IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 June 2022, accepted 27 June 2022, date of publication 6 July 2022, date of current version 26 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188792

RESEARCH ARTICLE

Reinforcement Learning-Based Visible Light Positioning and Handover Scheme With Stereo-Camera

BO ZHANG, DAHAI HAN, MIN ZHANG, LIQIANG WANG, AND XIAOYUN LI

State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Dahai Han (dahaihan@gmail.com)

This work was supported in part by the Shenzhen Science and Technology Innovation Commission, China, under Project 2021Szvup009; and in part by the Zhejiang Laboratory International Talent Fund for Young Professionals.

ABSTRACT Aimed at solving the problems of frequent handover and large overheads for positioning in 6G, this paper proposes a reinforcement learning (RL)-based visible light positioning (VLP) handover scheme by means of a stereo camera system in which a lower handover rate and higher positioning accuracy were achieved simultaneously. Because of the randomness of the distribution location of indoor light sources and obstacles, even if the parameters of light sources and receivers are consistent, users cannot rely on unitary and invariable parameters to determine whether to change the access point (AP) in the process of moving, especially in some special locations. The proposed scheme, which decomposed the user's moving track at different speeds in a VLP system to optimize selecting the AP by using RL at each step, was exhibited and tested. Experimental results show that the proposed scheme achieved millimeter-level positioning accuracy, improved the normalized reward by over 40% and reduced the handover rate by 87% and 78% compared to the immediate handover (IHO) and dwell handover (DHO) methods concurrently.

INDEX TERMS 6G, location-based services, visible light positioning, reinforcement learning, handover.

I. INTRODUCTION

A visible light communication (VLC) base station is deployed in the indoor base station of a 6G network [1], [2], providing users with an extremely high rate of data transmission and high-accuracy positioning conditions for indoor location services. Nevertheless, ultra-dense deployment of access points (APs) is required to achieve effective visible light coverage in large indoor spaces.

Indoor positioning technology based on VLC (also known as visible light positioning, VLP) employs LED lamps to transmit ID information, and the positioning terminal performs indoor positioning through a photodetector (PD) or camera. Owing to its advantages of high positioning accuracy and low cost, this technique has been widely studied [3]. However, most VLP systems require at least three VLP LED lamps for accurate position calculation. Therefore, this limitation may restrict the application of VLPs in practical scenarios. Zhang *et al.* [4] proposed a single-LED-based VLP system based on circular projection, with a red marker providing additional information regarding direction. Nevertheless, the computational cost of circular projection was high for a positioning accuracy of only 17.52 cm. Hao *et al.* [5] proposed using the calibrated inertial sensor of smartphones to help single-LED position calculations with projective geometry. Their system achieved a positioning accuracy of 16 cm. Han *et al.* [6] proposed a plane intersection-line scheme based on the geometric features of LED projection to improve positioning accuracy regardless of whether the receiver was horizontal or tilted. This approach achieved 3D average positioning errors of 5.58 cm.

In contrast, machine learning methods can be introduced into VLP to improve positioning performance. These methods may significantly improve the performance. In recent years, academia and industry have considered machine learning as an indispensable auxiliary tool for VLP systems [7], [8]. Lin *et al.* [9] proposed a VLP system to estimate the receiver's position approximately based on decoded block

The associate editor coordinating the review of this manuscript and approving it for publication was Khaled Rabie.

coordinates and then obtain the position precisely by using a typical backpropagation ANN. The experiments showed that the proposed scheme provided a mean positioning error of 1.49 cm in 2D. Wu *et al.* [10] proposed a received-signal-strength (RSS) based VLP system using sigmoid function data preprocessing (SFDP) method, and apply it to kernel ridge regression machine learning (KRRML) algorithm. The experimental average positioning error of about 2 cm in both horizontal and vertical directions. Du *et al.* [11] proposed a 3D indoor VLP system assisted by deep learning techniques to learn from a series of samples labeled with their 3D locations to estimate a new sample. Furthermore, a new method for relying on offline preparation was adopted to minimize the workload of VLP system deployment. However, these studies have high computational complexity and long positioning time. They are only suitable for offline processing and need the support of high-performance computing devices.

In addition to accuracy and robustness, the real-time performance of VLP is a major concern for its application. Li *et al.* [12] proposed an unbalanced single-LED VLP algorithm to increase the positioning accuracy and reduce the computational complexity. Simultaneously, a fast beacon searching algorithm was proposed to reduce the processing time for each captured image. Consequently, the average positioning time was reduced to 60 ms on a low-end embedded platform. In [13], Guan *et al.* proposed a triple-light positioning algorithm that requires solving binary linear equations. In particular, the LED-ID detection and recognition problems were processed by a machine learning algorithm, with a computational time of 65.50 ms. On this basis, in [14], an extended Kalman filter (EKF) was implemented for real-time 3D pose estimation (position and orientation) by fusing the relative pose measurements from the IMU with the absolute pose from the VLP measurement. Consequently, the average calculation time was approximately 33 ms on a low-cost embedded platform. Furthermore, the optical flow method used in [15] to track the LED substantially increased the computational cost and running time of the VLP system. Xie *et al.* [16] proposed an algorithm using the mean shift algorithm to track and locate LEDs in the image sequence and an unscented Kalman filter algorithm to predict the possible position of the LED in the next frame. As a result, the computational cost of running the complete algorithm was reduced, and the average processing time was 24.93 ms. The above study suggests that real-time performance can only be guaranteed when users access the same AP for continuous positioning. However, when users enter overlapping areas covered by different APs and need to hand over, the user devices demand extra computational operations because of re-decoding and identifying the target AP. Commercial cameras generally have a frame rate of 30 Hz, which leaves only 30 ms to process each frame of the image. The performance of image processing equipment at mobile receivers is limited, so the real-time detection, decoding and position estimation of target APs cannot be guaranteed due to the pressure of computing time. This attribute increases the delay, decreases positioning accuracy, and causes other problems, reducing the quality of the user experience.

To avoid frequent handovers in an ultra-dense network, the concept of handover skipping was introduced. The existing research mainly focused on handover mechanisms based on optimizing the QoS in the field of optical wireless communication. Wu *et al.* [17] proposed an RSRP-based handover skipping method that uses a weighted average of the value of RSRP and its rate of change. The authors determined whether to skip a certain AP. As the rate of change in the RSRP is related to the user velocity, this novel method is sensitive to velocity. Moreover, AI methods have been considered an important part of 6G networking [18] and have attracted considerable attention in optical wireless communication handover mechanisms. In particular, reinforcement learning (RL) algorithms have been applied in AP selection and handover mechanisms. Wang *et al.* proposed an adaptive optical wireless communication handover mechanism for 6G networks with hybrid architectures by employing RL to optimize the waiting time before the handover process and reduce the interruption time incurred by the handovers [19]. Seyed *et al.* [20] employed multiple millimeter-wave radio frequency (RF) transmitters as complementary APs for an optical wireless communication system in the case of blockage. The study proposes a convolutional neural network (CNN)-based algorithm consisting of offline and online modes to dynamically tune the waiting time (WT) and handover margin (HOM) based on alternating the SNR values related to the serving of cooperating VLC transmitters in consecutive time slots. These studies mainly focus on enhancing the quality of communication service when users are moving by optimizing the WT and HOM. Nevertheless, in visible light indoor positioning, the influence of changing AP on positioning accuracy when users move across regions remains to be studied.

In addition to the handover problem of the VLC+RF hybrid technology in a heterogeneous networking environment, the positioning handover problem based on VLC should also be considered. Nevertheless, the above studies focused on the boundaries where switching occurred and optimizing HOM and WT, neglecting the balance between positioning accuracy and handover rate under specific user trajectories.

Because the single LED group plays the role of a AP in a VLP with relatively limited coverage, an intensive deployment of APs is required in a 6G network [21], as shown in Fig. 1. Due to the randomness of the distribution location of indoor light sources and obstacles, as users move, the target APs are frequently handover to maintain connectivity, which leads to more operations on target identification and ID decoding. The higher the handover frequency is, the higher the system overhead, and the poorer the service quality. However, the ping-pong effect cannot be avoided by the immediate handover scheme, and the special path cannot be optimized by the dwell handover scheme. Therefore, this scenario requires an AP adaptive handover mechanism to

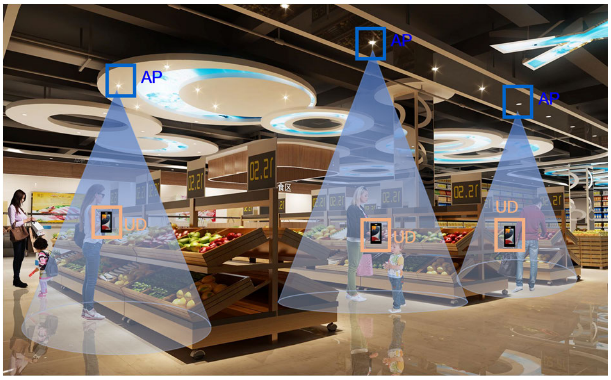

FIGURE 1. Indoor VLP scene with a random deployment of light sources and obstacles.

balance handover efficiency and positioning accuracy and reduce the impact of the AP handover on location services. In this study, we propose an RL-based VLP handover scheme with a stereo camera. Regarding the positioning mode, due to the sensitivity of the PD to the beam direction, the mobility of the positioning terminal is severely limited; therefore, a positioning mode based on a stereo camera is adopted in this study. Regarding the positioning handover strategy, a multi-target cross-region positioning handover scheme based on a stereo camera is proposed to balance the positioning accuracy and handover rate. Based on a previous study [22], the positioning accuracy of users in the overlapping regions of adjacent APs is evaluated, and the AP selection problem is solved using the RL algorithm. The performance of the proposed method is analyzed for the handover rate and positioning errors based on developed mathematical expressions. Compared to optimization with the traditional case, the simulation results show that optimization with machine learning is advantageous in terms of the handover rate and normalized reward of positioning accuracy.

The remainder of this paper is organized as follows. Section II describes the indoor VLP network system. Subsequently, the handover scheme is described in Section III. Then, a theoretical analysis using mathematical expressions is presented in Section IV. The simulation results are presented in Section V. Finally, Section VI summarizes the study.

II. VLP SYSTEM MODEL CONSIDERING HANDOVER

In a previous work [22], we introduced a positioning scheme based on a stereo camera. An LED group comprising at least three LEDs transmits their 3D coordinate information, encoded by flicker mitigation, dimming, and expansion code. As the LEDs are grouped, we assume that each LED is placed at one of the three vertices of an equilateral right triangle, lying in the same plane (*i.e.*, all have the same z-coordinate value). At the receiver, a stereo camera with a given field of view Φ was used to receive the light signals from the three LEDs spatially separated using two separate lenses and capture the finger images based on the rolling-shutter effect (RSE). Subsequently, the LED IDs, which were coded by Flicker-Free coding with extending support [22], were extracted quickly and effectively by image thresholding segmentation processing techniques [23]. Note that the image sensors are installed in the same plane at a known lateral distance B, with their major axes coinciding. After camera calibration, the lenses could have identical properties and focal length f. The axis of each lens, normal to the image sensor plane, intersects the center of the corresponding image sensor. Hence, the coordinates can be calculated from geometric relationships and optimized from the bilinear interpolation function. In this study, we mainly focus on the impact of AP handover on the quality of location service. Therefore, for the convenience of the study, the positioning accuracy is converted to the distance between the user and the center of the area covered by the AP, without considering the specific positioning accuracy value of the positioning scheme. Note that stereo cameras have obvious advantages for distance detection.

We present the VLP scheme for connecting stereo cameras to multiple LED groups (APs) as follows. The geometric relationship of the proposed system model is detailed in Fig. 2. consider an indoor positioning scenario comprising several LED groups. Each group has identical physical characteristics but different coordinate information. Generally, the camera sensor is rectangular, while the stereo camera adopted at the receiver captures visible LED groups simultaneously through two lenses. Hence, the moving range of the receiver is an inscribed square of the coverage range of the LED group, as shown in Figure 2 (a). The overlap region between the LED groups is $s_{i,j} = R\left(L-R\right) - \frac{(L-R)^2}{2}$, where i and j denote the adjacent groups, whose corresponding distance is denoted by R, and L is the edge length of the effective coverage region of the camera under the single LED group.

The user moves around the room according to the random waypoint (RWP) model. The handover process is initiated in the overlapping regions $s_{i,j}$. Moreover, no limitation is placed on either the direction or movement of the user, as shown in Fig. 2(b). Let P_1, P_2 and P_3 denote the centers of the three reference LEDs equipped in Group B ($\mathrm{G_B}$). Point O represents the midpoint of the straight line joining the centers of the two lenses. Moreover, the centroids of the projection area for $\mathrm{G_A}$ and $\mathrm{G_B}$ are denoted as P_A and P_B, respectively. Fig. 2(c) and Fig. 2(d) show the procedure for determining the horizontal distance between point O and the LED group.

If the distances between the center of the image and that of the image sensors are i_1 and i_2 and their projections on the major axis of the image sensors are x_1 and x_2, then k_1, k_2, and k_3, the horizontal distances of point O from the three LEDs, can be calculated from the geometric relationship between the distance and the position difference of the LED images on the two image sensors as follows:

$$h = \frac{f \times B}{|x_1 - x_2|}, \tag{1}$$

$$m = \frac{i_1 \times h}{f}, \quad n = \frac{i_2 \times h}{f}, \tag{2}$$

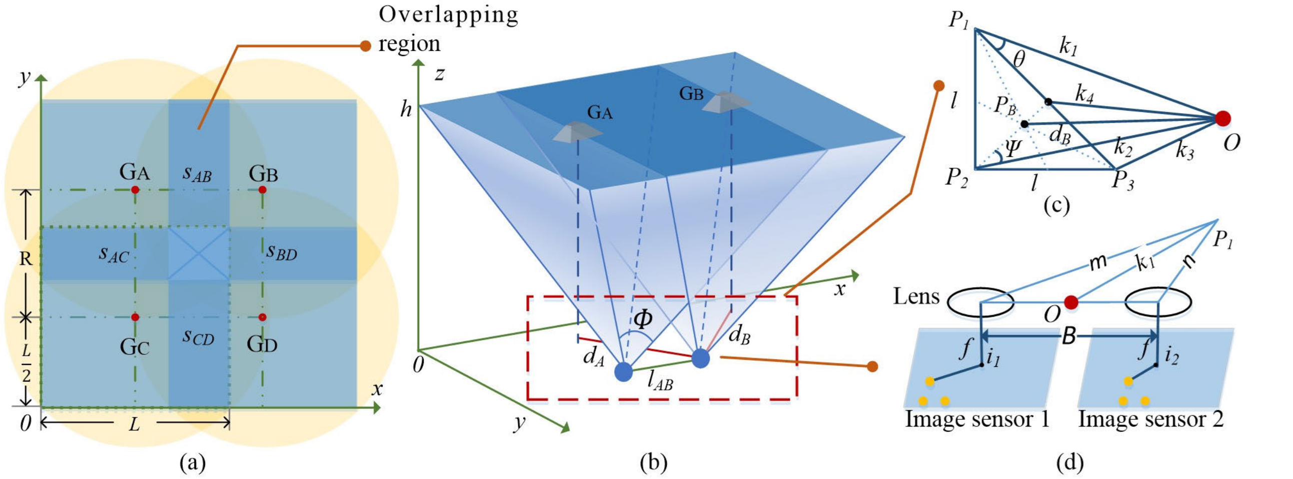

FIGURE 2. Diagram of the proposed system model: (a) and (b) camera detection range and overlapping region, (c) and (d) geometry diagram of the procedure for determining the distance d_i.

$$k_1 = \frac{\sqrt{2}}{2}\sqrt{m^2 - \frac{1}{2}B^2 + n^2} \tag{3}$$

where h is the height of the LED groups from the ground, f and B are known, and k_2 and k_3 can be obtained similarly. Here, $\Delta P_1P_2P_3$ is an isosceles right triangle; hence, the horizontal distance between O and P_B can be represented by d_B and calculated as follows:

$$\cos\theta = \frac{k_1^2 + 2l^2 - k_3^2}{2\sqrt{2}\, lk_1}, \tag{4}$$

$$k_4 = \sqrt{k_1^2 + \frac{1}{2}l^2 - \sqrt{2}\, lk_1 \cos\theta}, \tag{5}$$

$$\cos\psi = \frac{k_2^2 + \frac{1}{2}l^2 - k_4^2}{\sqrt{2}lk_2}, \tag{6}$$

$$d_B = \sqrt{k_2^2 + \frac{2}{9}l^2 - \frac{\sqrt{2}}{3}lk_2 \cos\psi}. \tag{7}$$

Similarly, the horizontal distance d_i between O and the centroid of the projection area can be obtained for the other LED groups.

III. MULTI-TARGET CROSS-REGION POSITIONING HANDOVER SCHEME

In indoor multi-LED group coverage positioning systems, the signal coverage range of a single LED group is limited; therefore, multi-user positioning switching between adjacent LED groups is a necessary component to maintain connectivity and positioning accuracy in a complex signal environment. In this section, two problems should be solved to select the best LED group and ensure QoS. The first problem is selecting the appropriate decision conditions providing the desired handover. In this regard, we propose a positioning accuracy (PA)-based handover scheme using the rate of change in the PA to indicate whether a user travels toward the central area of the LED group. The second problem is to establish a path for LED group selection to obtain the maximum PA. Thus, a heuristic RL-based optimization selection algorithm is designed to optimize the path of the LED group selection. In the previous section, the distance from the LED group was known, and the users' movement direction could be easily obtained by calculating the included angle between the position of the target AP imaging in the sensor plane and the center axis of the sensor. Fig. 3 illustrates the movement path of a user covered by several cells. Here, we considered three movement scenarios to evaluate the proposed algorithm.

Path 1: The user is directly transferred from LED Group J to LED Group I ($G_J \rightarrow G_I$) by crossing a single overlapping area (OA). The PA of the target LED group continues to increase until it is higher than the threshold, initiating handover. The immediate handover (IHO) method, which always chooses the group providing the highest PA, is applied to this simple straight-path scenario.

Path 2: The user moves from G_V to G_S via G_U, thus crossing the two OAs. The IHO hands over the user from G_V to G_U and then from G_U to G_S, even though the G_U user crosses rapidly. To suppress frequent handovers, the standard handover scheme in long-term evolution (LTE) [24] considers the idea of hysteresis, which delays the handover decision for a certain amount of time, as shown in Fig. 3. We set the dwell time t_{dw} to ensure continuity in tracking the reference LED group and maintain the original link as much as possible. Based on the dwell handover (DHO) scheme, two options are available for selecting the best LED group. In option I, G_U is selected, and the path from G_V to G_S is $G_V \rightarrow G_U \rightarrow G_S$. We found that the user is still in the G_U area when t_{dw} expires and switches to PA_{max}. In contrast, option II skips the G_U to suppress frequent switching. The path from G_V to G_S is $G_V \rightarrow G_S$. However, when the user crosses the border between G_V and G_U, G_U offers a higher PA than G_S. However, it offers a residence time of less than t_{dw}; this option lowers the handover cost. Note that for the two options, the PA from the source to the destination is monitored. A higher positioning accuracy or a lower HO rate is obtained by setting the value of t_{dw}. Nevertheless, the DHO scheme cannot skip

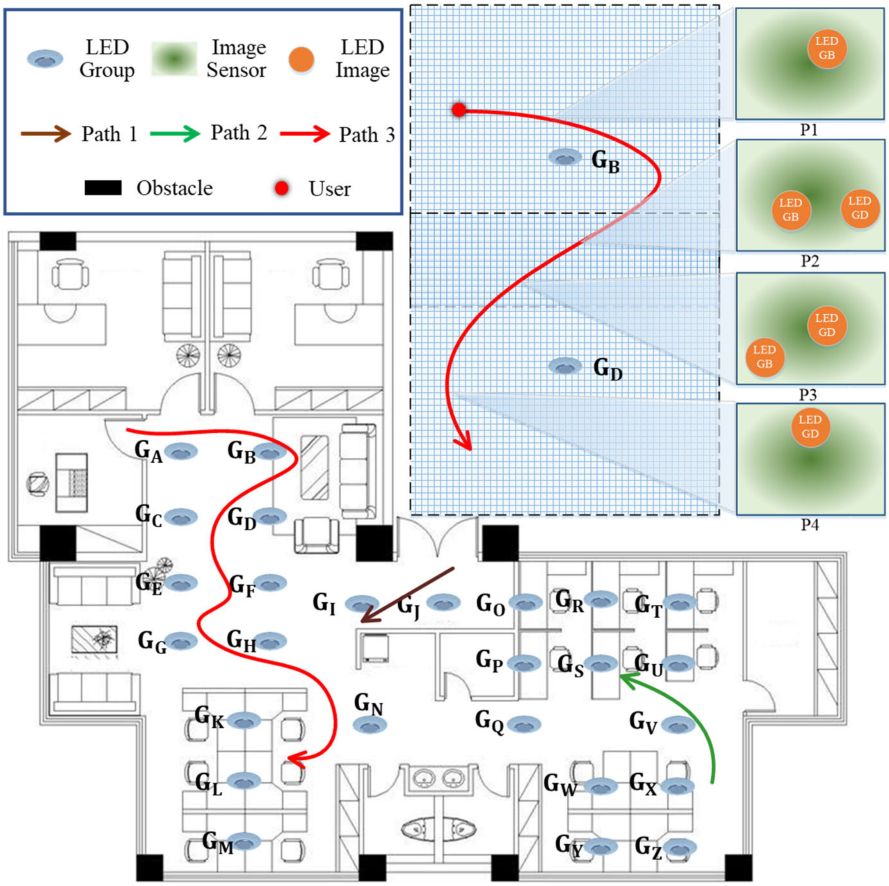

FIGURE 3. Schematic diagram of the visible light positioning handover scheme.

LED groups where the user stays longer, even slightly longer than t_{dw}.

Path 3: The user performs the reciprocating motion by constantly changing the direction from $\mathrm{G_A}$ to $\mathrm{G_L}$, crossing many OAs and resulting in a ping-pong effect. The DHO scheme can guarantee a minimum connection time, typically limited to hundreds of milliseconds. As shown in Fig. 2, obstacles block the user, and the path does not arrive at the center of the $\mathrm{G_C}$ and $\mathrm{G_E}$. However, this handover scheme cannot skip the LED groups because the stay duration of the user is much longer than t_T. Therefore, we propose a policy for selecting optimal LED groups based on an RL algorithm considering handover rates and PA. In this regard, policy evaluation computes the state- and action-value functions for a policy. Thus, the control problem involves determining the optimal policy. Moreover, planning constructs a value function or policy by using a particular model. As shown in Fig. 3, the coverage area of the camera under a single LED group is divided by a grid whose size represents the user's moving speed (the larger the grid is, the higher the speed). In each cell, the user is located in (c_t), and the group is accessed at that moment (G_t), constituting a state space S. Then, the user switches to an adjacent group or maintains the original, forming an action space $\mathcal{A}$. At each time step t, the user receives the state $s_t\,(c_t, G_t)$ in S and selects an action from $\mathcal{A}$, following a policy $\pi\,(a_t \mid s_t)$, selecting the LED groups. Subsequently, the scalar reward r_t, which transfers to the next state s_{t+1}, can be obtained according to the distribution of PA in the coverage area and handover delay for the reward function $\mathcal{R}\,(s, a)$ and state transition probability $\mathcal{P}\,(s_{t+1} \mid s_t, a_t)$, respectively. This process continues until the user reaches the terminal state and then restarts. The purpose is to find a policy that achieves the largest reward. Evidently, the reward is determined by the current state and subsequent states; thus, the proposed scheme can suppress the ping-pong effect.

IV. THEORETICAL PERFORMANCE ANALYSIS BY INTRODUCING THE RL ALGORITHM

In this section, the theoretical performance of the proposed handover scheme is analyzed for an arbitrary line trajectory. We assume that LED groups are evenly distributed on the ceiling, and several users are wandering on various routes. The positions of the users can be detected separately based on the spatial reusability of the image sensors, and the interference of different users can be ignored. As shown in Fig. 4, we employed the normalized PA to represent the distance from the user to the LED groups. A higher PA implies that the corresponding LED group is closer to the user. The positioning accuracy follows the cumulative distribution function (CDF) of the Poisson distribution and decreases with

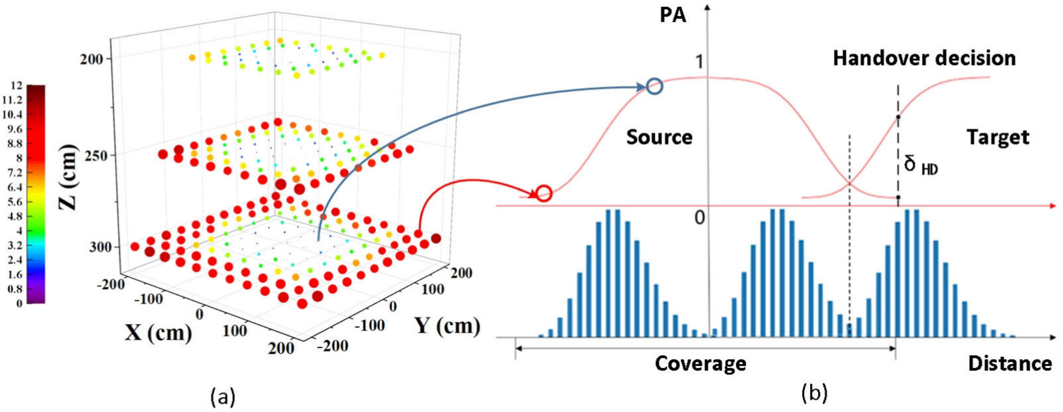

FIGURE 4. **(a) Positioning error distribution under single LED group obtained through experiments. (b) Illustration of LED group selection based on PA.**

increasing distance from the center of the LED coverage area. The PA of the user under LED Group i can be expressed as follows:

$$PA\left(d_i\right) = \int_0^{d_i} \frac{\lambda^{\mathcal{K}(x)}}{\mathcal{K}\left(x\right)!} e^{-\lambda} dx, \tag{8}$$

$$\mathcal{K}\left(d_i\right) = \frac{k\left(L/2 - d_i\right)}{L/2}, \tag{9}$$

where d_i represents the horizontal distance between the user and the centroid of one LED group, and λ and k are Poisson distribution parameters. Based on the experimental data, as shown in Fig. 4(a), the curve of positioning error variation was fitted. The vertical height was selected as 3m in subsequent simulation, λ and K were set as 10 and 20 respectively. Note that the length of the effective coverage region of the camera under the single LED Group L depends on the vertical height of the camera from the ceiling and the camera's field-of-view angle, given as $L = h * \tan\varphi$.

In particular, the positioning accuracy decreases when the distance d_i increases. Therefore, this study only focuses on the horizontal distance d_i between the user and LED groups; the difference in positioning accuracy caused by different positioning schemes is not discussed here.

The ping-pong effect is avoided by introducing the following two parameters into the handover scheme: handover delay (HD) and time to dwell (t_{dw}), as in LTE [24]. As shown in Fig. 4(b), the PAs of the source and target LED groups are denoted as $PA\left(d_S\right)$ and $PA\left(d_T\right)$, respectively. Let δ_{HD} denote the HD value. For comparison, all the handover schemes employ the same δ_{HD} and start the process when the following condition is satisfied:

$$PA\left(d_T\right) > PA\left(d_S\right) + \delta_{HD}. \tag{10}$$

A. HANDOVER RATE

Introducing the handover scheme inevitably leads to the receiver reidentifying and locating the new access LED group under the coverage of multiple LED groups. Frequent handovers may increase the computational load of the system and the communication delay of users in the VLP system, decreasing positioning accuracy, delayed updating of positioning information, and even interruption. Therefore, the handover rate, which directly reflects the handover frequency of users during the movement process, is a crucial indicator of the performance of the VLP system.

Let ξ denote the overall set of overlapping area boundaries of two adjacent LED groups, whose distances to the two LED groups are the same. This distance is less than or equal to the distances to all the other LED groups. Whenever an active user cross ξ, the reference LED group changes; thus, a handover occurs. Let $\mathcal{L}$ denote the trajectory of the user, which is finite in length. Handovers occur at the intersections between $\mathcal{L}$ and ξ, and the number of handovers experienced by the user equals the number of intersections between $\mathcal{L}$ and ξ, denoted by $\mathcal{N}\left(\mathcal{L}, \xi\right)$. To track this number, we should first study the length intensity of ξ, denoted by $\mu\left(\xi\right)$. As the developed system is single-tier [25], we have

$$\mu\left(\xi\right) = 2\sqrt{\rho} = 2\sqrt{\int_0^{R/2} x \sum\nolimits_{d_i \le x} PA\left(d_i\right) dx} \tag{11}$$

where ρ represents the expected intensity of PA in a single coverage area, R is the distance between adjacent LED groups, and $PA\left(d_i\right)$ is given by (8). Note that a higher intensity of ξ increases the boundary-crossing opportunities and thus increases the HO rate.

According to [26], the expected number of intersections between an arbitrary curve and a stationary boundary is $\pi/2$ multiplied by the length of the curve and the length intensity of the boundary. Therefore, the expected number of intersections between $\mathcal{L}$ and ξ is as follows:

$$\mathbb{E}\left(\mathcal{N}\left(\mathcal{L}, \xi\right)\right) = \frac{2}{\pi} \mu\left(\xi\right) \left|\mathcal{L}\right| \tag{12}$$

where $\left|\mathcal{L}\right|$ denotes the length of $\mathcal{L}$.

Finally, let R_{ho} denote the handover rate; then, we derive the handoff rates from (11)– (12):

$$R_{ho} = \frac{4v}{\pi} \sqrt{\int_0^{R/2} x \sum\nolimits_{d_i \le x} PA\left(d_i\right) dx \left|\mathcal{L}\right|}, \tag{13}$$

where v denotes the velocity of an active user.

Algorithm 1 RL Algorithm

Input: The position of each LED group, user's path, user's speed, and iteration number M.
Output: Q action-value function (from which a policy and select actions are obtained), the max reward of the policy, and the handover rate of the policy.

```
Initialize the simulation scenario env
Initialize the PA of each LED group at different positions as reward $r_t$
Initialize the length of the sampled path $T$
// Sample the user's path based on the user's speed
for episode = 1 to $M$ do
    Reset the simulation scenario $env$
    sum reward ⇐0
    Initialize sequence $s_1$ $(c_1, G_1)$ and selected LED group sequence $\phi_1 = \{\ \}$
        for t = 1 to $T$ do
    following the $\epsilon$-greedy policy, select
    $a_t = \begin{cases} \text{a random action} & \text{with probability } \epsilon \\ \text{argmax } Q(s_t, a_t) & \text{otherwise} \end{cases}$
            observe reward $r_t$ and next state $s_{t+1}$
            sum reward ⇐ sum reward $+r_t$
    $\phi_{t+1} \Leftarrow s_{t+1}\left(, G_{t+1}\right)$
            following $\gamma$, update Table $Q$
    $Q_t(s_t, a_t) \Leftarrow Q_t(s_t, a_t) + \alpha[r_t + \gamma Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)]$
        end
        if sum reward > max reward do
            max reward ⇐ sum reward
    end
end
```

B. REWARD FUNCTION

The reward function is defined as the sum of the PAs obtained by the user moving each step under the LED groups until the terminal state is reached. Moreover, the impact of the handover delay should be considered. The reward r_t at each step can be expressed as follows:

$$r_t = \begin{cases} PA(d_i) \times (\Delta t - t_{HD}) & \textit{Switch LED group} \\ PA(d_i) \times \Delta t & \textit{else}, \end{cases} \tag{14}$$

where Δt represents the dwell time for each step, which depends on the velocity of the active user v, and t_{HD} is the handover delay time. Note that when t_{HD} is greater than Δt, the reward for the next step is negatively affected. In an episodic problem, rewards are accumulated until the user reaches the terminal state and then restarts. If any LED group is found at each Δt, the connection is successfully established from the source to the destination. Otherwise, the location service is blocked. The action value is given as follows:

$$Q(s, a) = \mathbb{E}[R_t | S_t = s, a_t = a], \tag{15}$$

where $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ is the expected return for selecting action a in the state. The return is the discounted accumulated reward with a discount factor $\gamma \in (0, 1]$.

The aim is to find a policy maximizing the expected future reward. An optimal policy is to select the best time to switch to the nearby LED group for each state. In addition, the optimal reward value is the maximum positioning accuracy from the source to the destination. The problem was set up in a discrete state and action space. RL can operate in a model-free approach by wandering in the graph according to some policy without global information.

TABLE 1. Parameter configuration.

Room size (length × width × height)	15 m × 21 m × 3 m
FoV angle of the user device	70°
Frame Rate/FPS	30
Radius of the effective coverage region of a single AP, r	2 m
Dwell time t_{dw}	120 ms [19]
Handover overhead	60 ms
Simulation runs	20

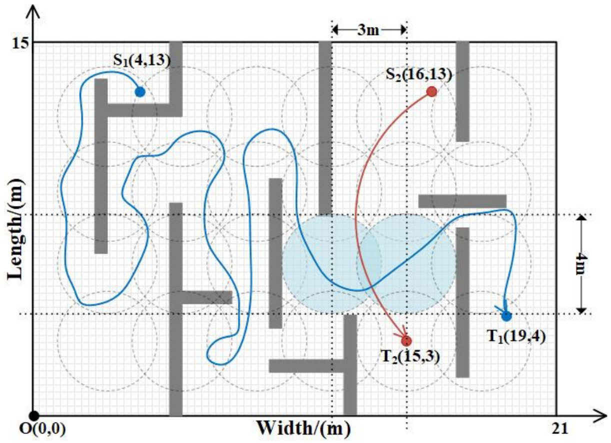

FIGURE 5. Simulation environment map and user-movement path diagram.

Furthermore, Q-learning is a temporal difference (TD) control method that is central to RL. Q-learning learns the action-value function $Q(s, a)$ direct experience with TD error, with bootstrapping, in a model-free, online, and fully incremental manner. The update rule is as follows:

$$Q_t(s_t, a_t) \Leftarrow Q_t(s_t, a_t) + \alpha [r_t + \gamma Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)], \tag{16}$$

where α is the learning rate and $r_t + \gamma Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$ is the TD error. Algorithm 1 presents the pseudocode of the Q-learning algorithm. Specifically, it is Q (0) learning, where "0" indicates that it is based on one-step returns.

The user's movement track was recorded by a stereo-camera at a fixed frame rate FPS, and the theoretical maximum value of PA at each step was 1. Therefore, the theoretical maximum value of the maximum reward of positioning accuracy is $|\mathcal{L}|/v \times FPS$. To facilitate comparison, normalized data processing was adopted to compare the effects of the proposed handover scheme on positioning accuracy, denoted as the normalized reward of PA.

V. SIMULATION AND ANALYSIS OF RESULTS

An experimental system was established using experimental VLP data and typical parameters reported for commercially available devices. The system was used to simulate the indicators of the proposed AP selection algorithm and evaluate the normalization positioning accuracy and handover rate of the proposed algorithm. The simulation environment comprising 24 Aps on the ceiling, as shown in Fig. 5, was developed in Python. The separation between the two nearest Aps was fixed at 3 m, and the coverage area of each AP was set to 4 m according to the camera's field of view angle and the distance from the roof to the camera (set as 3 m). The size of our map model was 15 × 21 m, divided into grids according to the movement speed of the user device. We set up two movement paths: a complex path $S_1(4,13) - T_1(19,4)$ and a simple path $S_2(16.13) - T_2(15,3)$, represented in blue and red, respectively. The user devices move from the starting position to the target position at different speeds along these paths. The user is assumed to have the same communication conditions within the coverage area of each AP. The parameter configurations are listed in Table 1. Moreover, the results were compared with those of the IHO and DHO algorithms. In the simulations, we used the handover rate and the normalized reward of positioning accuracy to measure the performance of the proposed algorithm.

As mentioned in the previous section, the handover rate could not be reduced due to the randomness of user movement in the coverage overlap area of different Aps (up to four Aps), and more importantly, the system performance in judging whether to switch target Aps based only on real-time positioning accuracy could not be improved. Therefore, a lower handover rate improves network performance when the normalized reward of PA of the user after completing the mobile process is sufficient.

A. THE IMPACT OF THE WEIGHT COEFFICIENT

First, we studied the effect of the RL parameters on the performance of the proposed method. The following factors were studied: ϵ-greedy, discount factors, γ, and the learning rate, α. The reward function was used as a standard to determine the parameter values.

α guides adjusting the gradient of the loss function. If α is substantially large, the gradient descent can overshoot the minimum value. However, it may fail to converge or diverge. The lower the learning rate is, the slower the loss function variation. Although using a low learning rate ensures that no local minima are missed, this attribute also implies that the convergence process is more time-consuming. As shown in Fig. 6, when the training episode is higher than 8000, the normalized reward of PA with either of three α values show an obvious convergence trend and the same limit value. Moreover, the fastest convergence rate is obtained when α is 0.3.

For the trade-off between exploring uncertain policies and exploiting the current best policy, we introduce a simple approach, that is, ϵ-greedy, where $\epsilon \in (0,1)$. In ϵ-greedy, the user selects a greedy action a_t for the current state s_t, with probability 1 - ϵ. Then, the user selects a random action with probability ϵ. That is, the user exploits the current value

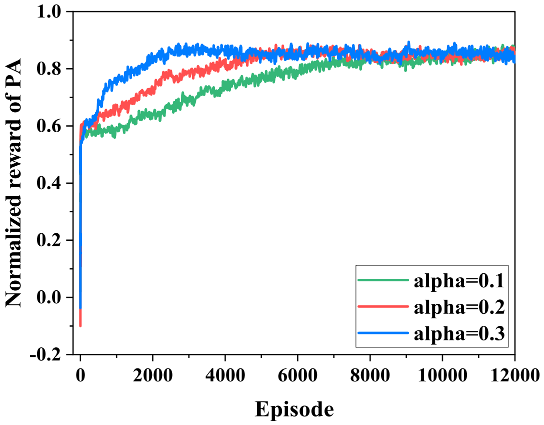

FIGURE 6. Convergence performance comparison for different values of α.

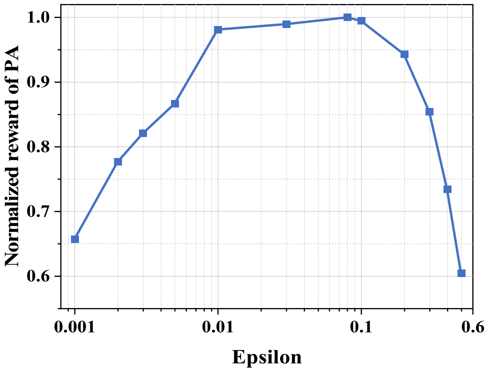

FIGURE 7. Normalization reward versus ϵ.

function $maxQ\,(s_t, a_t)$ estimation with probability 1 - ϵ and explores with probability ϵ. As shown in Fig. 7, a very small ϵ causes the algorithm to fall into a local minimum. Thus, obtaining the optimal solution is challenging. However, an excessively large ϵ causes unnecessary handover skipping.

As illustrated in Fig. 8, we selected three exploration rates ϵ (ϵ = 0.08, 0.2, and 0.3) to obtain their reward of PA. Evidently, the maximum reward of PA can be obtained when $\epsilon = 0.08$. Furthermore, with increasing ϵ, the convergence performance and the optimal value decreases. The reason is that although an increase in ϵ might bring more exploration opportunities to avoid converging to a poor local optimal value, it negatively affects the existing positive learning experience. This result is consistent with the conclusions presented in Fig. 7. In addition, we compare four discount factors γ regarding the reward of PA convergence with different exploration rates ϵ. Compared to the other three subgraphs, the convergence speed of the average return value is the fastest when $\gamma = 0.7$, and the normalized reward of PA shows an evident divergence trend with increasing γ. Note that the discount factor γ should be between zero and one, as stated in (15). If γ is 0, only the current reward is considered, implying that a shortsighted strategy is adopted. The larger the value of γ is, the greater the impact of future benefits on the value of the current action. However, a large value of γ implies that the future benefits considered by the algorithm are far beyond the scope of the current behavior,

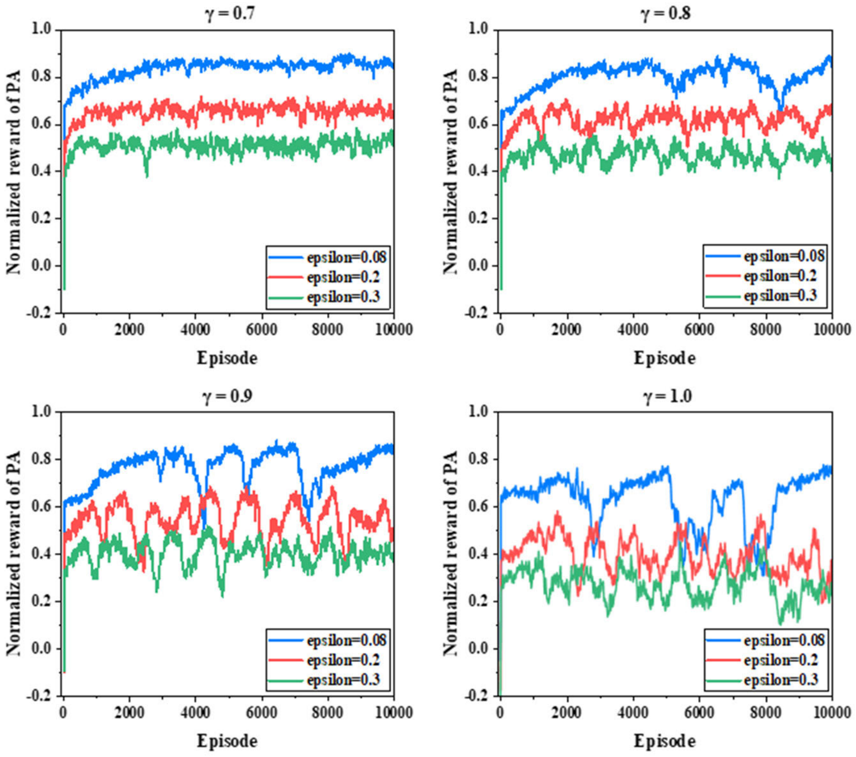

FIGURE 8. Convergence performance comparison for different values of γ and ϵ.

which is obviously unreasonable. Therefore, in subsequent experiments, the value of γ is set to 0.7.

B. HANDOVER RATE AND NORMALIZED REWARD OF PA

Second, considering the handover rate and normalized reward of PA, the positioning service quality of the proposed scheme was evaluated and compared with the IHO and DHO schemes. Fig. 9 and Fig. 10 show the quality of positioning service as users move at different speeds over simple and complex paths, respectively. Obviously, when the path is simple and has fewer obstacles, the gap between the proposed scheme and the other two schemes is small. The proposed scheme and DHO scheme can skip unnecessary handover, so the handover rate is better than IHO, as shown in Fig. 9(a). However, from the perspective of the normalized reward of the PA value, the overall positioning accuracy is not significantly improved, especially when the user moves slowly, as shown in Fig. 9(b). Nevertheless, for complex paths, the optimization effect of the proposed scheme is impressive. Fig. 10(a) shows a decrease in the handover rate when the proposed approach is compared with the IHO and DHO methods. Three outcomes are observed: i) compared with the baseline methods, the proposed scheme can effectively decrease the handover rate for different user speeds; ii) the IHO method shows the same performance as the DHO method for the handover rate when the UD movement speed is 1 m/s, which may be very slow for the dwell time t_{dw} in DHO; and iii) as the user's speed increases, the gap in the handover rate between different schemes increases. At v = 1 m/s, the proposed method achieves a handover rate that is 75% smaller than that of the DHO. When v is increased to 3 m/s, the gap decreases to 58%. For $v = 5$ m/s, the proposed approach reduces the handover rate by 78% and 87% compared with the DHO and IHO methods, respectively. However, a simple comparison of the handover rates does not completely reflect the advantages and disadvantages of the schemes; therefore, we also compared the normalized reward values of PA.

The results of the normalized reward of PA are shown in Fig. 10(b). Note that as the user speed increases, the normalized reward of PA decreases for all three schemes. This result is due to the higher handover rate of a shorter duration under a single AP coverage area with increasing user speed. The proposed scheme has the highest reward of PA and the lowest handover rate when the speed is fixed. The figure also shows that the IHO scheme outperforms the DHO scheme on the reward of PA at speeds lower than 2 m/s. However, the DHO scheme has a higher handover rate. The reason is that the dwell time makes the DHO scheme miss opportunities for a more accurate position. At speeds greater than 2 m/s, the dwell time can help the DHO scheme achieve a better reward of PA values. Fig. 10(b) shows that the proposed scheme outperforms the IHO and DHO schemes by 11% and 22%, respectively, in terms of the normalized reward of PA when the user movement speed is 1 m/s. Moreover, for $v = 5$ m/s, the proposed approach increases the normalized reward of PA by over 40% compared with the DHO and IHO methods.

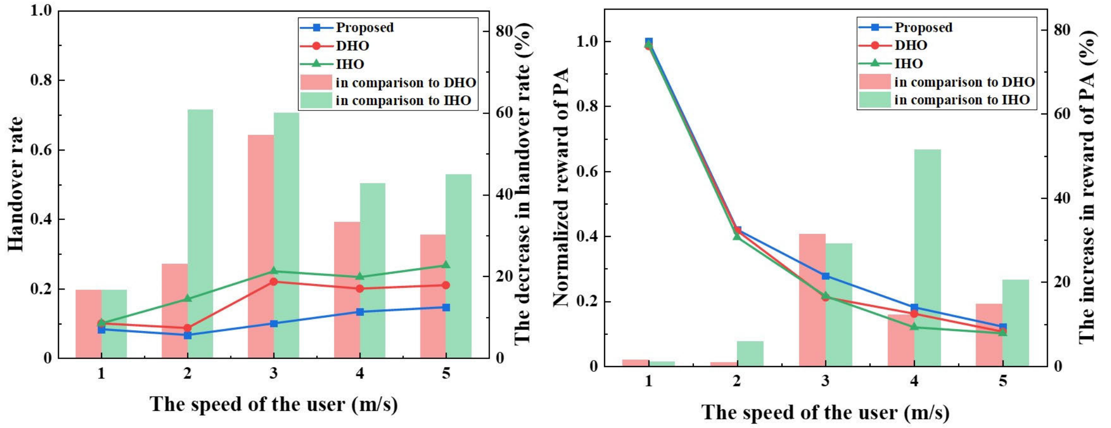

FIGURE 9. Simple path: (a) Handover rate versus the user speed. (b) Normalized reward of PA versus the user speed.

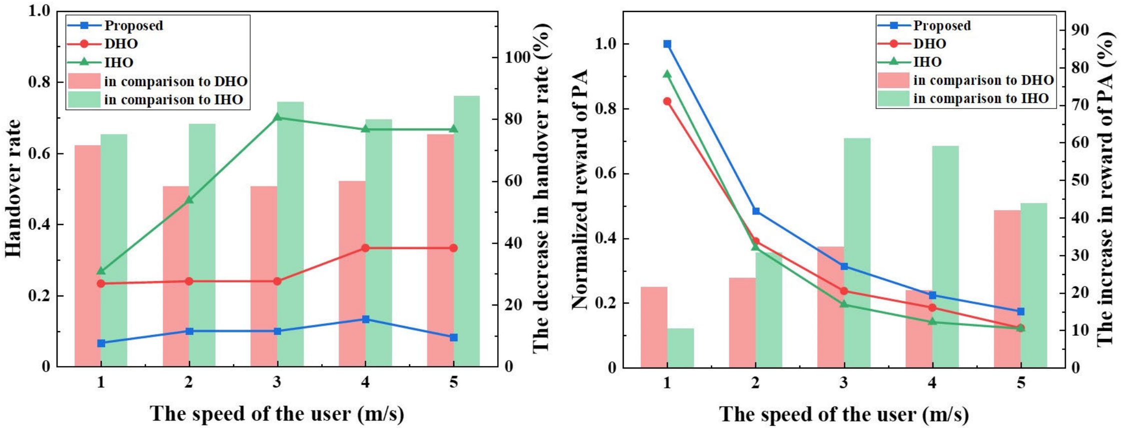

FIGURE 10. Complex path: (a) Handover rate versus the user speed. (b) Normalized reward of PA versus the user speed.

VI. CONCLUSION

Due to the intensive deployment of APs on 6G networks, to avoid frequent switching, this paper proposes an RL-based VLP handover scheme using a stereo camera. Since the accuracy of the positioning algorithm based on a camera sensor is related to the distance between the receiver and the target AP, a fast PA evaluation method was proposed based on the convenience of distance estimation by a stereo camera. Moreover, due to the randomness of indoor light sources and obstacle distribution locations, an AP adaptive handover mechanism based on the RL algorithm was proposed to balance the handover rate and positioning accuracy to reduce the impact of AP handover on location services. The simulation results showed that when the user moved rapidly, the proposed approach reduced the handover rate by 78% and 87% compared with the DHO and IHO methods, respectively. Finally, the proposed scheme improved the normalized reward of PA by over 40% compared to DHO and IHO.

REFERENCES

[1] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019, doi: 10.1109/MVT.2019.2921162.

[2] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019, doi: 10.1109/ACCESS.2019.2921522.

[3] Y. Zhuang, L. Hua, L. Qi, J. Yang, P. Cao, Y. Cao, Y. Wu, J. Thompson, and H. Haas, "A survey of positioning systems using visible LED lights," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1963–1988, 3rd Quart., 2018, doi: 10.1109/COMST.2018.2806558.

[4] R. Zhang, W.-D. Zhong, Q. Kemao, and S. Zhang, "A single LED positioning system based on circle projection," *IEEE Photon. J.*, vol. 9, no. 4, pp. 1–9, Aug. 2017, doi: 10.1109/JPHOT.2017.2722474.

[5] J. Hao, J. Chen, and R. Wang, "Visible light positioning using a single LED luminaire," *IEEE Photon. J.*, vol. 11, no. 5, pp. 1–13, Oct. 2019, doi: 10.1109/JPHOT.2019.2930209.

[6] H. Cheng, C. Xiao, Y. Ji, J. Ni, and T. Wang, "A single LED visible light positioning system based on geometric features and CMOS camera," *IEEE Photon. Technol. Lett.*, vol. 32, no. 17, pp. 1097–1100, Sep. 1, 2020, doi: 10.1109/LPT.2020.3012476.

[7] X. Wang and J. Shen, "Machine learning and its applications in visible light communication based indoor positioning," in *Proc. Int. Conf. High Perform. Big Data Intell. Syst. (HPBDIS)*, May 2019, pp. 274–277.

[8] H. Q. Tran and C. Ha, "Machine learning in indoor visible light positioning systems: A review," *Neurocomputing*, vol. 491, pp. 117–131, Jun. 2022, doi: 10.1016/j.neucom.2021.10.123.

[9] Y.-C. Wu, C.-W. Chow, Y. Liu, Y.-S. Lin, C.-Y. Hong, D.-C. Lin, S.-H. Song, and C.-H. Yeh, "Received-signal-strength (RSS) based 3D visible-light-positioning (VLP) system using kernel ridge regression machine learning algorithm with sigmoid function data preprocessing method," *IEEE Access*, vol. 8, pp. 214269–214281, 2020, doi: 10.1109/ACCESS.2020.3041192.

[10] B. Lin, Q. Guo, C. Lin, X. Tang, Z. Zhou, and Z. Ghassemlooy, "Experimental demonstration of an indoor positioning system based on artificial neural network," *Opt. Eng.*, vol. 58, no. 1, Jan. 2019, Art. no. 016104, doi: 10.1117/1.OE.58.1.016104.

[11] P. Du, S. Zhang, C. Chen, H. Yang, W.-D. Zhong, R. Zhang, A. Alphones, and Y. Yang, "Experimental demonstration of 3D visible light positioning using received signal strength with low-complexity trilateration assisted by deep learning technique," *IEEE Access*, vol. 7, pp. 93986–93997, 2019, doi: 10.1109/ACCESS.2019.2928014.

[12] H. Li, H. Huang, Y. Xu, Z. Wei, S. Yuan, P. Lin, H. Wu, W. Lei, J. Fang, and Z. Chen, "A fast and high-accuracy real-time visible light positioning system based on single LED lamp with a beacon," *IEEE Photon. J.*, vol. 12, no. 6, pp. 1–12, Dec. 2020, doi: 10.1109/JPHOT.2020.3032448.

[13] W. Guan, S. Wen, L. Liu, and H. Zhang, "High-precision indoor positioning algorithm based on visible light communication using complementary metal–oxide–semiconductor image sensor," *Opt. Eng.*, vol. 58, no. 2, Feb. 2019, Art. no. 024101, doi: 10.1117/1.OE.58.2.024101.

[14] W. Guan, L. Huang, B. Hussain, and C. P. Yue, "Robust robotic localization using visible light positioning and inertial fusion," *IEEE Sensors J.*, vol. 22, no. 6, pp. 4882–4892, Mar. 2022, doi: 10.1109/JSEN.2021.3053342.

[15] W. Guan, X. Chen, M. Huang, Z. Liu, Y. Wu, and Y. Chen, "High-speed robust dynamic positioning and tracking method based on visual visible light communication using optical flow detection and Bayesian forecast," *IEEE Photon. J.*, vol. 10, no. 3, pp. 1–22, Jun. 2018, doi: 10.1109/JPHOT.2018.2841979.

[16] Z. Xie, W. Guan, J. Zheng, X. Zhang, S. Chen, and B. Chen, "A high-precision, real-time, and robust indoor visible light positioning method based on mean shift algorithm and unscented Kalman filter," *Sensors*, vol. 19, no. 5, p. 1094, Mar. 2019, doi: 10.3390/s19051094.

[17] X. Wu and H. Haas, "Handover skipping for LiFi," *IEEE Access*, vol. 7, pp. 38369–38378, 2019, doi: 10.1109/ACCESS.2019.2903409.

[18] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020, doi: 10.1109/MWC.001.1900488.

[19] L. Wang, D. Han, M. Zhang, D. Wang, and Z. Zhang, "Deep reinforcement learning-based adaptive handover mechanism for VLC in a hybrid 6G network architecture," *IEEE Access*, vol. 9, pp. 87241–87250, 2021, doi: 10.1109/ACCESS.2021.3089521.

[20] S. M. Sheikholeslami, F. Fazel, J. Abouei, and K. N. Plataniotis, "Sub-decimeter VLC 3D indoor localization with handover probability analysis," *IEEE Access*, vol. 9, pp. 122236–122253, 2021, doi: 10.1109/ACCESS.2021.3108173.

[21] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020, doi: 10.1109/MCOM.001.1900411.

[22] B. Zhang, M. Zhang, D. Han, and C. Shi, "A visible light positioning system with improved positioning algorithm based on stereo camera," in *Proc. Asia Commun. Photon. Conf. (ACPC)*, 2019, Paper M4A.15.

[23] S. Mahajan, N. Mittal, and A. K. Pandit, "Image segmentation using multilevel thresholding based on type II fuzzy entropy and marine predators algorithm," *Multimedia Tools Appl.*, vol. 80, no. 13, pp. 19335–19359, Feb. 2021, doi: 10.1007/s11042-021-10641-5.

[24] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RCC); Protocol Specification (Release 13)*, document TS 36.331, Version 13.0.0, 3GPP, Valbonne, France, Jan. 2016.

[25] W. Bao and B. Liang, "Stochastic geometric analysis of handoffs in user-centric cooperative wireless networks," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[26] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M. Alouini, "Handover management in dense cellular networks: A stochastic geometry approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.

BO ZHANG received the B.S. degree in science and mechanical engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Photonics and Optical Communications. His research interests include optical wireless communication systems, visible light positioning, and wireless optical signal processing.

DAHAI HAN received the B.S. degree from Jilin University, China, in 2002, and the Ph.D. degree from the Beijing University of Posts and Telecommunication, Beijing, China, in 2007. He is currently an Associate Professor with the State Key Laboratory of Information Photonics & Optical Communications, Optical Wireless Communications Group, Beijing University of Posts and Telecommunications. His research interests include UV and visible light communication and detecting.

MIN ZHANG received the Ph.D. degree in optical communications from the Beijing University of Posts and Telecommunications (BUPT), China. He is currently a Professor with BUPT, the Deputy Director of the State Key Laboratory of Information Photonics and Optical Communications, and the Deputy Dean of the School of Optoelectronic Information. He holds 45 Chinese patents. He has authored or coauthored more than 300 technical papers in international journals and conferences and 12 books in the areas of optical communications. His current research interests include optical communication systems and networks, optical signal processing, and optical wireless communications.

LIQIANG WANG received the B.S. degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2017. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications (BUPT). His research interests include optical wireless communication systems and intelligent optical networks.

XIAOYUN LI is currently pursuing the M.S. degree with the Optical Wireless Communications Group, BUPT, Beijing, China. Her major contribution is to the experiment.