Proximal Trajectory Optimizations for UAV Supported FSO/RF Heterogeneous Networking with QoS Guarantee

Yong-Ce Liu, Zi-Yang Wu

***Abstract*—This letter investigates a hybrid FSO/RF wireless network consisting of UAV and ground-based multi-homing mobile vehicles in a complex environment. By taking advantage of the high flexibility of UAV to reduce the incidence of FSO links being blocked by buildings due to the movement of UAV and vehicles, a PPO-based reinforcement learning algorithm is proposed to optimize the UAV flight trajectories. Finally, under the principle of reducing the use of RF links, the high speed characteristics of the FSO link and the stable characteristics of the RF link are combined to ensure vehicles' quality-of-service.**

***Index Terms*—UAV, Hybrid FSO/RF, Multi-Homing, Reinforcement Learning, PPO**

I. Introduction

As the performance of UAV increases, its application in wireless network is becoming more and more widespread [1], [2]. With its high flexibility, UAV can quickly deploy wireless network in unexpected situations (e.g., natural disaster or war). For example, it can be used as a aerial base station [1] to achieve communication coverage to the ground or as relay [3] to provide relay service for ground users. Moreover, with the radio frequency (RF) band becoming scarce nowadays, due to the high rate, easy deployment and high confidentiality of Free space optical(FSO), the communication system for UAV wireless network based on FSO has been proposed as [4]. However, FSO is susceptible to weather conditions, and is easily blocked by tall buildings in complex environment. At the same time, the RF channel can provide relatively stable service despite its low capacity. Since multi-homing users have the ability to aggregate resources from multiple communication modes, hybrid FSO/RF communication systems are studied [5]. The combination of FSO (high rate ...) and RF (stability) allows UAV to always provide stable communication services to ground users in unexpected situations. However, most of the existing studies on UAV-Hybrid FSO/RF or UAV-Mixed FSO/RF systems [6] are focused on the scenarios with simple environments. That is, the UAV communicates with a fixed-location data center using FSO link and with mobile users using RF link, which allows us not to consider the case of blocked FSO links. [7] proposes a channel outage probability prediction algorithm for mobile optical wireless communications, but it focuses on indoor visible light communications (VLC). [8] points out the fact that FSO channels are easily blocked in urban scenarios, but it does not consider the motion of users or transmitting point. In complex environment, FSO link blocked by high building is a frequent event due to the movement of users and UAV, so it becomes important to study how to avoid FSO link blocking events by UAV movements to ensure the users' quality-of-service (QoS). Model free reinforcement learning is proposed to solve the Online Trajectory optimization problem in UAV optical wireless networks. In [9], Q-Learning is proposed to optimizes the transmit powers at the RF and VLC APs to ensure QoS satisfaction. In [10], Q-Learning is proposed to solve the Trajectory Design for Multi-UAV problem. To the best of the authors' knowledge, the problem of UAV trajectories optimization for ensuring QoS of mobile multi-homing users in complex environment has not been studied so far. In the system studied in this letter, the movement of UAV and users in complex environments can lead to a large number of communication outages. The UAV needs to plan how to fly in the future period to avoid communication outages. To solve this problem, a PPO-based algorithm is proposed to satify the users' QoS under a fixed power use principle.

II. System Model

A. Environment

We investigate a hybrid wireless network consisting of ground-based multi-homing mobile vehicles (called users later) and high-speed UAV in a complex space environment. In the case of known Channel State Information (CSI), the ground multi-homing mobile users build a communication link with the UAV via hybrid FSO/RF. Under a fixed power use principle, the UAV constantly optimizes the flight decision to ensure the users' QoS within a certain period of time. Based on the three-dimensional Cartesian coordinate system, the coordinate position of the UAV can be expressed as $\mathbf{q}_S(t) = [x_S(t), y_S(t), H_S]$ at timeslot t. Note that we assume that the UAV flies at a fixed altitude H_S. Similarly, the position coordinates of the user u can be expressed as $\mathbf{q}_D^{(u)}(t) = \left[x_D^{(u)}(t), y_D^{(u)}(t), 0\right], u \in [0, U]$. Note that we assume that there are a total of U mobile users. From the above we can obtain the distance $l^{(u)}(t)[m] = \|\mathbf{q}_S(t) - \mathbf{q}_D^{(u)}(t)\|, u \in [0, U]$ between the UAV and the user u. In addition, there are several buildings of different heights in the space. These buildings are model as Axially Aligned Bounding Box (AABB), i.e., we represent these buildings as cubes parallel to the coordinate plane. Therefore, the location information of the building n in detail can be expressed by the coordinates of the lower left vertex $Obj_L^n = [x_L^n, y_L^n, 0], n \in [0, N]$ and the coordinates of the upper right vertex $Obj_R^n = [x_R^n, y_R^n, H_R^n], n \in [0, N]$, where H_R^n is the height of the building n and N is the total number of buildings in the space. Considering the practical situation, we assume that the UAV flies under the limits of maximum acceleration $A_{S,max}$ and maximum velocity $V_{S,max}$, and the mobile users travel under the limits of maximum acceleration $A_{D,max}$ and maximum velocity $V_{D,max}$.

B. FSO Channel

Note that we assume that precise pointing, acquisition, and tracking (PAT) algorithm may compensate for the pointing error. At the same time, atmospheric turbulence is ignored due to the UAV flying at low altitude [11]. Therefore, the channel gain of the FSO link between UAV and user u at timeslot t can be expressed as $h_{FSO}^{(u)}(t) = \eta h_p^{(u)}(t)$, where $\eta, h_p^{(u)}(t)$ represent the responsivity of the photo-detector (PD) and atmospheric loss, respectively. Because η depends on the specific hardware device and has no effect on the dynamic environment proposed in this letter, it is assumed that $\eta = 1$.

Atmospheric loss: Based on the Beer-Lambert Law, the atmospheric loss $h_p^{(u)}(t)$ between UAV and user u at timeslot t can be expressed as

$$h_p^{(u)}(t) = e^{-\beta \cdot l^{(u)}(t)} \tag{1}$$

Where $\beta = \frac{3.91}{V \times 10^3} \cdot (\frac{\lambda_0}{550[nm]})^{-p} \cdot \frac{\ln 10}{10}[m^{-1}]$ is determined by the optical wavelength λ_0, which is set to $1550[nm]$ in this letter, $V[km]$ is the visibility, and p is a coefficient related to the size of the particles in the air, which is determined by the Kim Model [12].

Rate For FSO Link: While the capacity of the FSO channel has not been known in a closed-form, the capacity bounds of FSO have been proposed in several papers. In order to ensure the users' QoS as much as possible, we adopt the lower bound of the channel capacity of the FSO link proposed in [13] to represent the achievable rate at the user u. Therefore, at the timeslot t, the achievable rate [Mbps] at the user u can be expressed as

$$R_{FSO}^{(u)}(t) = \frac{B_{FSO}}{2} \cdot \log_2(1 + k\frac{|h_{FSO}^{(u)}(t)|^2 \mathcal{P}_{FSO}^{(u)}(t)^2}{\sigma_{FSO}^2})[Mbps] \tag{2}$$

where $B_{FSO}[MHz]$ represents the bandwidth of FSO link, $\mathcal{P}^{(u)}$ is the allowed peak power at the user u , σ_{FSO}^2 is the noise variance for FSO, and k is expressed as

$$k = \begin{cases} \frac{e^{2\alpha\mu^*}}{2\pi e}(\frac{1-e^{-\mu^*}}{\mu^*})^2 & 0 < \alpha < \frac{1}{2}, \\ \frac{1}{2\pi e} & \frac{1}{2} < \alpha < 1 \end{cases} \tag{3}$$

μ^* is the unique solution to $\alpha = \frac{1}{\mu^*} - \frac{e^{-\mu^*}}{1-e^{-\mu^*}}$ where $\alpha = \frac{\mathcal{E}}{\mathcal{P}}$ is the ratio between the allowed average power $\mathcal{E}$ and the allowed peak power $\mathcal{P}$.

C. RF Channel

The channel gain of the RF link between UAV and user u at timeslot t can be expressed as [14]

$$h_{RF}^{(u)}(t) = \sqrt{\tau_{RF}^{(u)}(t)} \cdot \tilde{h}_{RF}^{(u)}(t) \tag{4}$$

where $\tau_{RF}^{(u)}(t)$ and $\tilde{h}_{RF}^{(u)}(t)$ denote the effect of large-scale fading and the effect of small-scale fading with $\mathbb{E}|\tilde{h}_{RF}^{(u)}(t)|^2 = 1$, respectively. With the probilities of line-of-sight (LoS) and non-line-of-sight (NLoS) links, $\tau_{RF}^{(u)}(t)$ is expressed as [15]

$$\tau_{RF}^{(u)}(t) = \begin{cases} \beta_0 l^{(u)}(t)^{-\tilde{\alpha}}, & \text{LoS Link} \\ \kappa\beta_0 l^{(u)}(t)^{-\tilde{\alpha}} & \text{NLoS Link} \end{cases} \tag{5}$$

where β_0 represents the received power at the reference distance $d_0 = 1[m]$, $\tilde{\alpha}$ is the path loss exponent, and $\kappa < 1$ is an additional attenuation factor due to the NLoS link. The LoS probility between UAV and user u at timeslot t is modeled as [16].

$$P_{LoS}^{(u)}(t) = \frac{1}{1 + C \cdot \exp(-D\left[\theta^{(u)}(t) - C\right])} \tag{6}$$

where C and D are the parameters depend on the propagation environment. and $\theta^{(u)}(t) = \frac{180}{\pi} sin^{-1}(\frac{H_S}{l^{(u)}(t)})$ is the elevation angle in degree [16] Therefore, we can obtain the channel gain between UAV and user u at timeslot t.

$$\begin{aligned} h_{RF}(t) &= \sqrt{\beta_0 \hat{P}(t) l_{RF}^{-\tilde{\alpha}}(t)} \cdot \tilde{h}_r(t) \\ \hat{P}(t) &= P_{LoS}(t) + \kappa(1 - P_{LoS}(t)) \end{aligned} \tag{7}$$

At timeslot t, the achievable rate between UAV and user u can be expressed as

$$R_{RF}^{(u)}(t) = B_{RF} \cdot \log_2(1 + \frac{|h_{RF}^{(u)}(t)|^2 \mathcal{P}_{RF}(t)}{\sigma_{RF}^2})[Mbps] \tag{8}$$

where $B_{RF}[MHz]$ represents the bandwidth of the RF link, $\mathcal{P}_{RF}(t)$ represents the transmit power at timeslot t, and σ_{RF}^2 is the noise variance for RF.

III. Proximal Trajectory Optimizations for UAV

A. Problem Formulation

Ground multi-homing users can obtain data from both the RF and FSO transmitters on the UAV. Therefore, the total achievable rate at user u can be expressed as

$$R^{(u)}(t) = R_{FSO}^{(u)}(t) + R_{RF}^{(u)}(t) \tag{9}$$

Considering the practical situation, the UAV can directly decide the next acceleration based on the current flight state, not the next moving position. Therefore we assume that the movement decision of the UAV is acceleration. To simplify the problem, we assume that the UAV makes a moving decision every μ_T time until the end of time. Also, during the μ_T time, the motion of the UAV will be treated as uniformly accelerated linear motion, and when μ_T is taken small enough, the flight of the UAV can be considered as a continuous motion. Note that $(x)_t$ represents the tth decision time, $(x)(t)$ represents the timeslot t, i.e. $(x)_{t+1} = (x)(t + \mu_T)$.

Our goal is to guarantee the total achievable rate at any user u during a period of time T. $R^{(u)}(t)$ should be maintained at the target rate R_{targ}. Moreover, in order to use the FSO link as much as possible, we assume that a fixed power use principle is used to provide communication services to any user u, i.e., in the case of known CSI, the UAV will first provide FSO services at the target rate within the limits of maximum power $\mathcal{P}_{FSO}$, and will use the RF link within the limits of maximum power $\mathcal{P}_{RF}$ as a supplement when the FSO link is insufficient to provide the target rate.

Therefore, what the UAV has to do is to use its high flexibility to avoid a large number of FSO communication outages and to use the RF link as a supplement in case of an outage which need to ensure that the distance $l^{(u)}(t)$ is within the effective range for rf links. As a result, the problem can be formulated as:

$$\begin{aligned}
\min_{\{\mathbf{a}_{(S,t)}\}} \quad & \sum_{t=0}^{T}\frac{1}{U}\sum_{u=1}^{U}|R_t^{(u)} - R_{targ}| \\
\text{s.t.} \quad & \mathbf{v}_{(S,t+1)} = \mathbf{v}_{(S,t)} + \mathbf{a}_{(S,t)}\mu_T, && (10)\\
& \mathbf{q}_{(S,t+1)} = \mathbf{q}_{(S,t)} + \mathbf{v}_{(S,t)}\mu_T + \frac{1}{2}\mathbf{a}_{(S,t)}\mu_T^2, && (11)\\
& \|\mathbf{a}_{(S,t)}\| \le A_{(S,max)}, && (12)\\
& \|\mathbf{v}_{(S,t)}\| \le V_{(S,max)}, && (13)\\
& t = 0, 1, 2, \ldots, T && (14)
\end{aligned}$$

Where the constraints (10) and (11) describe the relationship between the position, velocity and acceleration of the UAV. Moreover, considering that the UAV cannot provide infinite power, we assume that the UAV flies under the limits of maximum velocity $A_{(S,max)}$ and maximum acceleration $A_{(S,max)}$, as the constraints (13) and (12). Finally we assume that the UAV works over a period of time. As in the constraints (14), $t = 0$ represents the start of time and $t = T$ represents the end.

B. Definition of State, Action and Reward

State Space: As the constraints (10) and (11), the velocity $\mathbf{v}_S(t)$ and coordinates $\mathbf{q}_S(T)$ of the UAV will change during the flight, which is very important for UAV movement decision. Considering that the UAV flies at a fixed altitude, we only need to consider the information about the motion of the UAV in the x-y plane at the altitude of H_S. Moreover, our goal is to guarantee QoS at user u, so the UAV can make better movement decisions only if it knows the communication state information with user u, i.e., the UAV needs to know the achievable rate $R^{(u)}(t)$ and the distance $l^{(u)}(t)$ to user u at timeslot t. Therefore, the state space can be expressed as

$$\begin{aligned}
\mathcal{S}(t) = [&\theta_{\mathbf{v}}(t), \|\mathbf{v}(t)\|, x_S(t), y_S(t),\\
&l^{(1)}(t), l^{(2)}(t), \ldots, l^{(U)}(t),\\
&R^{(1)}(t), R^{(2)}(t), \ldots, R^{(U)}(t)]
\end{aligned} \tag{15}$$

Where $\theta_{\mathbf{v}}(t) \in [-\pi, \pi]$ represents the angle between the direction of velocity of the UAV in the x-y plane and the positive direction of the prescribed x-axis, which can be expressed as

$$\theta_{\mathbf{v}}(t) = \begin{cases} \cos^{-1}(\frac{v_x(t)}{\|\mathbf{v}(t)\|}) & v_y(t) \ge 0 \\ -\cos^{-1}(\frac{v_x(t)}{\|\mathbf{v}(t)\|}) & v_y(t) < 0 \end{cases} \tag{16}$$

where $v_x(t), v_y(t)$ represent the velocity of the UAV in the x-axis and y-axis directions at timeslot t, respectively.

Action Space: The UAV changes its position by providing power, and since the mass of each UAV is variable, the amount of power it can provide and the amount of drag it is subjected to are different. Therefore, the action space is defined as the acceleration $\mathbf{a}(t)$ obtained by the UAV after applying power, i.e., the action space is expressed as

$$a_t = [\theta_{\mathbf{a}}(t), \|\mathbf{a}(t)\|] \tag{17}$$

where $\theta_{\mathbf{a}}(t)$ represents the angle between the current velocity direction of the UAV ($\theta_{\mathbf{v}}(t)$) and the acceleration direction. Like $\mathbf{v}(t)$, the acceleration of the UAV only has components in the x-axis and y-axis directions, i.e., $\mathbf{a}(t) = [a_x(t), a_y(t)]$.

Reward: In order to make the achievable rate $R^{(u)}(t)$ at user u as close as possible to the target rate R_{targ}. We define the reward obtained by the agent(UAV) at timeslot t, as

$$r(t) = \exp(-\omega\frac{1}{U}\sum_{u=1}^{U}|\frac{R^{(u)}(t) - R_{targ}}{R_{targ}}|) - 1 \tag{18}$$

where $\omega \in (0, 2)$, the larger the ω, the more sensitive the reward is to changes in $R^{(u)}(t)$.

C. Deep Reinforcement Learning

In this letter the UAV works to optimize each decision process over a fixed period of time, which can be considered as a Markov Decision Process (MDP). Proximal Policy Optimization (PPO) [17] is a robust and hyperparameter-insensitive on-policy algorithm. There are two neural networks in PPO, one is the policy network with parameters ϕ, which takes the state information of the UAV as input and outputs the policy distribution about the action space, i.e., it is used to fit the policy function $\pi_\theta(a_t|s_t)$; The other one is the value network with parameters θ, which takes the state information of the UAV as input and outputs the value V_t corresponding to the state s_t, i.e., it is used to fit the state-value function $V_\phi(s_t)$.

Algorithm 1 Trajectory Optimize With PPO

```
Input: initial policy network parameters $\theta_0$, value parameters $\phi_0$, buffer horizon $\mathcal{T}$.
while $k \le$ epoch do
    Reset the environment and generate the users' trajectories randomly.
    Get the initial state $s_0, r_0$.
    for $t = 0, 1, \ldots, T$ do
        Sample action $a_t$ by $\pi_{\theta_k}(s_t)$ and exceute it.
        if $\|\mathbf{v}_t\| \ge V_{S,max}$ then
            Fly at maximum speed $V_{max}$ in the previous direction.
        end if
        Collect $s_{t+1}, r_{t+1}, a_t$.
        $s_t \leftarrow s_{t+1}$
    end for
    Compute advantage estimates $\hat{A}_t$ and rewards-to-go $\hat{R}_t$, $t \in [0, T]$, and store them in the buffer.
    if buffer is filled then
        Update the parameters $\phi_k$ by minimizing (19) via stochastic gradient descent with Adam.
        Update the parameters $\theta_k$ by maximizing (20) via stochastic gradient ascent with Adam.
        Empty the buffer and $k = k + 1$.
    end if
end while
```

The loss of the value network can be expressed as

$$L_t^V(\phi) = \mathbb{E}_t\left[\left(V_\phi(s_t) - \hat{R}_t\right)^2\right] \tag{19}$$

TABLE I
CHANNEL PARAMETERS

Symbol	V	$\mathcal{P}^{u}_{FSO,max}$	σ_{FSO}	B_{FSO}	α	$\mathcal{P}_{RF,max}$	$\tilde{\alpha}$	κ	C	D	β_0/σ^2_{RF}	B_{RF}
Value	0.5 km	40 mW	-20 dBm	5 MHz	0.25	100 mW	2.3	0.2	10	0.6	37.5 dB	5 MHz

where $\hat{R}_t = r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_{T+1})$ as the objective of the state value function $V_\phi(s_t)$, which is also known as rewards-to-go. γ is a hyperparameter called the discount factor, which is related to the estimation of future rewards by the agent.

The loss of Policy network can be expressed as

$$L_t^{P+S}(\theta) = \mathbb{E}_t \left[L_t^P(\theta) + c S_{\pi_\theta}(s_t) \right] \tag{20}$$

The first term $L_t^P(\theta)$ in (20) is used for a better policy update, which can be expressed as

$$L^P(\theta) = \min \left(\text{clip}\left(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon\right) \hat{A}_t, \rho_t(\theta) \hat{A}_t \right) \tag{21}$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio, obviously $\rho_t(\theta_{old}) = 1$. The term clip $\left(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon\right) \hat{A}_t$ makes the policy update more stable, where ε is a hyperparameter which is often set to $0.2 \sim 0.3$, $\hat{A}_t$ represents an advantage estimator. General Advantage Estimation(GAE) [18] is used to estimate the advantage function in this letter, which is expressed as

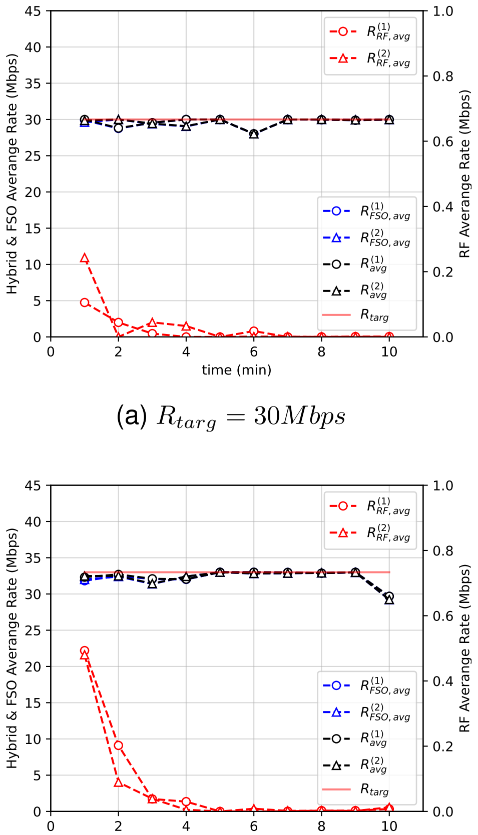
$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{22}$$

where λ is the gae factor, $\delta_t = r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t\right)$ is the TD-error. The second term $cS_{\pi_\theta}(s_t)$ in (20) is used in order to make the agent better at action exploration, which will avoid the agent from falling into a local optimum where c is coefficient, and $S_{\pi_\theta}(s_t)$ is the policy entropy.

Based on PPO, the algorithm (1) is proposed to solve the proposed model. The UAV chooses how to fly based on the policy network to avoid communication outages, and the policy network and the value network are updated regularly. The process described above is repeated until the UAV learns a satisfactory flight policy. Considering the practical situation, although the user u moves within a certain range, this letter does not limit the flight range of the UAV, which is more favorable for the UAV to explore a satisfactory policy when combined with the data-driven algorithm proposed in this letter.

IV. NUMERICAL RESULTS

The quality of the data is crucial to the training of the agent. Based on the method proposed by [19], we can obtain users' movement trajectories data to train the agent. All users move by a law in an area of size $1000m \times 1000m$, the center of which is set as the origin $[0, 0, 0]$. and the entrance to this area is $[-500, 0, 0]$. The height of the building ranges from $[60, 90]m$, with an area of $100m \times 100m$. Both the policy network and the value network are full connected networks with 2 layers of width 128, the learning rate of Adam's optimizer is 0.0002, and the other hyperparameters are set to $\mathcal{T} = 4096, \gamma = 0.97, \varepsilon = 0.25, \lambda = 0.95, c = 0.01, \omega = 1.3$. The initial position of the UAV is set to a location in the area above the entrance, i.e., $\mathbf{q}_S(0) = \left[-500 \pm \tilde{r}, \pm\tilde{r}, H_S\right], \tilde{r}[m] \in [0, 100]$. Moreover, $A_{S,max} = 5[m/s^2], V_{S,max} = 50[m/s], H_S = 100[m], \mu_t = 1[s], T = 600[s]$ is set to restrict the flight of the UAV. The number of vehicles is set to $U = 2$. The parameters related to the FSO channel and RF channel [14] are shown in Table I. In this letter different user maximum velocities $V_{D,max}$ and different target rates R_{targ} are set to verify the robustness of the algorithm 1.

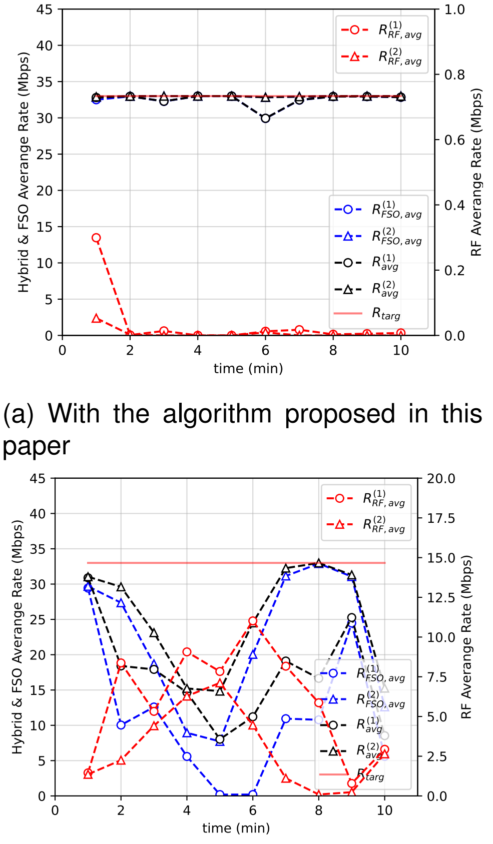


(a) $R_{targ} = 30Mbps$

(b) $R_{targ} = 33Mbps$

Fig. 1. Rate of users at maximum speed $V_{D,max} = 5m/s$

Figure 1 represents the achievable rate per minute for all users with the maximum rate $V_{D,max} = 5m/s$ limit. For example, the average rate per minute for an FSO link can be expressed as $R^{(u)}_{FSO,avg}(\tau+1) = \frac{1}{60}\sum_{t=60\tau}^{60(\tau+1)} R^{(u)}_{FSO,avg}(t)$, Where $\tau = 1, 2, \ldots, \frac{T}{60}$ represents the τth minute. From Figure 1, we can observe that the total rate at the user u is maintained around the target rate, which indicates that the algorithm proposed in this letter successfully meets the QoS requirements of mobile users. At the same time we can observe that the total rate and the rate provided by the FSO channel are approximately the same, Especially in the later time period. This indicates that UAV finds trajectories that can reach target rates using only FSO links.



(a) With the algorithm proposed in this paper

(b) $R_{targ} = 33Mbps$

Fig. 2. Rate of users under condition $V_{D,max} = 10m/s$, $R_{targ} = 33Mbps$

In addition, we set the maximum speed of the users to $V_{D,max} = 10m/s$, which corresponds to the scenario in which the vehicles needs to move fast to complete the tasks. As shown in Figure 2a, the user's QoS requirements are also satisfied, which verifies the robustness of the algorithm. Finally, due to the complexity of the environment presented in this letter, there is no certain conventional optimization algorithm as a reference.

Our common sense suggests that if the UAV always flies above the center of all mobile users, the QoS of the users may be well satisfied. However, as shown in Figure 2b, this idea does not meet the user's QoS requirements well due to the presence of tall buildings in the environment. Note that since in real scenarios the coordinates of each vehicle are not precisely available to the UAV at all times. This means that the UAV will not be able to execute this flight policy without other aids, and in this letter this algorithm is for comparison only.

V. Conclusion

In this letter, we focus on the problem of optimizing the UAV trajectories for ensuring QoS of mobile multi-homing vehicles in a complex environment. The movement of the vehicles and UAV leads to a large number of communication outages. Under the principle of fixed power use, a PPO-based reinforcement learning algorithm is proposed. By interacting with the environment, the UAV continuously improves the flight policy until it learns a better and robust flight policy. We can also see from the simulation results that the UAV's flight policy results in a significant reduction of communication outages, which successfully ensures the vehicles' QoS. The future research possiblly will focus on the communication coverage of Multi-UAVs to the ground users.

References

[1] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on uavs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

[2] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A commun. viewpoint," *IEEE Commun. Surveys Tutorials*, vol. 18, no. 4, pp. 2624–2661, 2016.

[3] P. Zhan, K. Yu, and A. L. Swindlehurst, "Wireless relay commun. with unmanned aerial vehicles: Performance and optimization," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 47, no. 3, pp. 2068–2085, 2011.

[4] H. Kaushal and G. Kaddoum, "Optical communication in space: Challenges and mitigation techniques," *IEEE Commun. Surveys Tutorials*, vol. 19, no. 1, pp. 57–96, 2017.

[5] F. Nadeem, V. Kvicera, M. S. Awan, E. Leitgeb, S. S. Muhammad, and G. Kandus, "Weather effects on hybrid fso/rf communication link," *IEEE Journal on Selected Areas in Commun.*, vol. 27, no. 9, pp. 1687–1697, 2009.

[6] J.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "Throughput maximization of mixed fso/rf uav-aided mobile relaying with a buffer," *IEEE Trans. on Wireless Commun.*, vol. 20, no. 1, pp. 683–694, 2021.

[7] Z.-Y. Wu, M. Ismail, E. Serpedin, and J. Wang, "Efficient prediction of link outage in mobile optical wireless commun." *IEEE Trans. on Wireless Commun.*, vol. 20, no. 2, pp. 882–896, 2021.

[8] M. T. Dabiri and S. M. S. Sadough, "Optimal placement of uav-assisted free-space optical communication systems with df relaying," *IEEE Commun. Letters*, vol. 24, no. 1, pp. 155–158, 2020.

[9] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid rf/vlc networks," *IEEE Wireless Commun. Letters*, vol. 9, no. 4, pp. 443–447, 2020.

[10] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-uav assisted wireless networks: A machine learning approach," *IEEE Trans. on Vehicular Technology*, vol. 68, no. 8, pp. 7957–7969, 2019.

[11] M. Najafi, H. Ajam, V. Jamali, P. D. Diamantoulakis, G. K. Karagiannidis, and R. Schober, "Statistical modeling of the fso fronthaul channel for uav-based commun." *IEEE Trans. on Commun.*, vol. 68, no. 6, pp. 3720–3736, 2020.

[12] M. A. Esmail, H. Fathallah, and M.-S. Alouini, "Outdoor fso commun. under fog: Attenuation modeling and performance evaluation," *IEEE Photonics Journal*, vol. 8, no. 4, pp. 1–22, 2016.

[13] A. Lapidoth, S. M. Moser, and M. A. Wigger, "On the capacity of free-space optical intensity channels," *IEEE Trans. on Information Theory*, vol. 55, no. 10, pp. 4449–4461, 2009.

[14] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE Trans. on Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, 2019.

[15] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device commun.: Performance and tradeoffs," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, 2016.

[16] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Commun. Letters*, vol. 3, no. 6, pp. 569–572, 2014.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[18] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2015. [Online]. Available: https://arxiv.org/abs/1506.02438

[19] Z.-Y. Wu, M. Ismail, J. Kong, E. Serpedin, and J. Wang, "Channel characterization and realization of mobile optical wireless commun." *IEEE Trans. on Commun.*, vol. 68, no. 10, pp. 6426–6439, 2020.