is a promising solution, effective UAV placement is required to guarantee their success [3,4]. In [5], K-means algorithm was employed for UAV placement in a distributed fashion. The method showed high computation load on the UAVs with slow convergence. A 3D trajectory planning and scheduling algorithm was proposed in [6] to minimize the pathloss. The study shows that while increasing the UAV's altitude may improve the probability of line-of-sight (LoS), it could lead to additional pathloss. Hence, there is a need for efficient and less computationally complex UAV 3D trajectory planning and placement algorithms. Furthermore, with the increase in the probability of the LoS path, algorithms for interference control to other UAVs and ground user may be required to achieve maximum network throughput [7].

Deep reinforcement learning (DRL) tools were also exploited in [6] for the joint trajectory optimization and frequency band allocation. Resource scheduling based on DRL for UAV-assisted communication was proposed in [8]. However, the authors in [6,8] fail to capture the dynamics of the ground users. In addition, the developed models capture only the LoS path, which is not realistic in ultra-dense networks.

In this paper, we propose a deep reinforcement learning (DRL) based multiple UAV placement and trajectory optimization. Note that when considering the joint association (UAV-UE) and UAV trajectory optimization in an ultra-dense small cell with mobile ground users, both discrete (association matrix) and continuous (3D positions) variables are involved. The problem is NP-hard, and hence, the existing results are mainly based on hybrid meta-heuristics solutions. To address, these issues, we employ UAVs with sensing capabilities, hence, the ground user location can be exploited in the UAV trajectory optimization.

2. System model and problem formulation

We consider a disaster scenario where the communicating base station is either non-functional or out of communication range. We assume M UAVs are deployed to serve K ground

* Corresponding authors.
E-mail addresses: oigbafe2@hanyang.ac.kr (I. Orikumhi), jsbae@etri.re.kr (J. Bae), stark95@hanyang.ac.kr (H. Park), remero@hanyang.ac.kr (S. Kim).
Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2023.05.007
2405-9595/© 2023 The Authors. Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

user equipment (UE). Furthermore, we assume that each UAV is equipped with two radios, where one of the radio is used for continuous 3D position sensing and the other radio is employed for data communication.

2.1. UAV dynamics and sensing

At timestep n, the state vector of the mth UAV is represented by $\mathbf{u}_m^n = [u_x^m, u_y^m, u_z^m]^T$ where x, y and z are the coordinate vectors of the UAVs. Let $\mathcal{S}$ be the set of ground users indices that need to be served by the UAVs. In the nth time step, the state vector of the mth UAV can be expressed as

$$\mathbf{u}_m^n = \mathbf{u}_{m-1}^n + b_l = \mathbf{u}_{m-1}^n + \begin{bmatrix} \delta\cos(l\Delta_\theta) \\ \delta\sin(l\Delta_\theta) \\ h_l \end{bmatrix}, \tag{1}$$

where δ is a constant step distance that the UAVs can travel at each timestep n, $\Delta_\theta = 2\pi/N_\theta$ is the unit steering angle, b_l, for $l = \{1, \ldots, N_\theta\}$ is the action control along the x, y, and z axis and h_l is the discrete altitude. The UAV determines trajectory by choosing its action from the discrete action space $\mathcal{B} \in \{b_1, b_2, \ldots, b_{N_\theta}\}$. The channel model includes LoS and NLoS with a probability that depends on both the UAV's altitude and the elevation angle between the user and the UAV. Given the mth UAV with an altitude h_m and the kth user with a distance $r_{k,m}$ from the UAV, the probability of LoS is given by [9]

$$p_{k,m}^{\mathrm{LoS}}(r_{k,m}, d_{k,m}) = \frac{1}{1 + \epsilon \cdot \exp\left(-\beta \frac{180}{\pi} \arctan \frac{\sqrt{r_{k,m}^2 - d_{k,m}^2}}{d_{k,m}} - \epsilon\right)}, \tag{2}$$

where $d_{k,m}$ is the horizontal distance from the projected position of the UAV on the 2D plane, ϵ and β are environment dependent parameters. Accordingly, the path loss between the mth UAV and the kth UE can be expressed as [10]

$$L_{k,m}(r_{k,m}, d_{k,m}) = \left(\frac{4\pi f r_{k,m}}{c}\right)^{-\alpha} (\zeta_{LoS} p_{k,m}^{LoS}(r_{k,m}, d_{k,m})) + \zeta_{NLoS}(1 - p_{k,m}^{LoS}(r_{k,m}, d_{k,m}))^{-1}, \tag{3}$$

where ζ_{LoS} and ζ_{NLoS} represent the losses from the LoS and NLoS links respectively.

2.2. Spectral efficiency

We assume that each UAV transmits with power P_{tx}. Hence, when a frame is transmitted by a mth UAV, it is received at the kth UE with the power $P_{tx} g_{k,m} L_{k,m}(r_{k,m} d_{k,m})$, where $g_{k,m}$ accounts for the multipath fading that is assumed to follow an exponential distribution. The quality of the wireless link is measured in terms of signal-to-interference-and-noise-ratio (SINR) defined as follows [11]

$$\gamma_{k,m} = \frac{P_{tx} g_{k,m} L_{k,m}(r_{k,m} d_{k,m})}{\sum_{j \neq m} P g_{k,j} L_{k,j}(r_{k,j} d_{k,j}) + \sigma^2}, \tag{4}$$

where σ^2 represents the power of an additive Gaussian noise. The average spectral efficiency received at a kth UE from the mth UAV ($C_{k,m}$) can be expressed as

$$C_{k,m} = \mathbb{E}[\log_2(1 + \gamma_{k,m})]. \tag{5}$$

In order to satisfy the kth user rate requirement R_k, the mth UAV needs to adjust the allocated bandwidth $b_{k,m}$ according to the quality of the link such that

$$R_{k,m} = b_{k,m} C_{k,m}. \tag{6}$$

2.3. Problem formulation

Let $\mathbf{A} \in \mathbb{R}^{K \times M} = (a_{k,m})$ denote the UAV-UE association matrix. The objective is to maximize the sum-rate by jointly optimizing the association matrix $\mathbf{A}$ and the 3D placement of the UAV $[\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z]$, subject to the bandwidth limitation for all UAVs and the constraint on the quality of service of all UEs. The problem can be formulated as

$$\max_{\mathbf{A}, (\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z)} \sum_{m \in \mathcal{U}} \sum_{k \in \mathcal{S}} a_{k,m} R_{k,m} \tag{7a}$$

$$\text{s.t} \quad R_{k,m} = b_{k,m} C_{k,m}, \quad \forall k \in \mathcal{S}, \forall m \in \mathcal{U} \tag{7b}$$

$$\sum_k a_{k,m} b_{k,m} \leq B_m, \quad \forall m \in \mathcal{U} \tag{7c}$$

$$a_{k,m} C_{k,m} \leq \frac{1}{C^{\min}} \quad \forall (k, m) \in \mathcal{U} \times \mathcal{S} \tag{7d}$$

$$\mathbf{x}^{\min} \leq \mathbf{u}_x^m \leq \mathbf{x}^{\max}, \quad \mathbf{y}^{\min} \leq \mathbf{u}_y^m \leq \mathbf{y}^{\max}, \quad \forall m \in \mathcal{U} \tag{7e}$$

$$\mathbf{u}_z^m \in \mathcal{H}, \quad \forall m \in \mathcal{U}, \tag{7f}$$

$$\mathbf{a}_{k,m} \in \{0, 1\}, \sum_m \mathbf{a}_{k,m} \leq 1, \quad \forall (k, m) \in \mathcal{U} \times \mathcal{S} \tag{7g}$$

where the constraint in (7b) guarantees that the minimum rates for each associated UAV-UE are satisfied, the constraint in (7c) ensures that the bandwidth on each UAV is not exceeded, (7d) ensures that the average spectral efficiency is not less that the minimum required rate ($R_{k,m}$) to each user, the constraint in (7e) and (7f) ensures that the service area of the UAVs is within the communication boundaries and the constraint in (7g) ensures that each user is only being served by a single UAV.

It should be emphasized that the minimum rate requirement for a UE must be satisfied for the UE to be associated to a UAV (in $\mathbf{A}$). Also, if the number of UEs (K) being served exceed the available bandwidth, user scheduling can be implemented.

3. Proposed method and solution

As shown in the problem formulation, the UAV placement and trajectory optimization is mathematically challenging. Finding a global optimal solution cannot be achieved using classical convex optimization methods. Our objective is to employ machine learning combined with the sensed information of the UAV's to design a low-complexity algorithm in order to achieve a stable solution. To this end, we introduce a deep reinforcement learning-based multiple UAV control algorithm for accurate UAV placement and trajectory optimization based on the UEs positions. First, we present details of the DRL algorithm, including the state, action, and reward. Then, we describe the UAV placement.

3.1. A DRL based UAV position placement

In the DRL-based UAV placement, the mth UAV observes state $\mathbf{s}_m^n$, moves its position $\mathbf{u}_m^n$ by taking action b_m^n, and gets a reward r_m^n through the UAV-environment interaction.

3.1.1. DRL state

The state vector of the mth agent at timestep n considers UAV position and UEs position, which is represented by $\mathbf{s}_m^n = [\mathbf{u}_m^n, \bar{\mathbf{u}}_{m1}^n, \ldots, \bar{\mathbf{u}}_{mM}^n, \mathbf{q}_{m1}^n, \ldots, \mathbf{q}_{mK}^n] \in \mathbb{R}^{3(M+K)}$, where $\mathbf{u}_m$ is the absolute coordinates of the mth UAV, $\bar{\mathbf{u}}_{mj}$ is the relative position between the mth UAV and the jth UAV and $\mathbf{q}_{mk}$ is the relative position between the mth UAV and the kth UE.

3.1.2. DRL action

In each timestep, the mth UAV selects action b_l^m from the discrete action space $\mathcal{B}$. The selected action determines the next position of the UAV. If the action combination selected by the DRL-based controller is likely to cause collisions between UAVs, a different action is selected from the action space. Note that the objective of this paper is to select the set of actions that maximize the reward function at each time step.

3.1.3. DRL reward

The UAVs move in the direction to maximize the cumulative reward by improving system sum-rate and minimizing the inter-cell interference observed from other UAVs.

The reward of the mth UAV at time step n is formulated as follows:

$$r_m^n = \mathcal{R}_{m,1} + \mathcal{R}_{m,2} \tag{8}$$

where $\mathcal{R}_{m,1} = \eta_1 G_1(\mathbf{s}_m^n, b_m^n)$, η_1 is the magnitude of the reward, $G_1(\mathbf{s}_m^n, b_m^n) = \mathcal{I}_m^{n-1} - \mathcal{I}_m^n$ compares the interference level caused by the UAV between time step $n-1$ and n and $\mathcal{I}_m^n = \sum_{j \neq m} P g_{k,j} L_{k,j}(r_{k,j} d_{k,j})$. UAVs get positive $\mathcal{R}_{m,1}$ when the UAV moves to a position which decreases the interference. $\mathcal{R}_{m,2} = \mathcal{C}^{n-1} - \mathcal{C}^n$ is a global reward related to the total sum-rate where $\mathcal{C} = \sum_{m \in \mathcal{U}} \sum_{k \in \mathcal{B}} a_{k,m} R_{k,m}$. The UAV gets a positive reward for $\mathcal{R}_{m,2}$ if it moves to a position that increases the sum-rate, it gets a negative reward if the sum-rate decreases and no reward if the sum-rate remains the same.

The optimal placement is taken when the UAV cannot improve its reward by taking any further action.

3.2. Reinforcement learning training

We describe the DRL employed in optimizing the placement and trajectory of the multiple UAVs, each UAV observes its state and selects an action from a global dueling network. The network is updated by the interaction of distributed agents (UAVs). The details of the multiple UAVs control algorithm are provided in Algorithm 1. In the beginning, the network is initialized with random weights of 0. Also, a replay memory, which stores the recent transition tuples, is initialized (lines 1–3). For a given episode, the DRL-environment is initialized, the multiple UAVs learn various position strategies for N consecutive time-steps without any stopping criterion (line 4–6). Each agent selects an action corresponding to the maximum Q value with probability $1-\epsilon$ or randomly with probability ϵ. The probability ϵ decreases as the training is repeated (lines 7–10). The weights of the Q network are updated, while reducing the loss function with the optimizer. Then, target Q network is updated, duplicating the weights of Q network every N time-steps (lines 12–15).

Algorithm 1: Multi-UAV Control for Spectral Efficiency Maximization based on DRL

Input: State vector of m-th agent
Output: Action of agent m-th agent
1 Initialize Q network with random weights;
2 Initialize target Q network Q^- with weights;
3 Initialize replay memory $\mathcal{D}$ to capacity N_D;
4 **for** $Episode = 1, \ldots, N_t$ **do**
5 Initialize environment;
6 **for** *Timestep* $n := 1, \ldots, N$ **do**
7 **for** *Agent* $m := 1, \ldots, M$ **do**
8 Select a greedy action;
9 $a_m^n = \underset{a_m}{\text{argmax}}\, Q^i(s_m^n, a_m)$ with probability $1-\epsilon$
 or a random action with probability $-\epsilon$
 Execute action a_m^n;
10 Store the transition sample $(s_m^n, a_m^n, r_m^n, s_m^{n+1})$ in replay memory;
11 **end**
12 Update Q network;
13 Update target Q network every N time steps;
14 **end**
15 **end**

4. Simulation results and discussion

In order to study the performance of the proposed algorithm, we consider a 100×40 meters area. The set of UEs $\mathcal{S}$ are distributed as a realization of a HPPP with intensity $\lambda = 16 \times 10^{-4}$, user/m^2. The deployed UEs are allowed to move about using a random walk model. The UAVs are deployed using a 3D PPP with intensity $\rho\lambda_u$. We assume the transmit power for each UAV $P = 10$ dBm. In order to compute the average spectral efficiency in (5), we employ Monte Carlo simulations with 5000 runs, the average is computed over the small-scale fading which is then used in (7b)–(7d). The simulation settings are summarized in Table 1.

The proposed scheme is compared with exiting UAV placement schemes. The benchmark schemes are described as follows;

1. Learn as you fly (LAYF) scenario: In this scenario, we implement the algorithm in [5]. Due to space limitation, we briefly describe the algorithm and invite readers to the algorithms in the reference. The LAYF-Nearest scenario assigns users to their nearest UAV based on the K-means algorithm. The LAYF scenario optimizes the UAV-UE association without necessarily associating the UE to the nearest UAV.

Table 1
Summary of Simulation Parameters.

Parameters	Values
Future time instance T_p	2 ms
Sampling time t_p	0.2 ms
Number of UEs K	3
Number of UAVs M	30
$R_{n,k}$ Random in	[90,100] Mbps
Transmit power P_{tx}	10 dBm
Simulation environment	100×40 m
UAV altitude	25 m

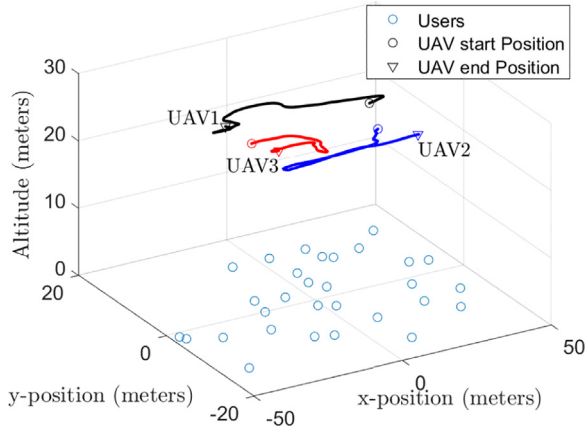

Fig. 1. 3D network configuration for ML based UAV trajectory training.

2. Random scenario: In this scenario, we generate a random feasible solution in each iteration. To this end, we first create a random matrix of UAVs 3D positions within the studied area. Then, we randomly assign users to UAVs such that each user is assigned to exactly one UAV. Then, we check if the constraints on bandwidth and spectral efficiency are satisfied. A random UE is disconnected if the spectral efficiency and bandwidth constraint are not satisfied until the constraints of the UAV are satisfied.

In Fig. 1, we plot the initial and final positions of multiple UAVs based on the UE positions. As depicted in the figure, the UAVs dynamically change their positions starting from their initial points, and move towards their served users in a few steps before reaching their optimal locations. It can also be seen from the figure that the UAVs can choose actions that allow them to explore the environment and choose the optimal position for the entire network. Although the heights of the UAVs are fixed for this result, the results can easily be extended to variable heights my training with different UAV positions in the z-direction.

Fig. 2 shows the convergence of the sum-rate function vs the number of steps under Rayleigh channels. The figure shows how the sum-rate evolves over the number of step. Clearly, the proposed ML-based approach significantly improves the overall sum-rate as compared with the baseline schemes. From the figure, it can be observed that the random deployment of the UAVs approach always yields a suboptimal result. When compared with the K-means LAYF approach

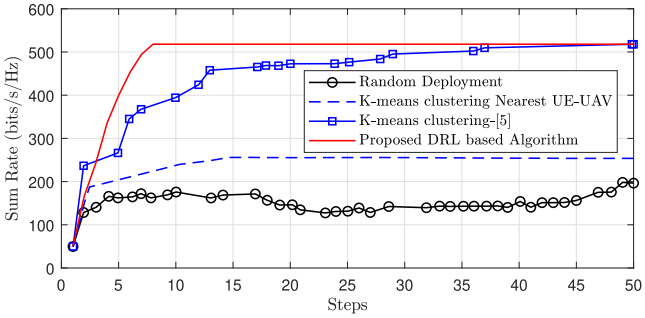

Fig. 2. Algorithms convergence over Rayleigh channels.

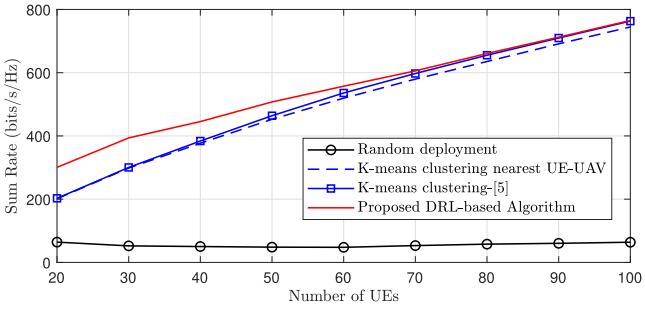

Fig. 3. Sum-rate versus number of users after algorithm convergence.

in [5], the proposed method achieves optimal rate much faster on the 8-th step, while the LAYF approach requires about 45 steps.

In Fig. 3 the sum-rate after convergence of the algorithms against the number of the ground users is plotted for a fixed bandwidth of the UAVs (400 MHz). It can be observed that the proposed DLR based approach provides a better performance as compared to the benchmark schemes.

5. Conclusion

In this paper, we propose a DRL-based UAV placement and trajectory optimization for UAV-assisted wireless networks. The DRL-based multiple UAVs placement can optimize the UAVs trajectory to improve the total sum-rate of the network. Each UAV obtains the relative position of other UAVs and users. Based on this information, we exploited a reward function consisting of interference difference reward and rate difference reward to quantify the effectiveness of the selected actions. Simulation results demonstrated that the proposed DRL-based multiple UAVs placement can be deployed for fast self organizing networks compared with the baseline schemes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP)

grant funded by the Korea government(MSIT) (No. 2020-0-00045, Development of movable high-capacity mobile communication infra for telecommunication disaster and rescue).

References

[1] M. Mozaffari, X. Lin, S. Hayes, Toward 6G with connected sky: UAVs and beyond, IEEE Commun. Soc. Mag. 59 (12) (2021) 74–80.

[2] J. Won, D.-Y. Kim, Y.-I. Park, J.-W. Lee, A survey on UAV placement and trajectory optimization in communication networks: From the perspective of air-to-ground channel models, ICT Express (2022).

[3] R. Bajracharya, R. Shrestha, S. Kim, H. Jung, 6G NR-U based wireless infrastructure UAV: Standardization, opportunities, challenges and future scopes, IEEE Access 10 (2022) 30536–30555.

[4] I.A. Elnabty, Y. Fahmy, M. Kafafy, A survey on UAV placement optimization for UAV-assisted communication in 5G and beyond networks, Phys. Commun. 51 (2022) 101564.

[5] H. El Hammouti, M. Benjillali, B. Shihada, M.-S. Alouini, Learn-as-you-fly: A distributed algorithm for joint 3D placement and user association in multi-UAVs networks, IEEE Trans. Wirel. Commun. 18 (12) (2019) 5831–5844.

[6] C. Wang, D. Deng, L. Xu, W. Wang, Resource scheduling based on deep reinforcement learning in UAV assisted emergency communication networks, IEEE Trans. Commun. 70 (6) (2022) 3834–3848.

[7] Y. Wang, Z. Su, N. Zhang, R. Li, Mobile wireless rechargeable UAV networks: Challenges and solutions, IEEE Commun. Mag. 60 (3) (2022) 33–39.

[8] R. Ding, F. Gao, X.S. Shen, 3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach, IEEE Trans. Wirel. Commun. 19 (12) (2020) 7796–7809.

[9] E.N. Almeida, A. Coelho, J. Ruela, R. Campos, M. Ricardo, Joint traffic-aware UAV placement and predictive routing for aerial networks, Ad Hoc Netw. 118 (2021) 102525.

[10] J. Moon, S. Papaioannou, C. Laoudias, P. Kolios, S. Kim, Deep reinforcement learning multi-UAV trajectory control for target tracking, IEEE Internet Things J. 8 (20) (2021) 15441–15455.

[11] A. Al-Hourani, S. Kandeepan, A. Jamalipour, Modeling air-to-ground path loss for low altitude platforms in urban environments, in: 2014 IEEE Global Commun. Conf., 2014, pp. 2898–2904.