PAPER

UAV-BS Operation Plan Using Reinforcement Learning for Unified Communication and Positioning in GPS-Denied Environment

Gebreselassie HAILE[†] *and* **Jaesung LIM**[††a)]**,** ***Nonmembers***

SUMMARY An unmanned aerial vehicle (UAV) can be used for wireless communication and localization, among many other things. When terrestrial networks are either damaged or non-existent, and the area is GPS-denied, the UAV can be quickly deployed to provide communication and localization services to ground terminals in a specific target area. In this study, we propose an UAV operation model for unified communication and localization using reinforcement learning (UCL-RL) in a suburban environment which has no cellular communication and GPS connectivity. First, the UAV flies to the target area, moves in a circular fashion with a constant turning radius and sends navigation signals from different positions to the ground terminals. This provides a dynamic environment that includes the turning radius, the navigation signal transmission points, and the height of the unmanned aerial vehicle as well as the location of the ground terminals. The proposed model applies a reinforcement learning algorithm where the UAV continuously interacts with the environment and learns the optimal height that provides the best communication and localization services to the ground terminals. To evaluate the terminal position accuracy, position dilution of precision (PDOP) is measured, whereas the maximum allowable path loss (MAPL) is measured to evaluate the communication service. The simulation result shows that the proposed model improves the localization of the ground terminals while guaranteeing the communication service.

key words: *unmanned aerial vehicle, communication, localization, reinforcement learning, PDOP*

1. Introduction

An unmanned aerial vehicle (UAV), also known as a drone, or an airborne relay, is an aircraft controlled by a computer system through a radio communication link. UAVs have become the center of research in the industry because of their paramount importance for military and civilian applications. In the civilian application, UAVs are highly demanded for public safety and rescue operations when natural and/or man-made disasters occur. In such cases, UAVs can be quickly deployed to serve as base station in the sky (UAV-BS) and provide communication as well as localization services [1], [2].

UAV has size, weight, and power (SWaP) limitations. Therefore, it is crucial to optimize the transmission power and bandwidth of UAV-BS communications. Various research issues and challenges regarding efficient UAV operation in wireless networks were introduced in [3], [4].

Manuscript received November 2, 2023.
Manuscript revised March 9, 2024.
Manuscript publicized May 6, 2024.

[†]Dept. of AI Convergence Network, Ajou University, Suwon, South Korea.
[††]Dept. of Military Digital Convergence, Ajou University, Suwon, South Korea.
a) E-mail: jaslim@ajou.ac.kr (Corresponding author)

DOI: 10.23919/transcom.2023EBP3174

The use of UAV-BS for communication service has been studied in [5]–[10]. In [5], the authors proposed an analytical approach to optimize the altitude of low area platforms (LAPs) which can deliver essential wireless communication for public safety agencies in remote areas or during the aftermath of natural disasters. The main goal of this research work is to provide maximum radio coverage on the ground. In [6], the authors proposed an energy efficient placement of a drone base station for minimum required transmit power. They formulated the problem in a way such that it minimizes the average transmit power of the UAV-BS that serves a set of ground users. The authors in [7] proposed 3-D placement of a directional-antenna equipped UAV-BS aiming to maximize the number of flying/hovering UAV-UEs under its coverage area.

In [8], the authors applied deep reinforcement learning to make drones behave autonomously inside a suburban neighborhood which has plenty of obstacles such as trees, cables, parked cars, houses, and other moving drones. The UAV learns about the environment to avoid these stationary and moving obstacles as it navigates through the neighborhood showing how it can be used to provide communication services safely. The authors in [9] proposed a Q-learning based UAV deployment algorithm in which the UAV makes its own decision for attaining an optimal 3-D position by learning from trial and mistake for maximizing the sum mean opinion score of ground users. In [10], the authors studied how to maximize the overall data rate through an intelligent deployment of an UAV-BS in the downlink of a cellular system. They apply a reinforcement learning algorithm to avoid collision between multiple UAVs and optimize the UAV-BS positions that provide maximum sum data rate of multiple user equipment.

The use of UAV-BS for localization service in non-GPS environments has been studied in [11]–[15]. In [11], the authors proposed the use of a single UAV to localize terminals in battlefield environments as the use of global navigation satellite system (GNSS) such as GPS is prone to jamming and has weak signal reception capability. They analyzed the localization service by varying the number of received navigation signals, and the velocity of the UAV. In [12], the authors proposed a Doppler shift-based user position detection system using UAV. They measured the statistical and quantitative performance of the positioning errors of a single ground user as the UAV moves in sinusoidal curve. The ground user sends continuous signal with a fixed frequency, the UAV receives it, and relays it to the terrestrial

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

control station where the position computation takes place. In [13], the authors proposed the use of multiple UAV-BSs and additional ground references to locate a ground user. In [14], the authors proposed UAV-assisted localization of wireless devices that are in network outage and have run out of power. To localize the inactive devices, the authors used wireless power transfer (WPT) based wireless charging in which a small amount of power is transferred by the UAV to enable the target device to broadcast a beacon. The beacon from the devices contains the information about the neighboring nodes and their signal strength. The paper in [15] proposed the use of UAV-BSs for localization of a connected autonomous vehicle (CAV). They applied reinforcement learning algorithm to find the best spatial configuration of the UAV-BSs to localize the CAV in an unknown environment.

In all the above research works, the focus of the authors is either on the use of the UAV-BS for communication or localization. In this study, we propose a reinforcement learning based UAV-BS deployment scheme to provide both communication and localization services to terminals in a suburban environment without cellular communication and GPS connectivity. To do so, the UAV-BS is first deployed within the defined minimum and maximum heights. Then, the maximum allowable path loss (MAPL) of the edge terminal is computed using the air to ground (ATG) path loss model to analyze the communication service. For the localization service, the UAV-BS periodically sends navigation signals to the ground terminals. Then, each ground terminal calculates its own position using the time difference of arrival (TDOA) algorithm. We assume that a single UAV-BS moves in a circular fashion and sends navigation signals at N navigation signal transmission points (NSTP) which define the UAV positions. The NSTPs of the UAV-BS serve as reference (anchor) signals. The UAV-BS to terminal geometry impacts the position accuracy of the terminals. The metric that is normally used for measuring the terminal position accuracy is known as dilution of precision (DOP) which represents the degree to which the UAV-BS to terminal geometry dilutes the position accuracy of the terminals [16].

The main contributions of this work are:

- Defined localization and communication models using a single UAV-BS in a GPS and cellular communication denied suburban environment.
- Defined an optimization problem that integrates both the localization and communication services.
- Proposed a reinforcement learning based model to solve the optimization problem.

The rest of the paper is organized as follows. Section 2 provides the proposed system model. From the system model, the proposed unified positioning and communication schemes are described in detail. Section 3 presents the reinforcement learning based approach. Section 4 provides the simulation results, and finally, the conclusion and future work are provided in Sect. 5.

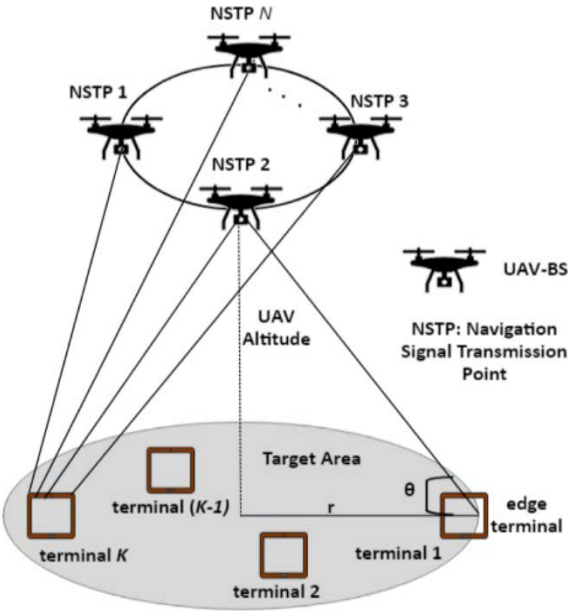

Fig. 1 Proposed system model.

2. System Model

Figure 1 shows the proposed system model. The UAV-BS is deployed to the target area where it moves in a circular route to provide the localization and communication services to the terminals. It's assumed that the UAV-BS is equipped with its own navigation equipment that provides accurate location information at any navigation signal transmission point, NSTP 1 to NSTP N as shown in Fig. 1. Also, it is assumed that the UAV platform consists of fixed-wing aircraft which can turn during flight in the sky and send downlink navigation signals to the terminals periodically. The terminals are assumed to be static (no mobility). Each navigation signal transmission point can provide different coverage area ranges as the UAV-BS moves in a circular path in the air. In this research work, however, we considered a fixed target area where the terminals are located. Only the edge terminal is located at the edge of the target area. Hence, the target area coverage and the location of the edge terminal are fixed as shown in Fig. 1.

2.1 Positioning Scheme

Let t_n be the time when the UAV-BS transmits a navigation signal from the n-th NSTP and, τ_n be the time when a ground terminal receives it.

The pseudo-range between the n-th UAV-BS NSTP and the k-th ground terminal is computed as

$$\rho_k^n = c \times (\tau_n - t_n) + \varepsilon_n \tag{1}$$

where $k = \{1, 2, \ldots, K\}$ is a ground terminal index, $n = \{1, 2, \ldots, N\}$ is a UAV-BS navigation signal transmission point index, c is the speed of light and ε_n denotes the error

that occurs during navigation signal transmission.

For any k-th ground terminal, the pseudorange difference between the n-th and the first UAV-BS navigation signal transmission points becomes:

$$\rho_k^n - \rho_k^1 = c \times ((\tau_n - \tau_1) - (t_n - t_1)) \tag{2}$$

where $\rho_k^1 = c \times (\tau_1 - t_1) + \varepsilon_1$ is the pseudo-range between the first UAV-BS navigation signal transmission point, $n = 1$, and the k-th ground terminal. The errors $\epsilon_1 = \epsilon_2 = \ldots = \epsilon_n$ are similar for the same environment, the suburban environment in our case. So, in the pseudo-range difference computations, these values cancel each other out.

In the 3-D Euclidean space orthogonal coordinate system, the pseudo-ranges ρ_k^1 and ρ_k^n are computed as follows:

$$\rho_k^1 = \left\| \mathrm{R}^1 - \mathrm{R}_k \right\| \tag{3}$$

$$\rho_k^1 = \sqrt{(x^1 - x_k)^2 + (y^1 - y_k)^2 + (z^1 - z_k)^2}$$

$$\rho_k^n = \| \mathrm{R}^n - \mathrm{R}_k \| \tag{4}$$

$$\rho_k^n = \sqrt{(x^n - x_k)^2 + (y^n - y_k)^2 + (z^n - z_k)^2}$$

where $\|\cdot\|$ represents the Euclidean norm vector, R^n is the position vector of the UAV-BS at t_n, and R_k is the position vector of the k-th terminal. Here, the UAV-BS location (x^n, y^n, z^n) is known at each navigation signal transmission time, t_n, whereas the position of the k-th terminal, $\mathrm{R}_k = (x_k, y_k, z_k)$, is unknown.

From Eq. (2), Eq. (3), and Eq. (4), the position of the k-th terminal, R_k, can be determined using the TDOA algorithm and the non-linear least squares method in the Levenberg-Marquardt algorithm [13].

From the pseudorange equation provided in Eq. (2), let's define matrices H and Z as follows:

$$H = \begin{bmatrix} \rho_k^2 & - & \rho_k^1 \\ \rho_k^3 & - & \rho_k^1 \\ \vdots & \vdots & \vdots \\ \rho_k^N & - & \rho_k^1 \end{bmatrix} \tag{5}$$

$$Z = \begin{bmatrix} -\dfrac{x^1 - x_k}{\rho_k^1} & -\dfrac{y^1 - y_k}{\rho_k^1} & -\dfrac{z^1 - z_k}{\rho_k^1} \\ -\dfrac{x^2 - x_k}{\rho_k^2} & -\dfrac{y^2 - y_k}{\rho_k^2} & -\dfrac{z^2 - z_k}{\rho_k^2} \\ \vdots & \vdots & \vdots \\ -\dfrac{x^N - x_k}{\rho_k^N} & -\dfrac{y^N - y_k}{\rho_k^N} & -\dfrac{z^N - z_k}{\rho_k^N} \end{bmatrix} \tag{6}$$

where matrix H is a column vector which consists of the pseudo-range differences between the first and the remaining $(N-1)$ UAV-BS positions for the k-th terminal, and matrix Z is a set of unit vectors of the N UAV-BS positions for the k-th terminal.

From Eq. (5) and Eq. (6), the terminal position is computed as:

$$\mathrm{R}_k = \frac{1}{2}(H^T H)^{-1} H^T Z \tag{7}$$

where R_k refers to the position of the k-th terminal, H^T is the transpose of matrix H, and $(H^T H)^{-1}$ indicates the inverse of the matrix $(H^T H)$. At least four UAV-BS positions ($N \geq 4$) are required to calculate the positions of each terminal, $\mathrm{R}_k = (x_k, y_k, z_k)$, using Eq. (7).

Position DOP (PDOP) is a metric used to measure the accuracy of terminal positioning in global navigation systems, particularly in the context of global positioning systems (GPS) and other airborne-relay based navigation systems. It is the uncertainty of 3-D parameters (latitude, longitude, and height) and depends on the geometric arrangement of the navigation signal transmission points and the altitude of UAV-BS from perspective of the ground terminals. To compute the PDOP, let's define the geometric matrix, G:

$$G = \begin{bmatrix} -\dfrac{x^1 - x_k}{\rho_k^1} & -\dfrac{y^1 - y_k}{\rho_k^1} & -\dfrac{z^1 - z_k}{\rho_k^1} & 1 \\ -\dfrac{x^2 - x_k}{\rho_k^2} & -\dfrac{y^2 - y_k}{\rho_k^2} & -\dfrac{z^2 - z_k}{\rho_k^2} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -\dfrac{x^N - x_k}{\rho_k^N} & -\dfrac{y^N - y_k}{\rho_k^N} & -\dfrac{z^N - z_k}{\rho_k^N} & 1 \end{bmatrix} \tag{8}$$

From the geometric matrix in Eq. (8), we define the covariance matrix $Q = (GTG)^{-1}$ which is a 4×4 matrix.

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} \\ Q_{41} & Q_{42} & Q_{43} & Q_{44} \end{bmatrix} \tag{9}$$

Then, the PDOP of each terminal is extracted from the covariance matrix, Q, as follows:

$$\mathrm{PDOP}_k = \sqrt{Q_{11} + Q_{22} + Q_{33}} \tag{10}$$

PDOP is a dimensionless number. A lower PDOP value indicates a more favorable geometric configuration, leading to higher position accuracy, while a higher PDOP value suggests less favorable geometry and potentially reduced accuracy.

Another metric used to evaluate the terminal position accuracy is root mean square error (RMSE). RMSE is the measure of the root of the mean of the squared errors between the predicted and true/actual terminal position values.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}} \left((\mathrm{x_k} - \hat{\mathrm{x}}_\mathrm{k})^2 + (\mathrm{y_k} - \hat{\mathrm{y}}_\mathrm{k})^2 + (\mathrm{z_k} - \hat{\mathrm{z}}_\mathrm{k})^2 \right)}{\mathrm{K}}} \tag{11}$$

where (x_k, y_k, z_k) is the true position and $(\hat{\mathrm{x}}_\mathrm{k}, \hat{\mathrm{y}}_\mathrm{k}, \hat{\mathrm{z}}_\mathrm{k})$ is the estimated position of the k-th terminal.

2.2 Communication Scheme

From the system model provided in Fig. 1, the ATG model is used to evaluate the communication service of the terminals located in the target area. From the ATG model [17], the average path loss between the UAV-BS and the ground terminal is computed as:

$$\mathrm{PL} = \mathrm{PL}_{LoS} \times p(LoS, \theta) + \mathrm{PL}_{NLoS} \times p(NLoS, \theta) \quad (12)$$

where PL_{LoS} is the line-of-sight (LOS) path loss, $p(LoS, \theta)$ is the LOS probability at elevation angle θ, PL_{NLoS} is the non-line-of-sight (NLOS) path loss, and $p(NLoS, \theta)$ is the NLOS probability at elevation angle θ.

Now, let's see how each of the parameters in the average path loss equation provided in Eq. (12) are computed. The elevation angle is defined by $\theta = \arctan\left(\frac{h}{r}\right)$ where h is the UAV height, and r is the horizontal distance between the center of the coverage area and the ground terminal.

The LOS and NLOS path loss parameters are given by:

$$\mathrm{PL}_{LoS} = \mathrm{FSPL} + \eta_{LoS} \quad (13)$$

$$\mathrm{PL}_{NLoS} = \mathrm{FSPL} + \eta_{NLoS} \quad (14)$$

where η_{LoS} and η_{NLoS} are the LOS and NLOS excessive path losses respectively. Their values are given in Table 1.

The free space path loss, FSPL, is given by:

$$FSPL = 20 \times log_{10}\left(\frac{4 \times \pi \times f \times d}{c}\right) \quad (15)$$

where f is the operating frequency, $d = \sqrt{h^2 + r^2}$ is the distance between the UAV-BS and the ground terminal, and c is the speed of light.

The LOS probability is given by:

$$p(LoS, \theta) = \frac{1}{1 + \alpha \times \exp(-\beta \times (\theta - \alpha))} \quad (16)$$

where α and β are environmental constants, whose values are shown in Table 1.

Equation (16) shows that the probability of having a line-of-sight connection between the UAV-BS and a ground terminal increases as the elevation angle increases. This decreases the mean path loss because the shadowing effect, which is the attenuation of the signal due to obstacles, decreases as the elevation angle increases. On the other hand, as the elevation angle increases, the distance between the UAV-BS and the ground terminal also increases which results in higher path loss. The NLOS probability at the given θ becomes:

$$p(NLoS, \theta) = 1 - p(LoS, \theta) \quad (17)$$

Now, we have all the parameters to compute the average path loss value, PL, at each terminal using Eq. (12).

The minimum received power at each ground terminal depends on the transmitted power of the UAV-BS, and the maximum path loss. In [18], the authors proposed a path loss and height optimization (PLaHO) model where they defined the maximum path loss for a given UAV height as follows:

$$h = \exp\left(a_0\left(\mathrm{PL}_{max} - u\right) - a_1\right) \quad (18)$$

where a_0 and a_1 are environmental constants, whose values are given in Table 1, and $u = 20 \times log_{10}\left(f \times 10^{-9}\right)$ where f is the operating frequency.

According to Eq. (18), the maximum path loss for an UAV-BS placed at 1400 m in a suburban environment is 110.4 dB. So, the maximum path loss value of 110.4 dB would be used as the threshold path loss in this study.

Table 1 Environment constants [1], [18].

Parameters	Suburban	Urban	Dense Urban
α	4.88	9.61	12.08
β	0.43	0.16	0.11
η_{LoS}	0.1	1	1.6
η_{NLoS}	21	20	23
a_0	0.1154	0.1150	0.1151
a_1	4.8008	4.5068	4.4511

2.3 Communication and Localization

By combining the communication and localization schemes, we define the following optimization problem.

$$\begin{aligned} &\text{min. } \mathrm{PDOP}_{ave} \\ &s.t.\ \mathrm{PL} \le \mathrm{PL}_{\max} \\ &\mathrm{h}_{min} \le \mathrm{h} \le \mathrm{h}_{max} \end{aligned} \quad (19)$$

where PDOP_{ave} is the average PDOP for the terminals in the target area, PL is the path loss of a terminal, and h is the UAV-BS height.

We are going to apply a reinforcement learning algorithm to solve Eq. (19) which will be described in the next sections in detail.

3. Communication and Localization Using RL

Reinforcement learning enables an agent to learn by continuously interacting with the eenvironment by trial and error using feedback from its own actions and experiences. In RL, the agent is not programmed what actions to take; instead, it learns the consequence of its actions. At each time step, the agent receives a state s_t from the state space and selects an action a_t from the set of possible actions in the action space. As a result of the action it takes, the agent gets a numerical reward r_{t+1} one time step later from the environment, and it finds itself in a new state s_{t+1} [19]. Figure 2 shows the agent-environment interaction in reinforcement learning algorithm.

Q-learning is a model-free RL algorithm which learns the value of an action in a particular state. Q-learning algorithms carry out an action multiple times and adjust the policy for optimal rewards based on the outcomes of the actions. Epsilon-Greedy action selection policy is applied

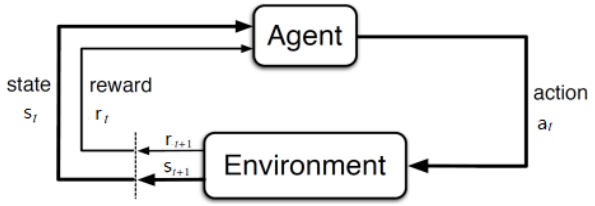

Fig. 2 Agent-Environment interaction in RL.

in the Q-learning algorithm where epsilon is a probability value that balances the exploration and exploitation of the action by the agent. Epsilon helps the agent to exploit the action with small probability of exploring.

In this paper, Q-learning algorithm is applied to solve the problem defined in Eq. (19). The goal is to apply Q-learning algorithm to acquire the minimum PDOP_{ave} value under the path loss and height constraints for unified communication and localization services using UAV-BS. For any Q-learning algorithm, the environment, agent, state, action, and reward should be defined. In this paper, these parameters are defined as follows:

Environment – An environment represents the system an agent interacts with. In this paper, the environment is a GPS and cellular communication denied suburban environment.

Agent – An agent is the entity that interacts with the environment to achieve a specific task. In this paper, the agent is the UAV-BS. So, UAV-BS and agent can be used interchangeably.

State space – The UAV-BS height (altitude) forms the state space in our model. Originally, the state space is continuous as the UAV-BS can take any value as it moves within the defined minimum height, h_{min} and the maximum height, h_{max}. This continuous UAV-BS height is then discretized to give an integer number of states. By defining Δh as the change of height after each action, the total number of states is computed as:

$$N_{states} = \left\lfloor \frac{(\text{h}_{max} - \text{h}_{min})}{\Delta\text{h}} + 1 \right\rfloor \tag{20}$$

where the floor function $\lfloor x \rfloor$ takes a real number x and gives the greatest integer less than or equal to x as an output.

In this paper, h_{min} and h_{max} are 400 m and 1400 m respectively, and Δh is 100 m. Applying these values to Eq. (20) provides $N_{states} = 11$ discrete number of states which define the state space. Figure 3 shows how each action the agent takes changes the state-space.

Action space: An action indicates what the agent (UAV-BS) does from the current state. An action space indicates the possible set of actions that the agent can take in the agent-environment interaction. In this paper, we have defined three types of actions the UAV-BS can take from the current state: An Upward Action, a Downward Action, and a Static Action, as illustrated in Fig. 3. Assuming that the current state of the agent is h, depending on the epsilon greedy policy,

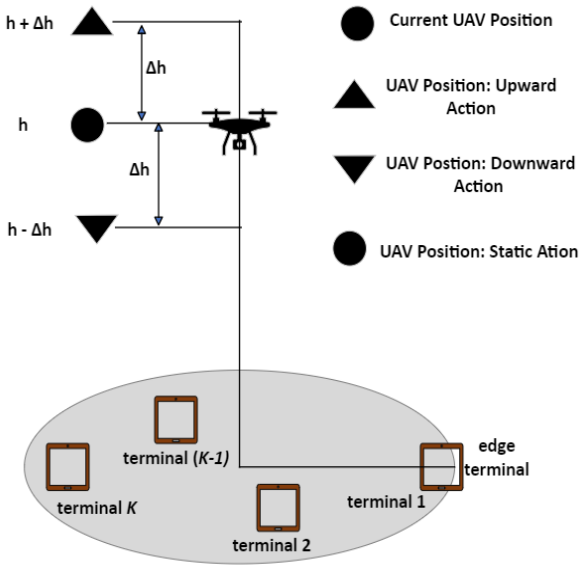

Fig. 3 Action space of the agent, UAV-BS.

the agent takes one of the three possible actions. An Upward Action takes the agent to the $(\text{h}+\Delta\text{h})$ state, a Downward Action takes it to the $(\text{h}-\Delta\text{h})$ state, and a Static Action causes the agent to remain int its current state at h.

Policy: Policy indicates how an agent chooses its actions. In Q-learning, an epsilon greedy strategy decides whether the agent should explore or exploit while interacting with the environment. The agent initially starts out by choosing a random action (exploration). As the episode progresses, the epsilon value provides a balance between exploration and exploitation. In exploitation, the agent chooses its action based on the highest Q-value from the Q-table for the given state.

Reward – A reward is a scalar value received after each action for transition to the new state. The average PDOP value is used as a reward in this paper. It measures the position accuracy of the terminals in the target area. The UAV-BS sends the navigation signals from the N UAV-BS positions of the current state, and each terminal computes the PDOP value as given in Eq. (10). When the number of terminals in the target area is more than one ($K > 1$), an average PDOP is used as the reward. The average PDOP is defined by:

$$\text{PDOP}_{ave} = \frac{\sum_{k=1}^{K} \text{PDOP}_k}{K} \tag{21}$$

The PDOP_{ave} decides the reward in the Q-learning algorithm. Low average PDOP shows good terminal position accuracy, and large average PDOP shows bad terminal position accuracy.

The Q-learning algorithm uses a Q-table which contains the state and action pair known as Q-values. At each state, the agent computes the numerical reward r_{t+1}, based on the average PDOP as follows:

$$r_{t+1} = \begin{cases} -1, & if\ PDOP_{ave}(s_{t+1}) > PDOP_{ave}(s_t) \\ 0, & if\ PDOP_{ave}(s_{t+1}) = PDOP_{ave}(s_t) \\ +1, & if\ PDOP_{ave}(s_{t+1}) < PDOP_{ave}(s_t) \end{cases} \quad (22)$$

where $PDOP_{ave}(s_{t+1})$ is the average PDOP at the next state, and $PDOP_{ave}(s_t)$ is the average PDOP at the current state.

Depending on the $PDOP_{ave}$, the UAV-BS decides r_{t+1} as shown in Eq. (22). If the average PDOP at the next state is greater than the average PDOP at the current state, the agent gets a negative reward. If the average PDOP at the next state is lower than the average PDOP at the current state, the agent receives a positive reward. If there is no change in the average PDOP values, the agent gets zero reward. The agent updates its Q-table and takes one of the 3 actions based on the epsilon greedy policy to move to the next state as shown in Fig. 3.

UAV-BS Q-table update – the UAV-BS has an action-value matrix which represents the value of being in a specific state s_t, while taking an action a_t. The UAV-BS updates the Q-value of the current state, $Q_n(s_t, a_t)$, through the Q-learning function defined by:

$$Q_n(s_t, a_t) = Q_o(s_t, a_t) + \mu \times \left(r_{t+1} + \gamma \times \max_a Q(s_{t+1}, a) - Q_o(s_t, a_t) \right)$$

which can be simplified to:

$$Q_n(s_t, a_t) = (1 - \mu) \times Q_o(s_t, a_t) + \mu \times \left(r_{t+1} + \gamma \times \max_a Q(s_{t+1}, a) \right) \quad (23)$$

where $Q_n(s_t, a_t)$ is the new Q-value of the current state, $Q_o(s_t, a_t)$ is the old Q-value of the current state, μ is the learning rate, r_{t+1} is the reward defined in Eq. (22), γ is the discount factor, and $\max_a Q(s_{t+1}, a)$ is the action that maximizes the Q-value of the next state. μ determines how much the agent adjusts its estimates based on new information obtained from the interactions with the environment. It's a value between 0 and 1. γ is a parameter that controls the importance of future rewards in the agent's decision-making process. It's a value between 0 and 1 and represents the extent to which the agent values future rewards compared to immediate rewards.

Stopping criteria: Initially, the UAV-BS is randomly located in one of the discrete states which correspond to the UAV-BS heights. Then, the interaction between the agent and the environment proceeds in sequences of steps until a stopping criterion is met. Stopping criteria decide when the agent should stop interacting with the environment. One way to define the stopping criteria is to let the agent continue until all the available states are visited. This is done to give the agent enough opportunity to interact with the environment and learn about it through exploitation and exploration following the e-greedy policy. Another way to outline the stopping criteria is to specify the number of steps in each episode. In this paper, we have defined the stopping criteria based on the number of steps. The agent runs for 200 steps and then stops. This is decided based on a repeated simulation observation where the reward does not improve when the number of steps exceeds 200.

DQN versus Q-learning: Deep Q-learning network (DQN) has become widespread in many of the RL-based research works recently. In this paper, however, Q-learning has been selected because the number of state and action spaces, (states = 11, and actions = 3), is very small and memory is not a problem. When the state and action spaces are large, using the Q-table is impractical because of memory limitations which affect the performance. In that case, DQN should be used as it addresses the memory limitation of Q-learning through *Replay Memory* technique where only limited number of state-action pairs are used instead of the whole state-action pairs.

4. Simulation Results

MATLAB is used to simulate the proposed UCL-RL model. We have developed a customized suburban environment that contains randomly generated terminals within a defined coverage area. The agent (UAV-BS) continuously interacts with the environment and learns about it using the reinforcement algorithm.

To the best of our knowledge, this is the first work that proposed the use of UAV for unified communication and localization services using reinforcement learning. To evaluate the performance of the proposed *UCL-RL* model, we have used two models for comparison. The first model is proposed in [11] which analyzed the use of single and dual UAV to localize terminals in battlefield environments. Since we are using one UAV-BS in this study, we have selected the single UAV-based localization (SUL) model of [11] for comparison. In the *SUL* model, the UAV-BS is placed at the middle of the UAV-BS height ranges which is 0.9 Km which is the average of the minimum (400 m) and maximum (1400 m) UAV-BS altitudes. The PDOP, RMSE, and PL metrics are then measured from this fixed altitude. As a second model, we have defined a *Basic* model where the UAV-BS is randomly placed within the minimum and maximum UAV-BS heights throughout all the simulation episodes. In the *Basic* model, the PDOP, RMSE, and PL metrics are computed from the random position the UAV-BS takes at each episode. For the proposed *UCL-RL* model, however, the UAV-BS learns the optimal UAV-BS height using the reinforcement learning algorithm through a continuous interaction with the environment.

Table 2 shows the simulation parameters. The simulation scenario consists of a 2.8284 km radius target area as shown in Fig. 4. Figure 4 illustrates one instance of the simulation scenario where the UAV-BS is placed at h = 1400. From this position, it moves in a circular path, generating navigation signals at each UAV position. Subsequently, the localization and communication services are measured. At another time step, the UAV-BS takes different UAV-BS heights according to the *UCL-RL* algorithm as depicted in

Table 2 Simulation parameters.

Parameters of the table	Values
UAV-BS speed	180 Km/h
UAV-BS turning radius	2000 m
UAV-BS height (Min, Max)	400 m, 1400 m
Δh	100 m
State space	11
Action space	3
Frequency	2 GHz
Learning rate	0.1
Discount factor	0.95
Number of terminals	1 and 20
Minimum received power	-80 dBm
UAV-BS transmitted power	15 W
Navigation signal transmission points	15

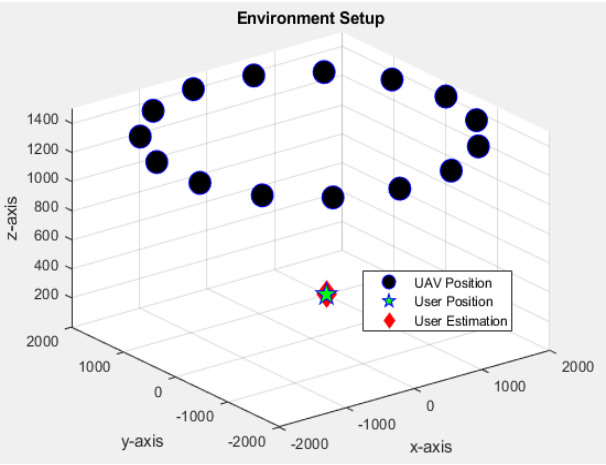

Fig. 4 Simulation scenario.

Fig. 3, to produce the required state, action, and reward. The agent continues interacting with the environment until it reaches a stopping condition.

There are two simulation scenarios: single user simulation scenario and multiple user simulation scenario. In the single user simulation scenario, one terminal ($K = 1$), which is an edge user, is used to evaluate the performance of the *SUL*, *Basic* and *UCL-RL* models. In the multiple user simulation scenario, 20 terminals ($K = 20$) are generated within the defined target area. Out of the 20 terminals, 19 terminals are randomly generated whereas 1 terminal is an edge user. The values for the learning rate $\mu = 0.1$ and the discount factor $\gamma = 0.95$ are selected because they are very common values in many of the Q-learning algorithm-based research works. The value for epsilon is initially 1 and decreases as the episode progresses to balance the exploitation and exploration strategies which are crucial to maximize the cumulative reward over time.

Figure 5 shows the PDOP simulation result for the *SUL*, *Basic* and *UCL-RL* models for the single user simulation scenario. Here, the dynamicity in the agent-environment interaction is the result of the change in altitude of the agent. The PDOP of the edge terminal is computed at each episode and serves as the reward. In Fig. 5, the PDOP for the *SUL* model doesn't vary throughout the episodes because it's measured from a fixed height. Initially, up until

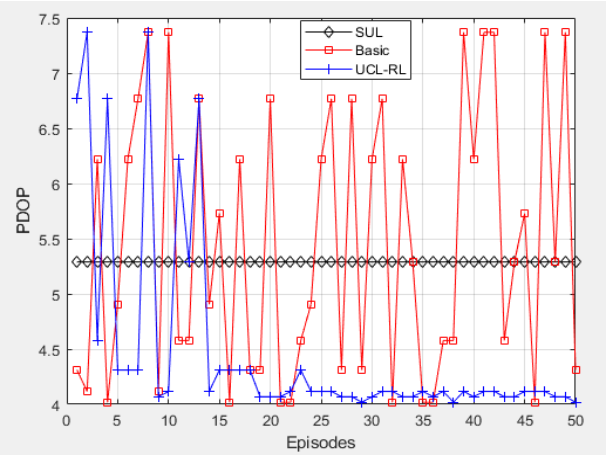

Fig. 5 PDOP comparison of SUL, Basic and UCL-RL models, K = 1.

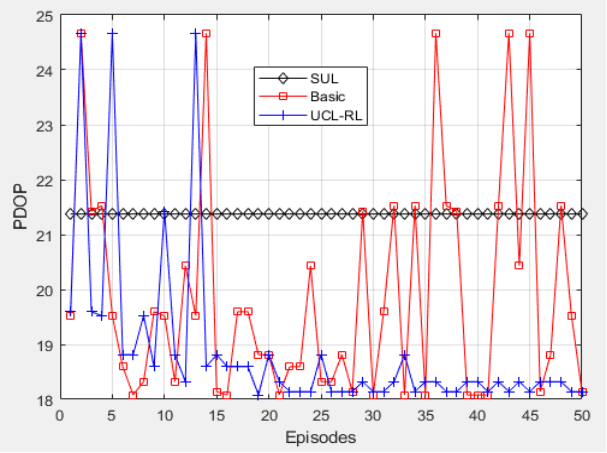

Fig. 6 $PDOP_{ave}$ comparison of SUL, Basic and UCL-RL models for K = 20.

episode 13, the proposed *UCL-RL* model has worse PDOP value than the *SUL* and *Basic* models as it has not interacted with the environment and learned the best reward yet. As the agent-environment interaction proceeds (defined by the episodes), the proposed *UCL-RL* model has resulted in an improved PDOP value compared to the *SUL and Basic* models. After 24 episodes, the *UCL-RL* model has converged to the best reward, while the *SUL* model has fixed value, and the *Basic* model has random values in every episode.

Figure 6 shows the average PDOP simulation result for the *SUL, Basic* and *UCL-RL* models for the multiple user simulation scenario. The average PDOP for the terminals is computed at each episode and serves as the reward. Initially, like in the single user scenario, the average PDOP value for the *UCL-RL* model is worse than the PDOP value of the *SUL* and *Basic* models. As the episode progresses, however, the average PDOP value for the *UCL-RL* model has improved. Starting from episode 14, the *UCL-RL* model provides better average PDOP compared to the *SUL* and *Basic* models as shown in Fig. 6.

The PDOP range in the single-user simulation scenario

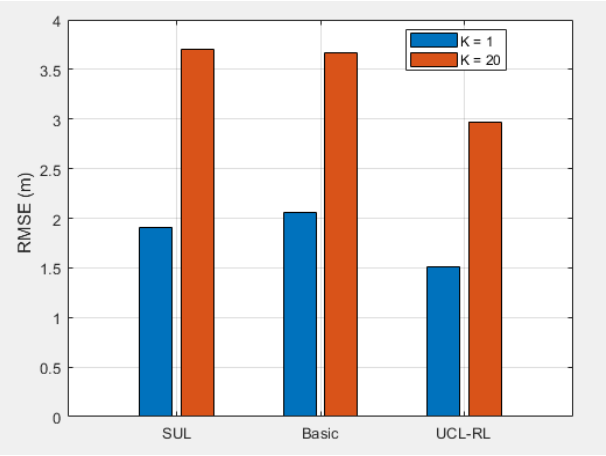

Fig. 7 RMSE comparison of SUL, Basic and UCL-RL models for K = 1 and K = 20 averaged over 50 episodes.

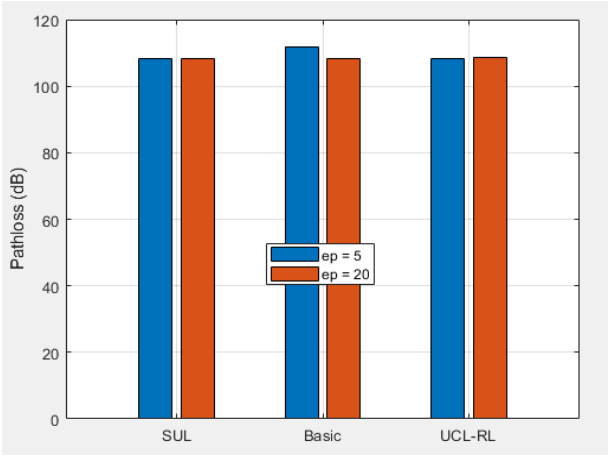

Fig. 8 MAPL comparison of SUL, Basic and UCL-RL models at episodes 5 and 20.

(PDOP = 4 to 7.5 in Fig. 5) is lower than the PDOP range in the multiple user simulation scenario ($PDOP_{ave}$ = 18 to 25 in Fig. 6). This variation in PDOP value in the two simulation scenarios come from the different UAV-BS to terminal geometries. The UAV-BS to terminal geometry is an important factor that affects the PDOP value. It describes the geometry of the navigation signal transmission points (NSTP) of the UAV-BS from the ground terminals perspective. In the single-user simulation scenario, the geometry of the terminal depends on the NSTPs and height of the UAV-BS only. This provides better UAV-BS to terminal geometry which corresponds to the lower PDOP range. In the multiple user simulation scenario, however, there are many UAV-BS to terminal geometries, one for each terminal, which depend on the NSTPs and the height of the UAV-BS as well as the positions of the terminals. These multiple geometries result in a large average PDOP value at each episode. That is the reason why the PDOP range in the multiple-user simulation scenario is larger than the PDOP range in the single-user simulation scenario. In both simulation scenarios, the proposed *UCL-RL* model provides better PDOP value as the episode increases compared to the *SUL and Basic* models as illustrated in Fig. 5 and Fig. 6.

Another way to measure the accuracy of the terminal positioning is to apply root mean square error (RSME). Figure 7 shows the RMSE for the *SUL, Basic and UCL-RL* models for the two simulation scenarios averaged over 50 episodes. For $K = 1$, the RMSE values for the *SUL, Basic* and *UCL-RL* models are 1.91 m, 2.06 m and 1.51 m respectively. For $K = 20$, the RMSE values for the *SUL, Basic* and *UCL-RL* models are 3.70, 3.67 m and 2.97 m respectively. To compute the estimated positions of the terminals, we apply a non-linear least square method using the Levenberg-Marquardt algorithm [13]. The algorithm iteratively adjusts the estimated terminal positions, leading to reduced errors and minimized RMSE values. Consequently, despite the high PDOP values shown in Fig. 6, the RMSE values in Fig. 7 are small. The improvement is due to the effectiveness of the Levenberg-Marquardt algorithm in minimizing errors.

In both simulation scenarios, the *UCL-RL* model has provided smaller RMSE values compared to the *SUL* and *Basic* models. This shows the proposed *UCL-RL* model provides better terminal position accuracy as the agent learns the best parameters that minimize the positioning error of the terminals.

To evaluate the communication, the maximum allowable path loss (MAPL) metric is measured. The purpose of this evaluation is to show the path loss of the proposed *UCL-RL* model lies within the given path loss range defined by the maximum path loss, PL_{max}, which is the threshold path loss. According to the PLaHO model defined in [18] and given in Eq. (18), the MAPL for an UAV-BS to ground terminal communication in suburban environment is 110.4 dB which is the threshold path loss value. Figure 8 shows the path loss for the *SUL*, *Basic* and *UCL-RL* models at two episodes (ep = 5, and 20) for the single user simulation scenario ($K = 1$) by considering the MAPL for the suburban environment. The values of ep = 5 and ep = 20 are carefully selected to demonstrate learning properties of the agent. The lower episode, ep = 5, represents the learning process at the beginning of the learning. At this episode, the agent has had limited interaction with the environment. At ep = 20, the agent demonstrates substantial learning about the environment due to increased interaction. The characteristics of the other episodes are similar to these two episodes.

At ep = 5, the path loss values for the *SUL, Basic* and *UCL-RL* models are 108.15 dB, 111.94 dB and 108.15 dB respectively. At ep = 20, the path loss values are 108.15 dB, 108.12 dB, 108.74 dB for the *SUL, Basic* and *UCL-RL* models respectively. The proposed *UCL-RL* model has produced path loss value below the MAPL as the episode increases from 5 to 20 as shown in Fig. 8. There is a slight increase in the path loss value for the *UCL-RL* model when the episode increases from 5 to 20, but it is still less than the threshold path loss value (110.4 dB) which shows that the proposed

UCL-RL model maintains the communication service. This proves that the proposed *UCL-RL* model provides improved localization service while enabling communication service to the terminals in the target area when compared to the *SUL* and *Basic* Models.

5. Conclusion

This paper proposed the use of a single UAV-BS to provide unified communication and localization services in suburban environment with no cellular and GPS connectivity by applying reinforcement learning. The UAV-BS is flown to the target area and deployed within the minimum and maximum heights where it moves in a circular path to send navigation signals to the terminals in the target area. The combination of the UAV-BS turning radius, navigation signal transmission points, UAV-BS height, and the position of the ground terminals provides a dynamic environment. The UAV-BS interacts with the environment and learns the average PDOP value as a reward through the Q-learning algorithm. The path loss of an edge terminal is also measured to assess the communication service. Simulation results have shown that the proposed model provides improved terminal positioning accuracy while guaranteeing communication service.

In this work, the UAV-BS turning radius is constant. In our next work, we will design the problem by varying the turning radius and assess how it affects the localization and communication capabilities. In addition to that, we will expand the UAV-BS height range to increase the state space and then apply other reinforcement learning algorithms, like DQN, to evaluate the performance.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2007112).

References

[1] M. Erdelj, E. Natalizio, K.R. Chowdhury, and I.F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," IEEE Pervasive Comput., vol.16, no.1, pp.24–32, Jan.-March 2017.
[2] Y. Zeng, R. Zhang, and T.J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," IEEE Commun. Mag., vol.54, no.5, pp.36–42, May 2016.
[3] 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 Tuts., vol.21, no.3, pp.2334–2360, 2019.
[4] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," IEEE Commun. Survey Tuts., vol.18, no.2, pp.1123–1152, 2016.
[5] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," IEEE Wireless Commun. Lett., vol.3, no.6, pp.569–572, Dec. 2014.
[6] L. Wang, B. Hu, and S. Chen, "Energy efficient placement of a drone base station for minimum required transmit power," IEEE Wireless Commun. Lett., vol.9, no.12, pp.2010–2014, Dec. 2020.
[7] N. Cherif, W. Jaafar, H. Yanikomeroglu, and A. Yongacoglu, "On the optimal 3D placement of a UAV base station for maximal coverage of UAV users," IEEE Global Communications Conference, Taipei, Taiwan, pp.1–6, 2020.
[8] E. Çetin, C. Barrado, G. Muñoz, M. Macias, and E. Pastor, "Drone navigation and avoidance of obstacles through deep reinforcement learning," 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), San Diego, CA, USA, pp.1–7, 2019.
[9] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," IEEE Trans. Veh. Technol., vol.68, no.8, pp.8036–8049, Aug. 2019.
[10] S.P. Gopi and M. Magarini, "Reinforcement learning aided UAV base station location optimization for rate maximization," Electronics, vol.10, no.23, p.2953, 2021.
[11] D.-H. Kim, K. Lee, M.-Y. Park, and J. Lim, "UAV-based localization scheme for battlefield environments," MILCOM 2013 - 2013 IEEE Military Communications Conference, San Diego, CA, USA, pp.562–567, 2013.
[12] H. Ishikawa, Y. Horoikawa, and H. Shinonaga, "Maximum positioning error estimation method for detecting user positions with unmanned aerial vehicle based on Doppler shifts," IEICE Trans. Commun., vol.E103-B, no.10, pp.1069–1077, Oct. 2020.
[13] K. Lee, H. Noh, and J. Lim, "Airborne relay-based regional positioning system," Sensors, vol.15, no.6, pp.12682–12699, 2015.
[14] M. Atif, R. Ahmad, W. Ahmad, L. Zhao, and J.J.P.C. Rodrigues, "UAV-assisted wireless localization for search and rescue," IEEE Syst. J., vol.15, no.3, pp.3261–3272, Sept. 2021.
[15] E. Testi, E. Favarelli, and A. Giorgetti, "Reinforcement learning for connected autonomous vehicle localization via UAVs," 2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Trento, Italy, pp.13–17, 2020.
[16] R.B. Langley, "Dilution of precision," GPS World, vol.10, no.5, pp.52–59, 1999.
[17] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," 2014 IEEE global Communications Conference, Austin, TX, USA, pp.2898–2904, 2014.
[18] I. Mohammed, I.B. Collings, and S.V. Hanly, "A new connectivity model for unmanned aerial vehicle communications and flying height optimization," Transactions on Emerging Telecommunications Technologies, vol.34, no.6, e4767, 2023.
[19] R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction, 2nd ed., The MIT Press, 2018.

Gebreselassie Haile received B.S. in Electronics & Communications Engineering from Mekelle Institute of Technology, Ethiopia, in 2007 and M.S. in computer Engineering from Ajou University, Korea, in 2013. During 2007–2020, he was with Information Network Security Administration of Ethiopia where he was involved in Telecom & Network signal Analysis, Speech Compression, and Wireless Network Audit. Starting from March 2020, he is a Ph.D. student at Ajou University, Korea, in the department of Artificial Intelligence Convergence Network. His research interests are Resource Management in Wireless Networks, UAV for Communication & Localization, and Machine Learning for Wireless Networks.

Jaesung Lim received B.S. in electronic engineering from Ajou University, Korea, in 1983, and M.S. and Ph.D. in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), in 1985 and 1994, respectively. In 1985, he started as a researcher at Daewoo Telecommunication. In April 1988, he joined the institute of DigiCom, and was engaged in research and development of data modem, radar signal processing and packet data systems. From 1995 to 1997, he served as a senior engineer in the Central Research and Development Center of SK Telecom. Since March 1998, he has been with Ajou University where he is a professor of the department of military digital convergence teaching and doing research in the areas of wireless, mobile, and tactical communications and networks.