Reinforcement learning-based intelligent resource allocation for integrated VLCP systems

Yang, Helin; Du, Pengfei; Zhong, Wen-De; Chen, Chen; Alphones, Arokiaswami; Zhang, Sheng

2019

Yang, H., Du, P., Zhong, W.-D., Chen, C., Alphones, A., & Zhang, S. (2019). Reinforcement learning-based intelligent resource allocation for integrated VLCP systems. IEEE Wireless Communications Letters, 8(4), 1204-1207. doi:10.1109/lwc.2019.2911682

https://hdl.handle.net/10356/142886

https://doi.org/10.1109/lwc.2019.2911682

Reinforcement Learning-Based Intelligent Resource Allocation for Integrated VLCP Systems

Helin Yang, *Student Member, IEEE*, Pengfei Du, Wen-De Zhong, *Senior Member, IEEE*,
Chen Chen, Arokiaswami Alphones, *Senior Member, IEEE*, and Sheng Zhang

***Abstract*—In this letter, an intelligent resource allocation framework based on model-free reinforcement learning (RL) is first presented for multi-user integrated visible light communication and positioning (VLCP) systems, in order to maximize the sum rate of users while guaranteeing the users' minimum data rates and positioning accuracy constraints. The learning framework can learn the optimal policy under unknown environment's dynamics and the continuous-valued space, and a reward function is proposed to take into account the strict communication and positioning constraints. Moreover, a modified experience replay actor–critic (MERAC) RL approach is proposed to improve the learning efficiency and convergence speed, which efficiently collects the reliable experience and utilizes the most useful knowledge from the memory. Numerical results show that the MERAC approach can effectively learn to satisfy the strict constraints and achieve the fast convergence speed.**

***Index Terms*—Visible light communication and positioning, intelligent resource allocation, reinforcement learning, experience replay, actor critic.**

I. INTRODUCTION

WHITE light emitting diodes (LEDs) have achieved much attention recently for the long lifetime, low power consumption and reliability [1]. Besides illumination, LEDs are used for visible light communications (VLC), which have the abundant bandwidth and high security properties [1]. In addition, visible light positioning (VLP) based on LEDs has become an attractive research topic due to the high positioning accuracy as compared with radio frequency (RF)-based localization systems [2]. Recently, resource allocation has also been widely investigated in VLC and VLP systems [3]–[10] to satisfy quality-of-service (QoS) constraints of users and improve the system performance.

On the one hand, the works [3]–[6] investigated the resource allocation approaches in VLC systems, where the authors of [3] and [4] designed optical resource optimization schemes aiming to enhance the sum rate of users under the practical optical power and users' QoS constraints. In addition, the optimal power allocation for a considerable number of multiplexed users was derived in non-orthogonal multiple access (NOMA) VLC systems [5], [6]. On the other hand, the literatures [7]–[9] proposed power allocation approaches among LED lamps in VLP systems to improve the positioning accuracy. Especially, the experimental results were presented to indicate the effectiveness of power allocation based on received signal strength (RSS) in VLP systems [8]. So far, several works [8]–[10] deployed the integration of VLC and VLP (called integrated VLCP) in indoor scenarios, so both communication and positioning can be achieved at the same time, but the works [8]–[10] did not investigate the different QoS requirements in multi-user integrated VLCP systems.

The practical integrated VLCP systems indicate the uncertainty in the accurate channel information and the complete model of the systems evolution due to the mobility of users, different QoS requirements, as well as the user arrival and departure dynamics. Furthermore, the above optimization technologies [3]–[9] are usually non-convex and NP-hard, leading to a difficult search of the optimum. Hence, the model-free reinforcement learning (RL) technique is adopted to solve optimization problems in wireless RF or VLC systems [11]–[15], where the optimal policy is learned for decision making by interacting with the environment. So far, the works [12] and [13] applied experience replay to enhance the learning speed, but they did not study how to collect the reliable experience and utilize the most useful policy from historical experience.

To address the above mentioned issues, this letter firstly investigates the resource allocation in multi-user integrated VLCP systems under the different users' QoS and positioning accuracy requirements, where the system uses filter bank multi-carrier-based subcarrier multiplexing (FBMC-SCM) [10] and RSS [8]. The decision making problem for power and subcarrier allocation is modelled as a Markov decision process (MDP), and a reward function is proposed to consider users' requirements. To improve the the learning efficiency and convergence speed, a modified experience replay actor-critic (MERAC) RL approach is proposed to efficiently collect the reliable experience and utilize the most useful knowledge. Numerical results verify that the proposed MERAC approach can efficiently deploy intelligent resource allocation for multi-user integrated VLCP systems.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Integrated VLCP System Model

We consider a downlink multi-user integrated VLCP system, which consists of four LED lamps, and K users are randomly distributed on the floor, as shown in Fig. 1. There exists a central controller to connect all LED lamps and the uplink feedback is offered by the Wi-Fi links. Each user is

---

Manuscript received March 6, 2019; accepted April 13, 2019. Date of publication April 16, 2019; date of current version August 21, 2019. This work was conducted within the Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc. and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme. The associate editor coordinating the review of this paper and approving it for publication was A. Liu. *(Corresponding author: Pengfei Du.)*

H. Yang, P. Du, W.-D. Zhong, A. Alphones, and S. Zhang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: hyang013@e.ntu.edu.sg; pfdu@ntu.edu.sg).

C. Chen is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China.

Digital Object Identifier 10.1109/LWC.2019.2911682

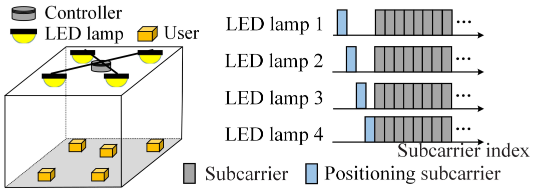


Fig. 1. An indoor multi-user integrated VLCP system.

equipped with a single photodetector (PD). The total bandwidth is equally allocated to N subcarriers, and there are four subcarriers being used for RSS positioning. Let $\mathcal{L}$, $\mathcal{N}$, and $\mathcal{K}$ denote the sets of LED lamps, subcarriers and users, respectively. The transmitted signal z_l at LED lamp $l \in \mathcal{L}$ is given by

$$z_l = \sum\nolimits_{k\in\mathcal{K}} \sum\nolimits_{n\in\mathcal{N}} \rho_{k,n}\sqrt{P_{n,l}}x_{k,n} + b_l > 0 \quad (1)$$

where $\rho_{k,n}$ is a binary variable and $\rho_{k,n} \in \{0,1\}$, $\rho_{k,n} = 1$ represents that the n-th subcarrier is allocated to user k; otherwise, $\rho_{k,n} = 0$. $P_{n,l}$ denotes the allocated transmit electrical power on the n-th subcarrier at the l-th LED lamp. In addition, let $P_{n',l}$ denotes the n'-th ($n' \in \mathcal{N}$) subcarrier which is assigned for positioning and modulated into the l-th LED lamp. $x_{k,n}$ is the data symbol for user k on the n-th subcarrier and we set that the mean of $x_{k,n}$ is zero, and it is in the range of $[-\sigma, +\sigma]$ with $\sigma > 0$ [4], [9]. b_l denotes the direct current (DC) offset of the l-th LED lamp and it is used to ensure the transmitted signal (shown in Eq. (1)) is positive.

For user k, its received electrical power on the n'-th positioning subcarrier from the l-th LED lamp is $P_{i,l}^{\text{rec}} = G_{k,l}^2 P_{i,l}$, where $G_{l,k}$ is the line-of-sight (LOS) channel gain from the l-th LED lamp to the k-th user. Then, the distance between the l-th LED lamp and the k-th user is calculated by [8]

$$d_{l,k} = \sqrt[4]{\frac{(m+1)A_r T_s(\psi_{k,l}) g(\psi_{l,k}) h^2}{2\pi} \sqrt[2]{\frac{P_{n',l}}{P_{n',l}^{\text{rec}}}}} \quad (2)$$

where A_r is the physical area of the PD, $d_{l,k}$ and $\phi_{l,k}$ denote the distance and the angle of incidence from the l-th LED lamp to the k-th user, respectively, $\psi_{k,l}$ is the incident angel. m is the order of Lambertian emission and h is the height. $T_s(\psi_{k,l})$ and $g(\psi_{k,l})$ are the gain of the optical filter and the optical concentrator gain at the PD, respectively. The user' position can be estimated based on the RSS algorithm [8]. The positioning root square error (RSE) of the k-th user is

$$RSE_k = \sqrt{(x_{k,\text{e}} - x_k)^2 + (y_{k,\text{e}} - y_k)^2} \quad (3)$$

where (x_k, y_k) and $(x_{k,\text{e}}, y_{k,\text{e}})$ are the real position and the estimated position of the k-th user, respectively.

At the receiver, the received signal-to-noise-ratios (SNRs) of the k-th user on the n-th communication subcarrier and the n'-th positioning subcarrier can be written as

$$SNR_{k,n} = \mu^2 \sum\nolimits_{l\in\mathcal{L}} P_{n,l}(G_{l,k,n})^2/\delta_{\text{total}}^2 \quad (4)$$

$$SNR_{k,n'} = \mu^2 P_{i,l}(G_{l,k,n'})^2/\delta_{\text{total}}^2 \quad (5)$$

respectively, where μ denotes the PD's responsivity. $G_{l,k,n}$ and $G_{l,k,n'}$ represent the channel gain from the l-th LED lamp to user k on the n-th and n'-th subcarrier [3], respectively. δ_{total}^2 is the total noise power, which contains the shot noise as well as the thermal noise [2]. For the k-th user, the data rate on the n-th subcarrier can be expressed as

$$R_{k,n} = \frac{B_{sub}}{2}\log_2(1 + SNR_{k,n}) \quad (6)$$

where the factor $\frac{1}{2}$ is due to the Hermitian symmetry [3], [5]. B_{sub} is the subcarrier bandwidth $B_{sub} = B/N$ with B being the system transmission bandwidth.

In practical integrated VLCP systems, users also have both the QoS and positioning accuracy requirements. These requirements can be guaranteed by controlling the probability of the unsatisfied communication service (p_k^{co}) and positioning service (p_k^{po}), where each user' data rate $R_k = \sum_{n\in\mathcal{N}} \rho_{k,n}R_{k,n}$ is below its minimum (min.) rate threshold $R_k^{\min}$, and RSE_k exceeds its maximum (max.) positioning error threshold $RSE_k^{\max}$, which can be expressed as

$$p_k^{\text{co}} = \Pr\{R_k < R_k^{\min}\},\ p_k^{\text{po}} = \Pr\{RSE_k > RSE_k^{\max}\}. \quad (7)$$

B. Problem Formulation

The policy learning problem of resource allocation in integrated VLCP systems can be modelled as a MDP with a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ [11], where $\mathcal{S}$ is the state space, $\mathcal{A}$ denotes the action space, $\mathcal{P}$ denotes the transition probability: $\mathcal{P}(s'|s,a)$ if the agent selects the action $a \in \mathcal{A}$ from the system state $s \in \mathcal{S}$ to the new state $s' \in \mathcal{S}$, r represents the reward, and $\gamma \in [0,1)$ means the discount factor.

1) Agent: Each communication and positioning service.

2) State: For each service (agent), the system state can be defined by the received SNR on the communication and positioning subcarriers as well as the subcarrier occupy status (ρ), which is given by $s = \{SNR_1, \ldots, SNR_N, \rho_1, \ldots, \rho_N\}$.

3) Action: The power and subcarrier allocation strategy is considered as the action of each agent in the RL framework, which can be defined as $a = \{P_{n,l}, \rho_{k,n}\}_{l\in\mathcal{L}, n\in\mathcal{N}, k\in\mathcal{K}}$.

4) Reward: The RL framework makes decisions by maximizing the immediate reward by interacting with the environment. Due to the requirements of users, a reward function is presented to consider the above requirements in multi-user integrated VLCP systems, which is calculated as follows

$$r = \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} \rho_{k,n}R_{k,n} - c_1 \sum_{k\in\mathcal{K}} p_k^{\text{co}} - c_2 \sum_{k\in\mathcal{K}} p_k^{\text{po}} \quad (8)$$

In (8), the first part is the utility (the sum data rate), and the second and third parts are the cost functions (unsatisfied QoS and positioning accuracy constraints). The coefficients c_1 and c_2 are the costs for the unsatisfied probabilities, and they are also used for balancing the utility and the costs.

The objective is to search a policy strategy π to maximize the reward of agents. Let $Q^\pi(s,a)$ denotes the state-action value function for starting the system state s with a given policy strategy π under action a, which is an accumulative discounted reward [11], and it is expressed as

$$Q^\pi(s,a) = r(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a)Q^\pi(s',a) \quad (9)$$

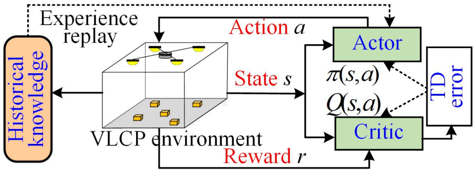


Fig. 2. MERAC learning framework for integrated VLCP systems.

The optimal policy with the selected action a is calculated recursively by applying the Bellman equation [11]

$$Q^*(s,a)=\max_{a\in\mathcal{A}}\{r(s,a)+\gamma\int_{s'\in S}P(s'|s,a)Q^{\pi^*}(s',a)\}. \tag{10}$$

III. MERAC FOR INTELLIGENT RESOURCE ALLOCATION

The problem modelled in Section II-B can be addressed by using the Q-learning and policy gradient algorithms [11]. However, Q-learning has a low convergent rate [11], [12], and the policy gradient algorithm may converge to the local optimal point [11]. Motivating by the idea of ER [11], we present the MERAC approach to improve the learning efficiency and convergence speed of finding the optimal solution in dynamic integrated VLCP systems. The learning framework of MERAC for the intelligent resource allocation is shown Fig. 2. In multi-user integrated VLCP systems, each agent observes its system state before executing resource allocation according to the learning policy strategy. After that, the new state and the immediate reward are feed back to agents from the environment. Finally, agents learn new policies and utilize the historical policy in the next step.

The learning framework based on MERAC for intelligent resource allocation is shown as follows:

1) *Experience Collection and Storing:* To avoid the risk of storing the unreliable experience, after an agent interacts with the environment, the agent stores the learned experience $e_t=(s_t,a_t,r_t,s_{t+1})$ with the best reward in the memory.

To simply the memory structure, we can combine some historical experience information into one information if they have the similar characteristics. The similarity level can be computed by using the Bregman Ball concept [11], where the Bregman Ball is regarded as the minimum manifold with a central e_{cen} (the experience information which has just been stored in the memory) and a radius e_{rad}. The agent searches the previous information point which has the most similarity with e_{cen}. The distance between two information is given by

$$D(e_{\text{cen}},e_{\text{rad}})=\left\{e_{\text{poi}}\in e : D(e_{\text{poi}},e_{\text{cen}})\le e_{\text{rad}}\right\} \tag{11}$$

where $D(\cdot)$ is the well-known Bregman divergence, which is also the manifold distance between two data points. From (11), if the two experience information have the high similarity $D(e_{\text{poi}},e_{\text{cen}})\le e_{\text{rad}}$, we can select the one with the best reward value in the memory and another one is dropped. To store the the new collected experience, the least used historical experience is dropped when the memory is full.

2) *Action Selection:* The agent selects an action a in state s with the the probability by the Boltzman distribution [11],

$$\pi(s,a)=\exp(\theta(s,a)/\tau)/\sum\nolimits_{a'\in\mathcal{A}}\exp(\theta(s,a')/\tau) \tag{12}$$

where θ (s,a) denotes the tendency to choose action a at state s and τ is the temperature with the positive value.

When a user newly joints in the system or applies new services or has poor performance, instead of building a MDP model, it can utilize the learned policy strategy from the historical knowledge by taking the following Step 5.

3) *State-Action Value Function Update:* The critic receives an instant reward from the environment at the end of stage t, and adopts the temporal difference (TD) error to evaluate the selected action, which can be expressed as [11]

$$\delta_t(s_t,a_t)=r(s_t,a_t)+\gamma Q_t(s_{t+1})-Q_t(s_t) \tag{13}$$

After that, the TD error is sent back to the actor to update the state-action value function by

$$Q_{t+1}(s_t)=Q_t(s_t)+\beta_c\delta_t(s_t,a_t) \tag{14}$$

where β_c is a positive parameter of the critic.

4) *Current Policy Update:* The policy is improved at the actor by using the TD error, and it can be updated as

$$\theta_{t+1}(s_t,a_t)=\theta_t(s_t,a_t)+\beta_a\delta_t(s_t,a_t) \tag{15}$$

where β_a is a positive parameter of the actor.

The policy will be improved with explorations, the value function Q(s,a) and the policy θ (s,a) will gradually converge to the optimal points, with the probability of 1 [11], [12].

5) *Overall Action Selection:* Different from [12] with randomly utilizing the historical knowledge, the system calculates the similarity level between the information of the learning agent and the historical information by evaluating the following three metrics: 1) service information, which mainly includes the communication services and positioning services; 2) the users information, which refers to the user location, activation behavior, mobility pattern, etc.; 3) the subchannel information, which contains the subchannel quality and subchannel assignment indicators. Similarly, we apply the Bregman Ball concept to calculate the similarity level by (11).

Once the learning agent finds the historical learning information with the highest similarity, the agent utilizes the historical action strategy a^{er} and the current native action a^{na} to generate an overall action, which is expressed as

$$a^{\text{ov}}=\omega a^{\text{er}}+(1-\omega)a^{\text{na}} \tag{16}$$

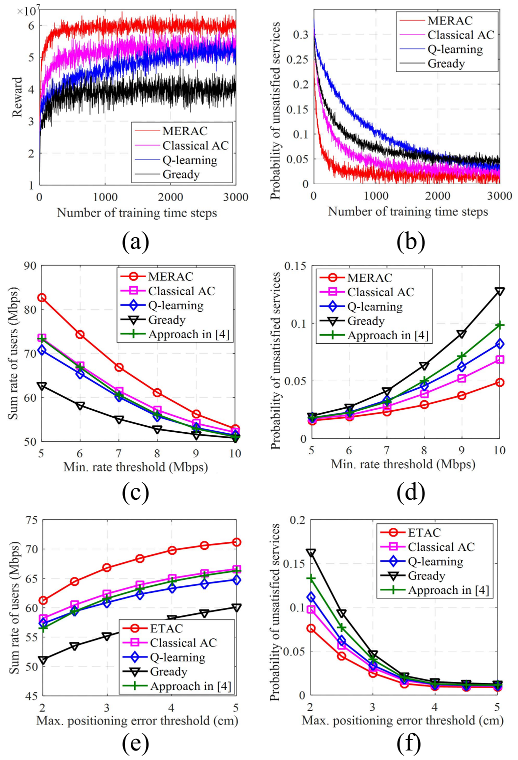
where $\omega\in[0,1]$ denotes the transfer rate, which is reduced after each iteration stage step to gradually remove the effect of the historical policy on the new action choice. We denote the sets of the historical stored state space and action space in the memory as $\mathcal{S}'$ and $\mathcal{A}'$, respectively. The action selection and learning update complexities of the proposed MERAC approach are equal to $O(|\mathcal{S}'|\times|\mathcal{A}'|+|\mathcal{S}|\times|\mathcal{A}|)$ [15], which is higher than that of the classical AC leaning approach with $O(|\mathcal{A}|)$. The proposed MERAC approach based intelligent resource allocation is provided in **Algorithm 1**.

IV. SIMULATION RESULTS AND DISCUSSIONS

This section present and discuss the performance of our presented intelligent resource allocation based on the MERAC approach in integrated VLCP systems, and compare it with the following approaches: 1. Classical AC learning [11] (denoted as Classical AC); 2. Q-learning (denoted as Q-learning); 3: Optimizing the sum rate with the constraints (7), refers to [4].

Algorithm 1 MERAC Based Intelligent Resource Allocation

```
Input: Learning rate factor $\beta_a$ and $\beta_c$, discount parameter $\gamma$, all integrated VLCP environment simulators.
Initialize: Initialize $s_0$, $Q(s_t, a_t)$ and $\pi_t(s_t, a_t)$
for each time stage $t$=0, 1, 2, ..., do
    Select an action $a_t^{\text{na}}$ at state $s_t$ based on $\pi_t(s_t, a_t)$ by (12);
    Calculate the reward $r_t$ in (8) and update state $s_{t+1}$;
    Find the historical learned action $a^{\text{er}}$ with the highest similarity, execute $a^{\text{er}}$ if the agent is new or has poor performance;
    Update (13), (14), (15) and (16), respectively;
    Update the policy function $\pi_{t+1}(s_t, a_t^{\text{ov}})$ by (12).
end for
```



(a) (b)

(c) (d)

(e) (f)

Fig. 3. The performance evaluations and comparisons. (a) Reward vs. training. (b) Unsatisfied services vs. training. (c) Sum rate of users vs R^{min}. (d) Unsatisfied services vs. R^{min}. (e) Sum rate of users vs. RSE^{max}. (f) Unsatisfied services vs. RSE^{max}.

We consider a 5m × 5m × 3m indoor room with the receiving plane 0.5m above the floor. The locations of four LED lamps are (1.25, −1.25, 3), (1.25, 1.25, 3), (−1.25, −1.25, 3) and (1.25, −1.25, 3) in meter. $K = 5$, $A_r = 0.5$ cm^2, $B = 20$ MHz and $T_s(\psi_{k,l}) = 1$. The LED lamp's semi-angle at half power is 60°, and the field of view of PD is of 120°. The PD responsivity is 0.5 A/W. The system transmission bandwidth $B = 20$MHz with 64 subcarriers, but only 31 subcarriers are used for data transmission due to Hermittian symmetry [3]. The transmit electrical power per LED lamp is 20 mW. All users have $R^{\text{min}} = 8$ Mbps and $RSE^{\text{max}} = 3$ cm. We set $c_1 = c_2 = 3 \times 10^8$ to balance the utility and the costs in (8) [12], [13]. Other relevant parameters can be found in [6], [12], and [14].

As shown in Fig. 3 (a) and (b), the MERAC approach has the best reward and satisfied services performance with the fastest convergence speed compared with other RL approaches. From Fig. 3 (c) and (d), we can observe that the performance of all approaches decreases as the minimum rate threshold $R^{\min}$ increases. This is due to that under the limited resource, more resource needs to be allocated to users with poor channel gains to guarantee their increased QoS requirements, consequently leading to the performance degradation. Fig. 3 (e) and (f) depict that the sum rate enhances and the probability of unsatisfied links reduces when the maximum positioning error threshold $RSE^{\max}$ increases. Such the performance improvement results from the looser requirement on the positioning accuracy as $RSE^{\max}$ increases, the requirement is easy to be satisfied in the case. From Fig. 3 (c) to (f), the approach in [4] can achieve the comparable sum rate performance to the RL approaches (except MERAC and greedy), but it obtains the lower probability of the satisfied services, especially when users have higher communication or positioning accuracy requirements.

V. CONCLUSION

This letter has studied the intelligent resource allocation based on RL in integrated VLCP systems. Moreover, the MERAC approach was presented to effectively learn the optimal policy by utilizing the historical learning knowledge. Numerical results verify the effectiveness of the presented MERAC approach compared with other approaches.

REFERENCES

[1] D. Karunatilaka, F. Zafar, V. Kalavally, and R. Parthiban, "LED based indoor visible light communications: State of the art," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1649–1678, 3rd Quart., 2015.

[2] M. F. Keskin, A. D. Sezer, and S. Gezici, "Localization via visible light systems," *Proc. IEEE*, vol. 106, no. 6, pp. 1063–1088, Jun. 2018.

[3] R. Jiang, Q. Wang, H. Haas, and Z. Wang, "Joint user association and power allocation for cell-free visible light communication networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 1, pp. 136–148, Jan. 2018.

[4] D. Bykhovsky and S. Arnon, "Multiple access resource allocation in visible light communication systems," *J. Lightw. Technol.*, vol. 32, no. 8, pp. 1594–1600, Apr. 15, 2014.

[5] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.

[6] C. Chen, W.-D. Zhong, H. L. Yang, and P. F. Du, "On the performance of MIMO-NOMA-based visible light communication systems," *IEEE Photon. Technol. Lett.*, vol. 30, no. 4, pp. 307–310, Feb. 15, 2018.

[7] M. F. Keskin, A. D. Sezer, and S. Gezici, "Optimal and robust power allocation for visible light positioning systems under illumination constraints," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 527–542, Jan. 2019.

[8] Y. Xu *et al.*, "Accuracy analysis and improvement of visible light positioning based on VLC system using orthogonal frequency division multiple access," *Opt. Exp.*, vol. 25, no. 26, pp. 32618–32630, Apr. 2018.

[9] M. Aminikashani, W. Gu, and M. Kavehrad, "Indoor positioning with OFDM visible light communications," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, 2016, pp. 505–510.

[10] H. L. Yang *et al.*, "Demonstration of a quasi-gapless integrated visible light communication and positioning system," *IEEE Photon. Technol. Lett.*, vol. 30, no. 23, pp. 2001–2004, Dec. 1, 2018.

[11] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[12] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor–critic learning framework for energy saving in cellular radio access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

[13] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33275–33284, 2018.

[14] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on efficient transfer actor-critic reinforcement learning for IoV communication networks," *IEEE Trans. Veh. Technol.*, to be published.

[15] N. Mastronarde and M. van der Schaar, "Fast reinforcement learning for energy-efficient wireless communication," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6262–6266, Dec. 2011.