01 Jan 2023

Q-Learning for Sum-Throughput Optimization in Wireless Visible-Light UAV Networks

Yuwei Long

Nan Cen
Missouri University of Science and Technology, nancen@mst.edu

Recommended Citation

Y. Long and N. Cen, "Q-Learning for Sum-Throughput Optimization in Wireless Visible-Light UAV Networks," *IEEE INFOCOM 2023 - Conference on Computer Communications Workshops, INFOCOM WKSHPS 2023*, Institute of Electrical and Electronics Engineers, Jan 2023.
The definitive version is available at https://doi.org/10.1109/INFOCOMWKSHPS57453.2023.10225783

IEEE INFOCOM 2023 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) | 978-1-6654-9427-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOMWKSHPS57453.2023.10225783

Q-Learning for Sum-Throughput Optimization in Wireless Visible-Light UAV Networks

Yuwei Long, Nan Cen
Department of Computer Science, Missouri University of Science and Technology, Rolla, MO
Email:{ylw22, nancen}@mst.edu

***Abstract*—Unmanned aerial vehicles (UAVs) have been adopted as aerial base stations (ABSs) to provide wireless connectivity to ground users in events of increased network demand, and points-of-failure infrastructure (such as in disasters). However, with the existing crowded radio frequency (RF) spectrum, UAV ABSs cannot provide high-data-rate communication required in 5G and beyond. To address this challenge, visible light communication (VLC) is proposed to be equipped on UAVs to take advantage of the flexible and on-demand deployment feature of the UAV, and the high-data-rate communication of the VLC. However, VLC has strong alignment requirements between transceivers, therefore, how to determine the position and orientation of the UAV is critically important for sum-throughput improvement. In this paper, we propose two Q-learning based methods to maximize the sum throughput of the wireless visible-light UAV network by jointly controlling the position and orientation of the UAV. The results show that the proposed approaches can achieve a network-wide data rate very close to the optimal solution obtained by exhaustive search and outperform up to 18% compared with the intuitive centroid-based method. Computation complexity is also evaluated, where results showing that the proposed two Q-learning based methods can both consume less computational time, i.e., approximately 9 times and 210 times less on average than that of the exhaustive search approach.**

***Index Terms*—Visible Light Networking, Unmanned Aerial Vehicles, Throughput Optimization, Q-learning.**

I. Introduction

Wireless Unmanned aerial vehicle (UAV) networks have been envisioned as a key technology in 5G and beyond, because of their mobility, flexibility, and on-demand deployment nature, which enables a diverse set of applications, including military, surveillance and monitoring, wireless communication enhancement, medical goods delivery, and post-disaster operations [1] [2] [3]. Figure 1 illustrates a number of scenarios of UAVs acting as aerial base stations (ABSs) to provide temporary emergency network services. Taking the example of an extreme event (e.g., festival and sports events), UAV can be deployed to alleviate data traffic congestion, thus assisting the ground base stations in improving connectivity, coverage, and capacity. However, the increased altitude and favorable propagation conditions of wireless UAV networks along with overcrowded radio frequency (RF) spectrum will result in stronger interference among the neighboring cells.

In recent years, visible light communication (VLC) has been proposed as a promising alternative technology to address RF spectrum scrunch because of its massive amounts of the unregulated spectrum ranging from 400 THz – 800 THz [4]. Besides, compared with conventional RF communication

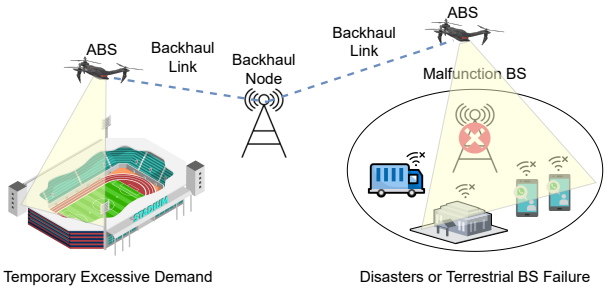

Fig. 1: UAV based aerial base station application scenarios.

technologies, VLC has some unique advantages, such as inherent security due to low penetration, no interference to the RF channel, dual illumination and communication capability, and high-data-rate potential. VLC also has some drawbacks, including limited coverage due to high attenuation, and inflexible deployment due to strong alignment requirement [1]. Thus, the combination of VLC and wireless UAV networks will not only meet the high-speed communication requirement in 5G and 6G, but also improve the flexibility of visible light communication [5], i.e., the dynamic mobility of the UAVs can help maintain a clear line of sight (LoS) requirement of visible light communication.

In the past few years, wireless visible-light UAV networks have attracted more and more attention from researchers, where how to deploy the UAVs in an optimal way to achieve different network control objectives is challenging. A number of works have been conducted to determine the optimal UAV deployment in terms of UAV positions [6]–[13], aiming at improving power efficiency [12] or spectral efficiency [8] [13]. However, to the best of our knowledge, none of these existing works consider controlling the orientations along with the positions of the UAV, which is also significantly important for alignment among transmitters and receivers communicating via VLC. In this paper, we design a new UAV deployment approach for wireless visible-light UAV networks to maximize the sum-throughput by controlling the position and orientation of the UAV. We claim the following three main contributions:

- We formulate mathematically the deployment control problem with the objective of maximizing the network-wide throughput of all ground users by jointly determining the position and orientation of the VLC-based UAV

[1]In VLC, the field of view of the transmitter (e.g., light emitting diode) and receiver (e.g., photodiode) are both limited, therefore, alignment is required to maintain continuous communication.

978-1-6654-9427-4/23/$31.00 ©2023 IEEE

in the drone hot-spot networks. The resulting problem is a mixed integer nonlinear nonconvex programming problem (MINLP), which cannot be solved within polynomial time.
- We propose two Q-learning based solution algorithms to solve the formulated MINLP problem, where two different approaches are designed to construct the action space to control the movement and rotation of the UAV.
- We conduct extensive simulations to evaluate the proposed solution algorithms in terms of optimality and computation time complexity by comparing them with the other two methods, i.e., exhaustive search-based and intuitive centroid-based methods. Results show that the proposed solution algorithms can achieve sum throughput up to about 99.9% optimality obtained by the exhaustive search method but with significantly reduced computation time up to approximately 303 times as shown in Sec. V.

The organization of this article is as follows. In Section II, the related works are reviewed. The system model and problem formulation are then presented in Section III. In Section IV, two solution algorithms are designed and discussed in detail. Simulation results are presented and discussed in Section V. Finally, we draw the main conclusions in Section VI.

II. Related Works

In recent years, VLC-enabled UAV networking has drawn increasing research attention [6]–[11], where how to deploy the UAVs to improve the system performance is one of the hot research topics [6]–[11] [14]. In [6], the authors propose an iterative algorithm to optimize the deployment of VLC-enabled UAVs, aiming at minimizing power consumption by considering the interference caused by the signal transmission as well as the illumination of UAVs. [7] presents a deep learning-based approach to dynamically deploy visible-light UAVs, thus optimizing energy-efficiency, where a learning framework of gated recurrent units (GRUs) with convolutional neural networks (CNNs) is adopted. The work in [8] considers a UAV-assisted wireless visible light network using nonorthogonal multiple access (NOMA), where a joint problem of UAV power allocation and placement is formulated to maximize the sum data rate by jointly considering the maximum power consumption constraint, quality of services of users, and UAV position. The authors also propose a Harris hawks optimization-based algorithm to obtain the sub-optimal solutions. The authors in [10] propose a federated learning framework based on convolutional auto-encoder to predict the illumination distribution in VLC-enabled UAV networks, based on which the optimal UAV deployment and user association policy are then determined to minimize the total transmission power of the UAVs. The work in [11] studies optimal deployment of UAVs over reconfigurable intelligent surfaces (RISs) assisted VLC system, where UAVs are designed to provide data transfer and illumination simultaneously. The authors formulate a power minimization problem by jointly controlling the UAV positions, the phase shift of RISs, and user and RIS association, and then design two solution algorithms.

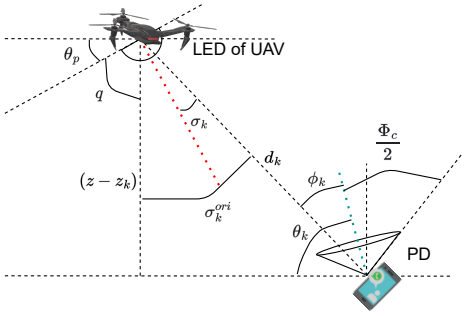

Fig. 2: Geometry LoS propagation model in VLC.

These papers are mainly focused on the UAV position determination problem to improve the system performance, with respect to either the total power consumption of UAVs or the sum data rate. In our paper, instead, besides the UAV location information, we also consider the orientation of the UAVs, which plays a significant role in maintaining the line of sight communication link between the UAV and users.

III. System Model and Problem Formulation

A. Model Description

We consider an infrastructureless wireless VLC-enabled UAV downlink access network, where a set of ground users $\mathcal{K} = \{1, \ldots, K\}$ are served by the UAV within the predefined geographical area. We assume that the location and orientation information of the ground users can be obtained by the devices themselves [15] and shared with the drones. The information of the location and orientation of the UAV and the k-th user are expressed as $\mathbf{p} = (x, y, z, \theta_p)$ and $\mathbf{N}_k = (x_k, y_k, z_k, \theta_k), k \in \mathcal{K}, \mathbf{N}_k \in \mathbb{R}^{4\times 1}$, respectively. Our objective is to *maximize the network-wide sum throughput by jointly controlling the movement of the aerial drones with respect to location and orientation.*

B. Transmission Model

We consider the Lambertian radiation pattern [16] in the visible light UAV networking system. Without loss of generality, we do not consider the diffusion of visible light (i.e., multipath propagation) in outdoor environments. Therefore, the only considered LoS channel impulse response of the VLC link for user k is given as [16]:

$$h_k^{LoS} = \begin{cases} \frac{A(m+1)}{2\pi d_k^2} P_t cos^m(\sigma_k) T_s(\phi_k) g(\phi_k) \cos(\phi_k) & 0 \le \phi_k \le \Phi_c, \\ 0 & \text{otherwise}, \end{cases} \quad (1)$$

where $m = \frac{\ln 2}{\ln(\cos \phi_{1/2})}$ represents the Lambertian emission order with $\phi_{1/2}$ being the half illuminance power of a transmitter. A denotes the physical area of the photodiode (PD) at the receiver side. d_k denotes the distance between the light emitting diode (LED) transmitter of UAV and the PD receiver of ground user k, calculated as $d_k = \sqrt{(x - x_k)^2 + (y - y_k)^2 + (z - z_k)^2}$. P_t is the transmitted power of UAV. σ_k and ϕ_k represent the irradiance angle and the incidence angle between UAV and ground user k, respectively. Φ_c denotes the field of view (FoV) of the PD. $T_s(\phi_k)$ represents the optical filter gain. $g(\phi_k)$ is the optical

concentrator gain, given as:

$$g(\phi_k) = \begin{cases} \frac{n^2}{\sin^2 \Phi_c}, & 0 \le \phi_k \le \Phi_c, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

We denote σ_k^{ori} being the irradiance angle to user k when UAV is facing directly downwards, calculated as $\sigma_k^{ori} = \arccos(\frac{z-z_k}{d_k})$. Given σ_k^{ori}, the incidence angle ϕ_k from UAV to ground user k can be obtained as:

$$\phi_k = \sigma_k^{ori} - (90^\circ - \theta_k). \quad (3)$$

As shown in Fig. 2, the geometric relationship among UAV rotation angle θ_p, the original irradiance angle σ_k^{ori}, and the instantaneous irradiance angle σ_k is given as below:

$$\theta_p + q = q + \sigma_k^{ori} - \sigma_k = 90^\circ, \quad (4)$$

where q is the included angle between the perpendicular line to the ground and the extension line of the rotation angle of the UAV. Finally, we can obtain the instantaneous irradiance angle σ_k with respect to θ and σ_k^{ori} as:

$$\sigma_k = |(\theta_p - \sigma_k^{ori})|. \quad (5)$$

The channel capacity C_k of user k is then calculated as:

$$C_k(\mathbf{p}) = B\log_2\left(1 + \frac{e}{2\pi}\left(\frac{\xi P_t h_k^{LoS}}{n_k^w}\right)^2\right), \quad (6)$$

where B is the bandwidth, e is the Euler's number, ξ is the illumination target, and n_k^w is the standard deviation of the additive white Gaussian noise for user k.

Problem Statement. The network control objective can then be stated as maximizing the sum data rate of all ground users in K by jointly determining the position and orientation of the UAV. The problem is formulated as:

$$\begin{aligned} \text{Problem 1: Given: } & \mathbf{N}_k, P_t, \Phi_c, \Sigma_c \\ & \underset{\mathbf{p}}{\text{Maximize}} \ \ f = \sum_{k \in \mathcal{K}} C_k(\mathbf{p}) \\ & \text{Subject to: } \ \phi_k \le \frac{\Phi_c}{2}, \\ & \qquad |(\theta_p - \sigma_k^{ori})| \le \frac{\Sigma_c}{2}, \\ & \qquad 0 \le x_k \le x_b, \ \ k \in \mathcal{K}, \\ & \qquad 0 \le y_k \le y_b, \ \ k \in \mathcal{K}, \end{aligned} \quad (7)$$

where x_b and y_b represent the predefined area boundaries, Σ_c denotes the field of view of the LED transmitter of the UAV. The problem is a mixed integer nonlinear nonconvex programming problem because of the nonlinear nonconcave function $C_k(\mathbf{p})$ with respect to $\mathbf{p}$ in (1) [13]. Given an arbitrary such problem, how to design a solution algorithm to achieve the global optimum is still an open problem.

IV. Q-LEARNING BASED SOLUTION ALGORITHMS

To solve the resulting MINLP problem in Sec. III, a heuristic solution algorithm is required to find the near-optimal solutions. Traditional widely adopted heuristic methods, such as genetic algorithms, may get stuck at some local minima. In recent years, learning-based methods have gained increasing attention to solve MINLP problems. In this paper, we propose solution algorithms based on Q-learning. This is motivated by that Q-learning is model-free and can learn from trial and error to obtain optimal and nearly-optimal solutions without knowing the environment in advance compared to model-based methods [17].

In Q-learning models, the goal of the agent is to learn the best policy from the Q-values in the Q-table. At each iteration time step, we update the Q-table values using Bellman optimality equation [18] after completing every transition and observing the current state-action pair (s_t, a_t) at time t, given as:

$$Q^{new}(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha r_t + \gamma Q(s_{t+1}, a^*), \quad (8)$$

where α is the learning rate. γ is the discount factor within the range [0, 1], representing how important a future action would be. s_{t+1} is the next state after taking action a_t at state s_t and a^* is the action that results in the maximum Q-value of all state-action pairs in state s_{t+1} [19].

The generalized solution algorithm framework is shown in Algorithm 1, where we assume that the visible-light UAV is equipped with an autonomous agent which takes an action and in turn, receives a reward and then makes a transition to a new state. Next, we explicitly define the *States, Actions, and the reward function* of the agent.

A. States

We consider a combination of the three-dimensional (3D) position and the orientation of the UAV as *States*. The agent discretizes the continuous state space with respect to the predefined serving area as well as the feasible rotation angles. The state space is expressed as a tuple, i.e., $\mathcal{S} := (X : \{0, 0+\eta, 0+2\times\eta, \ldots, x_b\}, Y : \{0, 0+\eta, 0+2\times\eta, \ldots, y_b\}, Z : h, \Theta_p : \{-r_b^\circ, -r_b^\circ + \delta^\circ, \ldots, 0^\circ, \ldots, r_b^\circ - \delta^\circ, r_b^\circ\})$, where (X, Y, Z) and Θ_p represent the state in terms of the agent's location and orientation, respectively. In X and Y, η reflects on how small we discretize our environment, i.e., the grid size. In the current environment, we consider the agent to be set at a certain altitude h, resulting in only one state for Z. For Θ_p, $-r_b^\circ$ and r_b° define the maximum rotation angle to the left or

Algorithm 1 Solution Algorithm

Data:
Predefine $\mathbf{N}_k$, P_t, Φ_c, Σ_c.
Initialize $Q(s, a)$, $\forall s = (x, y, z, \theta) \in S$, $\forall a \in A$, $\{\gamma\}$, $\{\epsilon\}$.
Result: Obtain $\{f\}$ and $\{\mathbf{p}\}$ when stopping criterion is met.
Initialize all $Q(s, a)$ table to zero
for *each Episode i* **do**
  Initialize $s_0 \leftarrow (0, 0, h, 0)$; $t = 1$.
  while *true* **do**
    Observe state s_t, and choose action a_t:
    Generate random number ρ in $(0, 1)$
    if $\rho > \epsilon$ **then**
      $a_t = argmax_a Q(s_t, a)$;
    else
      $a_t = Random(A)$;
    end
    Take action a_t;
      $s_t \leftarrow s_{t+1}$
    Update UAV's location to (x_t, y_t), and orientation to θ_t.
    Calculate r_t with the new location and orientation of the agent using (7)
    $Q^{new}(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r(s_t, a_t) + \gamma Q(s_{t+1}, a^*)$
    $t++$;
    if *meet the stopping criterion* **then**
      Output $\{f\}$ and $\{\mathbf{p}\}$
      break;
    end
  end
end

the right of the agent, while δ° represents the rotation step size. In summary, the total number of states for our scenario is $|\mathcal{S}| = |X| \times |Y| \times |Z| \times |\Theta_p|$.

B. Actions

At each time-step t, the UAV takes action $a_t \in \mathcal{A}$, where $\mathcal{A}$ is divided into two subspaces, i.e., UAV movement space $\mathcal{A}_\mathcal{M}$ and UAV rotation space $\mathcal{A}_\mathcal{R}$.

Movement Space: We discretize the UAV agent's moving action space [19] as: $\mathcal{A}_\mathcal{M} := \{up, right, down, left, stay\}$, with $|\mathcal{A}_\mathcal{M}| = A_M$. Our motivation for discretizing the action space is to ensure quick convergence of the agent to quickly deploy high-quality service to the ground users.

Rotation Space: We design two methods to conduct the rotation of the UAV: (i) Q-learning discrete continuous rotation action space, i.e., *Q-DCA*, denoted as $\mathcal{A}_\mathcal{R}$; and (ii) Q-learning discrete reduced rotation action space, i.e., *Q-DRA*, denoted as $\mathcal{A}_\mathcal{R}^\mathcal{R}$. The details are discussed in the following.

- Q-DCA: In Q-DCA, we discretize the feasible rotation space by splitting it into a set of bins with step size δ°, denoted as $\mathcal{A}_\mathcal{R} := \{-r_b^\circ, -r_b^\circ + \delta^\circ, \ldots, 0^\circ, \ldots, r_b^\circ - \delta^\circ, r_b^\circ\}$. The total number of the rotation actions in Q-DCA is $|\mathcal{A}_\mathcal{R}| = (r_b^\circ - (-r_b^\circ))/\delta^\circ + 1) = 2r_b^\circ/\delta^\circ + 1$. We can see that the smaller δ° is, the larger the rotation action space is and thus the slower the convergence of the Q-learning model is as proven in [20].
- Q-DRA: Motivated by the fact that reducing the number of actions can help with exploration, as there are fewer actions to try, which will help improve the sample efficiency of Q-learning model training. Therefore, we further reduce the discretized continuous rotation action space by just defining three discrete rotation choices: negative, zero, and positive [2], denoted as $\mathcal{A}_\mathcal{R}^\mathcal{R} := \{a_t^r, a_t^r \pm \delta_r^\circ\} \bigcap \Theta_p$, where a_t^r represents the rotation state at time t. In each time-step, the UAV agent can only turn left or right respectively by δ°, or stay at the current orientation. The total number of the rotation action space is $|\mathcal{A}_\mathcal{R}^\mathcal{R}| = 3$.

In summary, the numbers of the whole action space for Q-DCA and Q-DRA are $|\mathcal{A}_\mathcal{M}| + |\mathcal{A}_\mathcal{R}|$ and $|\mathcal{A}_\mathcal{M}| + |\mathcal{A}_\mathcal{R}^\mathcal{R}|$, respectively.

C. Reward

The goal of the learning agent is to learn a policy that maximizes the sum throughput of the visible-light enabled UAV network by controlling the position and orientation of the UAV in a real-time fashion. The reward obtained by the UAV agent at time t is given as:

$$R(s_t, a_t, s_{t+1}) = \begin{cases} \sum\limits_{k \in \mathcal{K}} C_k(\mathbf{p}) & \text{if constraints in (7) met,} \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

where $\sum\limits_{k \in \mathcal{K}} C_k(\mathbf{p})$ is the objective function in (7). If the learned position and orientation satisfy the constraints in (7), the sum throughput of the network is set as the reward, otherwise, the reward is 0, which will help force the agent to avoid learning unfeasible solutions.

[2]This is commonly used in camera rotation, where the agent can only select to turn the camera left or right or not at a fixed rate per step [21], which is also suitable for the UAV agent in the proposed scenario.

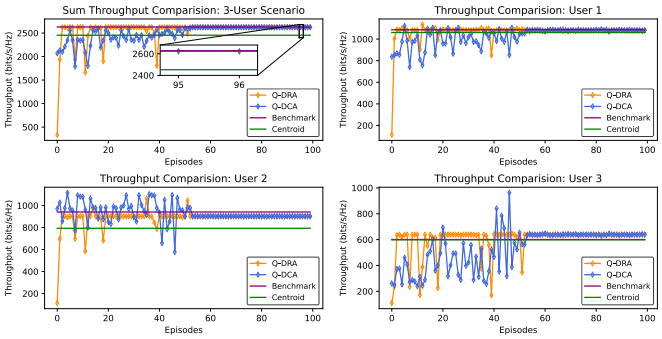

Fig. 3: Sum throughput and individual throughput comparison for the 3-user scenario.

D. Exploration and Exploitation Strategy

In Q-learning, the exploration-exploitation dilemma is significantly important. For moderately realistic problems, problem sizes are vast and computations are expensive. Thus, we want to learn accurate values for good states, rather than wasting precious (computational) budget on low-quality ones. In our Q-learning model, we adopt ϵ-greedy strategy that has been shown to be able to achieve better performance [22]. At the beginning of learning, ϵ-greedy strategy will force the agent to operate in exploration mode since the agent has no idea of the environment. As the algorithm runs, the ϵ value is decreasing and the chance of being in exploration mode is also decreasing, which allows the agent to be in the exploitation mode to exploit what it has learned. The detailed rules of the ϵ-greedy method is given below:

$$\pi_{a_t \in \mathcal{A}}(s_t \in \mathcal{S}) = \begin{cases} \text{argmax}_{a \in \mathcal{A}} Q(s_t, a) & \text{probability } 1 - \epsilon, \\ \text{random } a \in \mathcal{A} & \text{probability } \epsilon, \end{cases} \quad (10)$$

where ϵ is the time-variant parameter used for updating the ϵ-greedy strategy, and a represents any feasible action. The adopted ϵ-greedy scheme can help balance the exploration and exploitation in the learning process, thus improving learning efficiency [22].

E. Computational Complexity

Based on the theorem in [20], the time complexity of the proposed Q-DCA and Q-DRA are $O(|\mathcal{S}| \times (|\mathcal{A}_\mathcal{M}| + |\mathcal{A}_\mathcal{R}|))$ and $O(|\mathcal{S}| \times (|\mathcal{A}_\mathcal{M}| + |\mathcal{A}_\mathcal{R}^\mathcal{R}|))$, respectively.

V. Performance Evaluation

We evaluate the performance of the proposed solution algorithms by considering a network area of $7 \times 7 \times 10$ m^3, with $K = \{1, 2, \ldots, 10\}$ users served by one UAV. The altitude

TABLE I: Summary of Parameters

Parameter	Value
Bandwidth (B)	$B = 20$ MHZ
Transmitted electrical power (P_t)	$P_t = 1$ W
Optical filter gain ($T_s(\phi_k)$)	$T_s(\phi_k) = 1$
Field of view of the PD (Φ_c)	$\Phi_c = \pi$
Field of view of the LED (Σ_c)	$\Sigma_c = 2\pi/3$
Optical concentrator gain ($g(\phi_k)$)	$g(\phi_k) = 2.25, 0 \leq \phi_k \leq \pi; g(\phi_k) = 0, \phi_k \geq \pi$
Area of PD (A)	$A = 1$ cm^2
Learning rate (α)	$\alpha = 0.8$
Discount factor (γ)	$\gamma = 0.1$

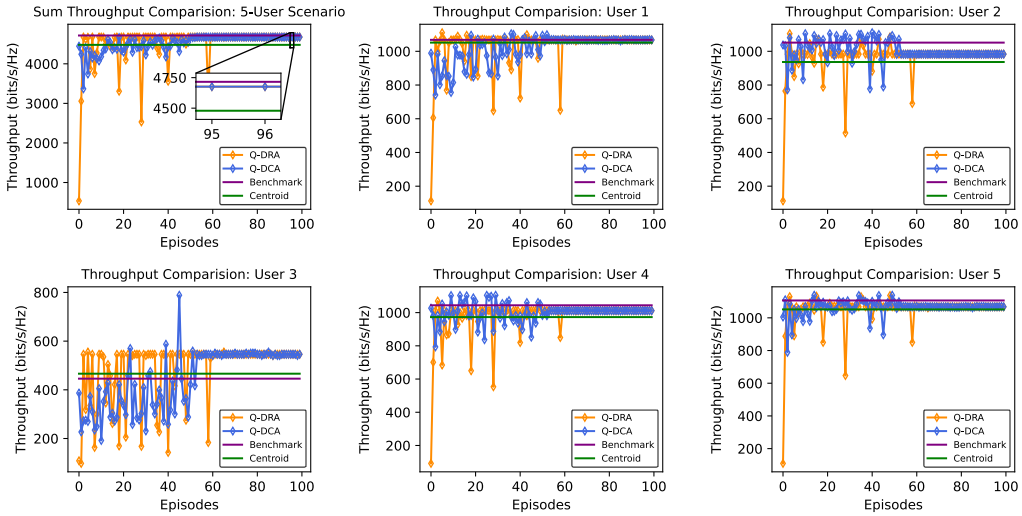

Fig. 4: Sum throughput and individual throughput comparison for the 5-user scenario.

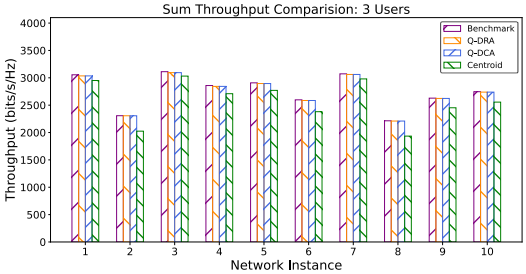

Fig. 5: Sum throughput comparison for the 3-user scenario.

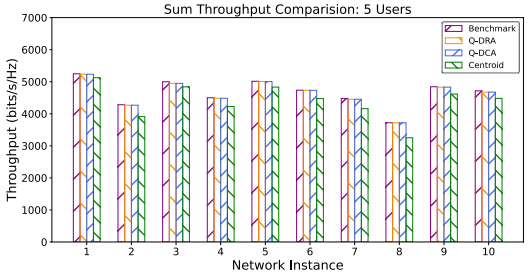

Fig. 6: Sum throughput comparison for the 5-user scenario.

of the UAV is set to $z = h = 10$ meters. Without loss of generality, θ_k is set to 90°, denoting that all the ground users are facing directly upwards. Table I includes the networking-related parameters for performance evaluation and the learning parameters used in the proposed Q-learning models.

A. Compared Methods

To comprehensively validate the proposed two solution algorithms, i.e., *Q-DCA* and *Q-DRA*, we also implement two other methods to compare: (i) exhaustive search and (ii) intuitive centroid-based method, denoted as *Benchmark* and *Centroid*, respectively.

Exhaustive search: Due to the limited computation capability of the computer, it is difficult to obtain the global optimal solution in the continuous state space. Therefore, we divide the state space with much smaller step sizes than that in Q-DCA and Q-DRA, i.e., 0.01 m and 1° for moving and rotating, respectively. The exhaustive search method will enumerate all the combinations within the state space to find the optimal solution, which will be considered as the benchmark in the following performance comparisons.

Centroid-based method: The location of the UAV is determined by the geometric center of the topology of the ground users. In this method, the rotation of the UAV is not considered and the UAV is set to face directly downwards.

B. Performance Comparison

Figures 3 and 4 report the sum throughput and individual throughput for each user achieved by the proposed two methods and the other two compared methods for the 3-user scenario and 5-user scenario, respectively. It can be seen that the proposed Q-DCA and Q-DRA can nearly reach the results obtained by Benchmark within 60 episodes and outperform Centroid by up to 18%. To further evaluate the effectiveness of Q-DCA and Q-DRA, we also compare the achievable sum-throughput by testing 10 different network topology instances for both 3-user and 5-user scenarios. The results are shown

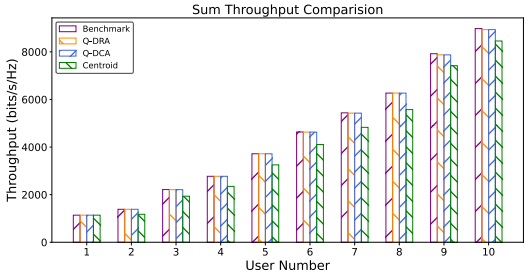

Fig. 7: Sum throughput comparison for scenarios with different user numbers.

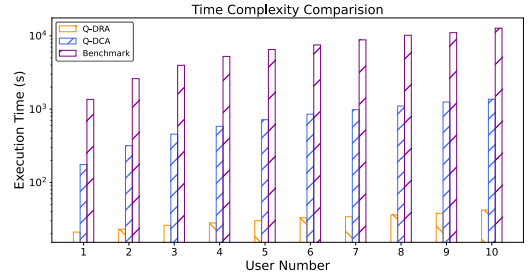

Fig. 8: Time complexity comparison for scenarios with different user numbers (A base-10 log scale is used for the Y axis).

in Figs. 5 and 6, where we can observe that Q-DCA and Q-DRA both achieve performance very close to the optimum obtained by Benchmark and outperform Centroid in all tested instances. For some network instances, big differences are not shown in the results because users are densely deployed within a small area, thus the solutions obtained by all four methods are very close. We can also observe the fluctuations of the sum throughput among all the test instances. This is caused by the deployment of users, i.e., if users are sparsely deployed (e.g., further from each other) within the predefined area, the achieved sum data rate is smaller due to the larger attenuation, otherwise, the sum-throughput is larger.

We further validate the feasibility of our proposed Q-DRA and Q-DCA algorithms with different numbers of users, ranging from 1 to 10 as shown in Fig. 7. We can see that Q-DCA and Q-DRA achieve better results compared with Centroid and achieves a performance very close to the optimum obtained by Benchmark in all tested scenarios besides the 1-user setting (where UAV is directly located on top of the user for all methods).

We finally evaluate the computation complexity for Q-DCA, Q-DRA, and exhaustive search based Benchmark method. Centroid is not included because its computation complexity is constant. Results are shown in Fig. 8, where we can clearly see that the proposed Q-DRA has the lowest computation time complexity compared with Q-DCA and Benchmark in all tested scenarios with user number ranging from 1 to 10. The computation complexity of Q-DCA takes about 23 more times on average than that of Q-DRA, which help further validate that reducing the action space can accordingly help decrease the computation complexity. It also clearly shows that the exhaustive search based Benchmark will result in significant computation complexity compared with Q-DCA and Q-DRA, especially as the scenarios become more complex, such as in terms of user number.

VI. Conclusion

In this paper, we have investigated the problem of optimal deployment of VLC-enabled UAVs, by jointly considering the position and orientation of the UAV to maximize the sum-throughput. We have proposed two Q-learning based solution algorithms: Q-DCA and Q-DRA. Extensive simulations are conducted to validate the performance of the proposed algorithms in terms of optimality and computation time complexity by comparing them with exhaustive search and intuitive centroid-based methods. Numerical results show that both proposed Q-DRA and Q-DCA can increase sum throughput by up to 18% compared with intuitive centroid-based methods, and achieve up to 99.9% optimality of the benchmark. Results also show that Q-DCA and Q-DRA can obtain the solutions about 9 times and 210 times faster on average than that of the exhaustive search method, respectively.

References

[1] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV Cellular Communications: Practical Aspects, Standardization Advancements, Regulation, and Security Challenges," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3417–3442, Fourth Quarter 2019.

[2] M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2334–2360, Third Quarter 2019.

[3] I. Bor-Yaliniz, M. Salem, G. Senerath, and H. Yanikomeroglu, "Is 5G Ready for Drones: A Look into Contemporary and Prospective Wireless Networks from a Standardization Perspective," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 18–27, February 2019.

[4] N. Cen, J. Jagannath, S. Moretti, Z. Guan, and T. Melodia, "LANET: Visible-light ad hoc networks," *Ad Hoc Networks*, vol. 84, pp. 107–123, 2019.

[5] N. Cen, "FLight: Toward Programmable Visible-Light-Band Wireless UAV Networking," in *Proc. of the Workshop on Light Up the IoT*, London, United Kingdom, September 2020.

[6] Z. Z, Y. Yang, C. Guo, M. Chen, S. Cui, and H. V. Poor, "Power Efficient Deployment of VLC-enabled UAVs," in *Proc. of IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, London, UK, August 2020.

[7] Y. Wang, M. Chen, Z. Yang, T. Luo, and W. Saad, "Deep Learning for Optimal Deployment of UAVs With Visible Light Communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7049–7063, 2020.

[8] Q.-V. Pham, T. Huynh-The, M. Alazab, J. Zhao, and W.-J. Hwang, "Sum-Rate Maximization for UAV-Assisted Visible Light Communications Using NOMA: Swarm Intelligence Meets Machine Learning," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10 375–10 387, 2020.

[9] A. Amantayeva, M. Yerzhanova, and R. C. Kizilirmak, "UAV Location Optimization for UAV-to-Vehicle Multiple Access Channel with Visible Light Communication," in *Proc. of Wireless Days (WD)*, Manchester, UK, April 2019.

[10] Y. Wang, Y. Yang, and T. Luo, "Federated Convolutional Auto-Encoder for Optimal Deployment of UAVs with Visible Light Communications," in *Proc. of IEEE International Conference on Communications Workshops (ICC Workshops)*, Dublin, Ireland, June 2020.

[11] Y. Cang, M. Chen, Z. Yang, M. Chen, and C. Huang, "Optimal Resource Allocation for Multi-UAV Assisted Visible Light Communication," *arXiv preprint arXiv:2012.13200*, 2020.

[12] S. Khairy, P. Balaprakash, L. Cai, and Y. Cheng, "Constrained Deep Reinforcement Learning for Energy Sustainable Multi-UAV Based Random Access IoT Networks With NOMA," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1101–1115, 2021.

[13] Y. Long and N. Cen, "Sum-Rate Optimization for Visible-Light-Band UAV Networks Based on Particle Swarm Optimization," in *Proc. of the Annual Consumer Communications Networking Conference (CCNC)*, Virtual, January 2022.

[14] X. Liu, Y. Liu, and Y. Chen, "Reinforcement Learning in Multiple-UAV Networks: Deployment and Movement Design," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8036–8049, 2019.

[15] M. Hazas, C. Kray, H. Gellersen, H. Agbota, G. Kortuem, and A. Krohn, "A Relative Positioning System for Co-Located Mobile Devices," in *Proc. of the International Conference on Mobile Systems, Applications, and Services*, Seattle, USA, June 2005.

[16] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, *Optical Wireless Communications: System and Channel Modelling with MATLAB*. Boca Raton, FL, USA: CRC Press, Inc., 2012.

[17] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine, "Model-based value estimation for efficient model-free reinforcement learning," *arXiv*, vol. abs/1803.00101, 2018.

[18] R. Bellman, "Dynamic Programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[19] C. Watkins and P. Dayan, "Q-Learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[20] S. Koenig and R. G. Simmons, "Complexity Analysis of Real-Time Reinforcement Learning," in *AAAI*, vol. 93, Washington, D.C., USA, July 1993.

[21] A. Kanervisto, C. Scheller, and V. Hautamäki, "Action Space Shaping in Deep Reinforcement Learning," in *IEEE Conference on Games (CoG)*, Virtual, August 2020.

[22] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," *Cambridge, MA*, vol. 22447, 1998.