Received April 1, 2018, accepted May 28, 2018, date of publication June 7, 2018, date of current version July 6, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2844882

Context-Aware Indoor VLC/RF Heterogeneous Network Selection: Reinforcement Learning With Knowledge Transfer

ZHIYONG DU[1], (Member, IEEE), CHUNXI WANG[2], YOUMING SUN[2], AND GUOFENG WU[2]
[1]College of Information and Communication, National University of Defense Technology, Wuhan 430010, China
[2]National Digital Switching System Engineering and Technological Research Center, Zhengzhou 450002, China

Corresponding author: Chunxi Wang (firefiy211@126.com)

This work was supported by the NSF of China under Grants 61601490 and 61671477. A preliminary version is published in the 2018 International Conference on Smart Materials, Intelligent Manufacturing, and Automation (SMIMA 2018).

ABSTRACT For the converged use of LTE, WLAN, and visible light communication in indoor scenarios, fine-grained and intelligent network selection is essential for ensuring high user quality of experience. To tackle the challenges associated with dynamic environments and complicated service requirements, we propose a context-aware solution for indoor network selection. Specifically, three-level contextual information is revealed and exploited in both the utility and algorithm designs. In particular, the contextual information about the asymmetric downlink-uplink features of network performance is used to design a fine-grained utility model. A context-aware learning algorithm sensitive to traffic type-location-time information is proposed. The time-location dependent periodic changing rule of load statistical distributions is further used to realize efficient online network selection via knowledge transfer. The simulation results show that the proposed algorithm can achieve much better performance with faster convergence speed than traditional reinforcement learning.

INDEX TERMS Indoor network selection, visible light communication, context-awareness, reinforcement learning, transfer learning.

I. INTRODUCTION

Currently, the proliferation of multimedia applications is increasing the demand for high data rate wireless services, which poses a great challenge for the emerging fifth generation (5G) mobile networks. Meanwhile, a general observation in [1] has shown that approximately 80 percent of data communication occurs indoors. Thus, improving wireless service quality for indoor users is an important issue. The converged use of different wireless networks by dynamically accessing LTE, WLAN and Visible Light Communications (VLC) [2], [3] could be an effective solution for improving indoor wireless communication quality. LTE is an evolving commercial mobile communication network that provides basic wireless access. WLAN is today's most widely used indoor wireless network. VLC is a newly emerging indoor wireless access solution. Many researchers agree that the emerging VLC is a promising solution in the 5G era for its tremendous value and potential. VLC possesses multiple advantages, such as high data rates, huge bandwidth, no electromagnetic interference and high security [4].

However, the complexity of the involved factors makes the access network selection challenging. First, as many new traffic types, such as virtual reality and online ultra-high definition video, emerge, characterizing the network selection utility becomes more complex. Second, the available wireless networks with different access technologies and owners show diversity and uncertainty in their performances due to channel conditions and user arrival and departure dynamics. Third, the traffic type is time-varying since user application changes and network performances may vary across time and locations. In other words, the optimal network selection choice varies with the environment and ensuring high user Quality of Experience (QoE) requires fine-grained, intelligent network selection methods.

In this paper, a context-aware solution is proposed for indoor network selection. Specifically, three-level contextual information is explored to understand the task. On the first level, the information about the asymmetric downlink-uplink features of network performance and traffic requirements is modeled in the utility. On the second level, the traffic

type-location-time information is used to design a learning algorithm. Finally, the periodic changing rule of load statistical distributions is used to further assist the learning algorithm. In particular, such information enables us to present knowledge transfer for reusing learning experiences, providing an effective and fast algorithm for network selection with contextual evolution.

Our main contributions are two-fold. First, we propose a fine-grained network selection model that takes the diverse and asymmetric downlink-uplink features of network performance and traffic requirements into account. Although many works on network selection, e.g., [5], have considered service requirements, the utility designs that differentiate uplink and downlink requirements of different traffic types as proposed in this paper are rare. Second, we propose a context-aware learning algorithm. The algorithm is sensitive to traffic type-location-time information, and thus is able to actively adapt the contextual evolution. In addition, the idea of transfer learning [6] is used in network selection. Even though some works such as [7] have studied the context-aware network selection, they worked in different ways and did not employ learning algorithm or knowledge transfer. Compared with some existing works that use reinforcement learning [9], [10], the introduction of transfer learning could significantly enhance the algorithmic performance, which can be found from the simulation results in Section VI. This method may provide a new perspective on endowing contextual awareness in solutions for self-organization and online optimization related problems [12].

The rest of this paper is organized as follows. We briefly review some related works in Section II. Then, the system model and designed traffic utility models will be introduced in Section III and Section IV, respectively. Next, we detail the proposed reinforcement learning with knowledge transfer in V and give related simulation results in Section VI. The final conclusions are drawn in Section VII.

II. RELATED WORK

The study of network selection in VLC heterogeneous networks is still in the infancy but has attracted much attention. In [13], Bao *et al.* analyzed the hybrid VLC and femtocell network and designed a protocol for access and handover control. In the proposed simple mechanism, the user switches to a VLC network as long as the user is in the VLC coverage and the channel gain is larger than a predefined threshold value, which does not fully consider the users' real achievable rates. Rahaim *et al.* [14] also presented a network handover scheme that improves the total throughput of a WiFi/VLC hybrid network, where the VLC is regarded as a compensatory access and the user will be allocated to the VLC network only when the WiFi is overloaded. In [15], Wu *et al.* proposed a fuzzy-logic based selection algorithm. However, it depends on preliminary training work. In [16], Wang *et al.* formulated the problem as a Markov decision process, but their work mainly focused on how to achieve the optimal trade-off between energy consumption and delay requirements. In [17], Liang *et al.* utilized both analytic hierarchy process (AHP) and cooperative games (CGs) to propose a AHP-CG algorithm for VLC heterogeneous networks.

Contextual information has been taken into account in some network selection methods. Generally, the contextual information mainly includes the traffic types, user demands, hardware conditions and so on. In [7], the application types and hardware conditions are considered in small cell association and the problem is formulated as a matching game between small cell base stations and users. The exploration of additional contextual information extracted from users' devices, such as the typical set of active applications, is proposed in [8]. We consider finer-grained contextual information, which is a vector consisting of the user's traffic type, location, time and available network set. This could guide network selection. Specifically, the uplink and downlink requirements of different traffic types and the diverse uplink and downlink performances of networks are modeled. The contextual information about the time-location dependent network load distribution inspired us to introduce transfer learning, which enables the reuse of learning experience and is able to significantly accelerate the learning convergence. Although reinforcement learning has been used in some recent works [9]–[11], the idea of reusing learning experience was not found. Therefore, to the best of our knowledge, this work is the first to introduce reinforcement learning with knowledge transfer into network selection.

III. SYSTEM MODEL

We consider an indoor heterogeneous wireless access environment that consists of N networks $\mathcal{N} = \{1, 2, \ldots, N\}$ of LTE, WLAN and VLC. Fig. 1 shows an example of the considered network. For simplicity, we use the term "network" to represent a base station (BS) in the LTE or an access point (AP) in the WLAN and VLC. We assume that a user is located in the overlapping area of N wireless networks. In a slotted system with the epoch duration of l seconds, the user can dynamically change its access network, but

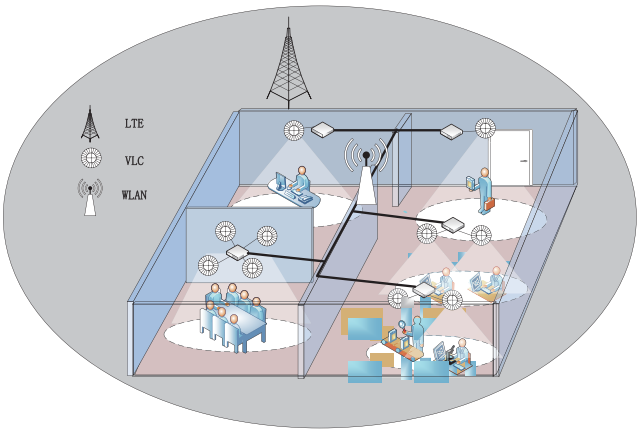


FIGURE 1. The system model of an indoor VLC/RF heterogeneous wireless network.

only one access network can be accessed at any given time slot.

We use throughput as the main performance metric of the networks. Many other performance metrics could be involved, which is beyond the focus of this paper. The maximal instantaneous rate of a user that is determined by the SNR (signal to noise ratio) according to the Shannon formula constitutes the upper bound of its throughput. Meanwhile, the multi-user access behavior determines the real-time network load distribution and thus affects the achieved throughput of each user in the network. Therefore, the achieved throughput $\Theta(i,n)$ of user i in network n is a function of the instantaneous rate R and the network load K_n (the total number of users in network n) $\Theta(i,n) = f(R, K_n)$ for a given slot. The function $f(\cdot)$ could be modeled depending on the specific network. In the following, the uplink and downlink throughput models of LTE, WLAN, and VLC are given.

A. LTE

The OFDMA is the downlink multiple access technology of LTE. According to the model in [5], the throughput under weighted-proportional fairness can be expressed as

$$\Theta_{\mathrm{DL}}(i,n) = \frac{\omega_i R_{n\to i}}{W_k} \tag{1}$$

where ω_i is user i's weight, $W_k = \sum_{i\in\mathcal{K}_n} \omega_i$ is the sum of weights of users, $\mathcal{K}_n$ is the set of users in network n and $K_n = |\mathcal{K}_n|$, $R_{n\to i}$ is the instantaneous downlink rate of user i.

In the uplink, LTE uses the SC-FDMA based MAC protocol with fair subcarrier sharing. Hence, the throughput of user i is roughly dependent on the total number of users sharing the same network,

$$\Theta_{\mathrm{UL}}(i,n) = \frac{R_{n\leftarrow i}}{K_n} \tag{2}$$

where $R_{n\leftarrow i}$ is the instantaneous uplink rate of user i

B. WLAN

In 802.11 WLAN MAC protocols, the distributed coordination function (DCF) leads to a fair access opportunity to uplink users. Hence, the low rate user capturing the channel will use it for a long time, thus penalizing high rate users. The uplink throughput of a WiFi user can be expressed as

$$\Theta_{\mathrm{UL}}(i,n) = \frac{L}{\sum_{j\in\mathcal{K}_n} \frac{L}{R_{n\leftarrow i}}} \tag{3}$$

Here, L is the packet size. The throughput that a user can obtain on the downlink is related to the scheduling mechanism of the access point. According to [19], when a round-robin(RR) scheme is used, then the downlink throughput can also be derived by replacing $R_{n\leftarrow i}$ with $R_{n\to i}$ in formula (3).

C. VLC

We consider an all-optical VLC network. Downstream data transmission and illumination are combined. Currently, there is no common view on the MAC protocol specified for VLC. In most existing works, it is assumed that the system uses TDMA with RR scheduling. Thus, if user i is assigned to the n-th VLC AP, the achieved throughput becomes [20]

$$\Theta_{\mathrm{DL}}(i,n) = \frac{R_{n\to i}}{2\cdot K_n} \tag{4}$$

Note that the intensity modulation with direct detection (IM/DD) is used in VLC and only real-valued signals can be transmitted to receivers. Thus, at least half of the subcarriers must be used to realize the Hermitian conjugate of the complex-valued symbol after modulation. Consequently, the formula is divided by 2.

Using visible light in uplink may not be practical, since it would constrain equipment power and users' psychological feelings. Referring to [21], we use infrared in the uplink. The main limitation of the infrared link is its low power transmission, which often leads to a low data transmission rate (up to 4 Mbps or 1.152 Mbps in [18]). Since visible light and infrared light exhibit very similar qualitative behavior, the uplink throughput model could also be derived by replacing $R_{n\to i}$ with $R_{n\leftarrow i}$ in formula (4).

IV. UTILITY FRAMEWORK DESIGN

Considering the diverse features of various traffic types, we propose a general utility model with differentiatec uplink and downlink performance requirements. Note that we mainly focus on the throughput, but this model can be easily extended to incorporate many other performance metrics. The achieved utility $u(\Theta_{\mathrm{UL}}, \Theta_{\mathrm{DL}})$ is designed from a novel perspective.

A. UPLINK-DOMINATED TRAFFIC

For traffic such as sending files or backing up files on the cloud, the uplink throughput is the main factor affecting the performance. The downlink throughput is negligible since it is just for transmitting some control and feedback messages (no less than a small threshold, e.g., Θ_0). As an example, it can be defined using a similar utility representing file transfer.

$$u(\Theta_{\mathrm{UL}}, \Theta_{\mathrm{DL}}) = \mathrm{I}\{\Theta_{\mathrm{DL}} \geq \Theta_0\}\,\lambda log(\beta\cdot\Theta_{\mathrm{UL}}) \tag{5}$$

where $\mathrm{I}\{x\} = 1$ when $x = 1$, and otherwise $\mathrm{I}\{x\} = 0$. $\mathrm{I}\{\Theta_{\mathrm{DL}} \geq \Theta_0\}$ is the minimal downlink throughput requirement and $\lambda log(\beta\cdot\Theta_{\mathrm{UL}})$ models the utility-throughput function [5].

B. DOWNLINK-DOMINATED TRAFFIC

On the contrary, downloading files and watching online videos mainly utilize the downlink throughput and can be classified as downlink dominant traffic. Since most existing works focus on this traffic type, the utility $u(\Theta_{\mathrm{DL}})$ can be easily derived by explicitly indicating the downlink throughput Θ_{DL} in existing utility models. For instance, the file download utility can use the above model by replacing Θ_{UL} with Θ_{DL}. Video traffic shows a threshold effect on throughput. Then,

a piecewise function of the downlink throughput plus the basic uplink throughput requirement is

$$u(\Theta_{UL}, \Theta_{DL}) = \begin{cases} 0 & \Theta_{DL} \le \Theta_1 \\ \dfrac{c(\Theta_{DL} - \Theta_1)}{\Theta_2 - \Theta_1} \mathrm{I}\{\Theta_{UL} \ge \Theta_0\} & \Theta_1 < \Theta_{DL} < \Theta_2 \\ c\mathrm{I}\{\Theta_{UL} \ge \Theta_0\} & \Theta_{DL} \ge \Theta_2 \end{cases} \quad (6)$$

where c is a constant, and Θ_1 and Θ_2 are two throughput thresholds determined by the traffic requirements.

C. UPLINK-DOWNLINK SYMMETRIC TRAFFIC

Video calls and video conference traffic have high requirements on both the downlink and uplink throughput. Either uplink or downlink throughput can be the bottleneck. We can replace Θ_{DL} with $\Theta_{min} = \min(\Theta_{UL}, \Theta_{DL})$ in formula (6) to get a utility function.

V. PROPOSED SOLUTION

A. LEARNING PROBLEM FORMULATION

Due to channel fading and the shadowing effect, the instantaneous rates $R_{n\leftarrow i}(t)$ and $R_{n\rightarrow i}(t)$ are time-varying. Moreover, the network load K_n is a random variable since the active user number in a network is dynamic. Consequently, the achieved throughput $\Theta(i, n)$ and the resulting $u(\Theta_{UL}, \Theta_{DL})$ are dynamic and random variables. Hence, it is reasonable to select the network that provides the best average performance. However, since the user has no prior knowledge of the average performance of the available networks, he has to learn the optimal selection from the interaction with the environment. Mathematically, this learning problem can be formed to select a network selection policy π^* that maximizes the long term average reward. In other words, it selects a series of actions $\{a(1), a(2), \cdots\}$ that can maximize the total expected return as

$$V^* = \max E\left\{\sum_{t=0}^{\infty} \gamma^t u\left[\Theta_{UL}(a(t)), \Theta_{DL}(a(t))\right]\right\} \quad (7)$$

where $\gamma \in (0, 1)$ represents the discount factor that reflects future returns relative to their current importance. $u[\Theta_{UL}(a(t)), \Theta_{DL}(a(t))]$ is the instant reward received at time t, and $\Theta_{UL}(a(t))$ and $\Theta_{UL}(a(t))$ are the instant uplink and downlink throughput, respectively.

B. REINFORCEMENT LEARNING BASICS

The problem mentioned above can be regarded as a large-scale constrained dynamic optimization problem embedded in a stochastic environment. Thus, reinforcement learning is one of the effective ways to find a solution. Among many learning algorithms, Jiang *et al.* [23] pointed out that Q learning is the most suitable for the small cell learning problem. In Q learning algorithm, the controller (learner) has to learn how to optimize its decision through historical experience by repeatedly interacting with the controlled environment in a manner of sensing, selecting an action, and obtaining a reward. Finally, the agent learns an optimal policy to maximize the total expected return as (7) over a time period.

Equation (7) can be rewritten in the form of Bellman equation [22]. Obtaining the optimal policy π^* requires solving Bellman's optimality criterion:

$$V^* = V^{\pi^*} = \max_{a\in\mathcal{N}_s}[u(t) + \gamma V^*] \quad (8)$$

For a policy π, define the Q-value corresponding to an action as:

$$Q^{\pi}[a(t)] = u(t) + \gamma V^{\pi}[a(t+1)] \quad (9)$$

where $a(t+1)$ is the action at time $t+1$. The optimal Q-value $Q^*(a(t))$ is defined as

$$Q^*[a(t)] = Q^{\pi^*}[a(t)] = u(t) + \gamma V^*[a(t)] \quad (10)$$

Then, (8) can be rewritten as

$$V^*[a(t)] = \max_{a\in\mathcal{N}}[Q^{\pi^*}[a(t)]] \quad (11)$$

Thus, $Q^*(a(t))$ can be expressed as

$$Q^*(a(t)) = u(t) + \gamma[\max_{m\in\mathcal{N}} Q^*(m)] \quad (12)$$

where m is the optional action in the action set $\mathcal{N}$. The Q learning algorithm finds the value of $Q^*(a(t))$ in an iterative manner at each t by updating the Q-value as follows

$$Q[a(t)] = (1-\alpha) Q[a(t)] + \alpha\left[u(t) + \gamma \max_{m\in\mathcal{N}} Q(m)\right] \quad (13)$$

where, α is the learning parameter. A Q learning agent tries an action, and then evaluates the consequences of the action through the sum of the immediate reward and the future reward. By trying one action at a time and decreasing the learning rate to zero in a suitable way, then as $t \to \infty$, $Q(a(t))$ converges to $Q^*(a(t))$ with a probability of 1. It learns the best action that maximizes the long-term discounted rewards.

C. REINFORCEMENT LEARNING WITH KNOWLEDGE TRANSFER

However, the standard Q learning algorithm may show slow convergence speed and poor performance due to the exploration. When the available strategy set is relatively large, there will be significant random exploration costs on bad strategies. Nevertheless, the idea of transfer learning [6] provides a feasible way to enhance the Q learning algorithm. The transfer learning transfers knowledge learned in certain *source-tasks* and uses it to improve the efficiency of machine learning in a related *target-task* apart from existing data/samples, as illustrated in Fig. 2. For reinforcement learning, the transfer learning enables us to accelerate the algorithm's convergence by using some knowledge or contextual information. One straightforward and effective transfer method is to set the initial solution in the target task based on a source task. In this way, the starting-point of the learning process could be much closer to the final target-task solution, compared to

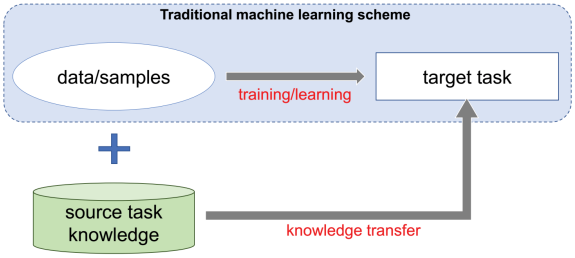


FIGURE 2. The transfer learning framework in machine learning.

the standard reinforcement learning starting at fully randomly searching.

The next concern is how to define the source-task and target-task and map the two tasks. Fortunately, we notice that the following observations may be useful.

Observation 1 (Not All Networks Are Inherently Suitable for All Traffic Types): There may be mismatches between the asymmetric downlink/uplink network performance and traffic requirements. For instance, VLC itself has poor uplink throughput due to the inherent limitations that we have mentioned. Thus, it is not suitable for the traffic with strict requirement on uplink performance.

Observation 2 (Not All Networks Are Preferred by the User for All Scenarios): The user may prefer some certain networks. For instance, if the user frequently changes his/her posture or moves around the room, the smartphone could not maintain stable VLC access, and thus VLC may not be preferred. For fee consideration, the user may not want the LTE access due to its relatively high fees.

Observation 3 (Network Load Statistical Distribution Is Time-Location Dependent): Recent literature [24] has revealed that the traffic/load shows the spatial and temporary distribution law. In other words, there is periodic changing rule with the time of the load statistical distribution for a given location. This periodic changing rule regarding the load dynamics of networks may be used. For example, the load statistical distributions at a specific location at the same duration on different weekdays are generally the same.

With these observations, we propose the Q learning algorithm with knowledge transfer as shown in Algorithm 1. To this end, we introduce a vector $(s, \mathcal{N}^*, i)$ to represent the traffic type-location-time contextual information, where $s \in \mathcal{S}$, $\mathcal{N}^* \subseteq \mathcal{N}$ and $i \in \mathcal{I}$ are the current traffic type, available network set and time period index, respectively. $\mathcal{S}$ is the set of traffic types, e.g., the three types defined in Section IV, and $\mathcal{N}$ is the maximal available network set as introduced in Section III. Note that since the available networks may vary across different locations, we use the available network set to indicate the "location" instead of exact coordinates. This type of location label is an efficient location discrimination method tailored for network selection. One day is divided into several time periods. For example, the daytime of weekdays from 8:00 am to 5:00 pm could be divided into 9 periods each corresponding to 1 hour duration. The load statistical distributions of all networks are assumed to remain unchanged in each time period.

Algorithm 1 Q Learning With Knowledge Transfer

1: **Inputs**: the discount factor γ, the learning parameter α, two initial exploration probabilities ε' and ε'', the stored learning record database $\boldsymbol{D}$.
% Initiation Stage: two initiation cases. If there is past learning experience, the stored Q table could be used.
2: **if** Current context $(s, \mathcal{N}^*, i)$ has corresponding learning record in the database **then**
3: Initialize Q table with previously learned value $\boldsymbol{Q} = \boldsymbol{D}\left[\boldsymbol{Q}_{(s,\mathcal{N}^*,i)}\right]$ and set $\varepsilon = \varepsilon'$.
4: **else**
5: Initialize Q table with $\boldsymbol{Q} = \boldsymbol{0}$ and set $\varepsilon = \varepsilon''$.
6: **end if**
% Loop Stage: algorithm and context update.
7: **loop**
8: For each slot t, based on the traffic type, select network $a(t)$ from the refined action set $\mathcal{N}_s \subseteq \mathcal{N}^*$ as follows
9: • With probability ε, choose an action at random;
10: • Else, choose $a(t) = \max_{m \in \mathcal{N}_s} Q(m)$.
11: Receive the reward $u(t)$.

$$Q[a(t)] = (1-\alpha)\, Q[a(t)] + \alpha \left[u(t) + \gamma \max_{m \in \mathcal{N}} Q(m) \right]$$

12: Update Parameters: In each iteration, the learning rate and the exploration probability could be gradually decreased in order to meet the convergence requirements.
13: Update $(s, \mathcal{N}^*, i)$.
14: **if** $(s, \mathcal{N}^*, i)$ has changed to a different $(\widehat{s}, \widehat{\mathcal{N}}^*, \widehat{i})$ **then**
15: Store the learned Q-value as $\boldsymbol{D}\left[\boldsymbol{Q}_{(s,\mathcal{N}^*,i)}\right] = \boldsymbol{Q}$
16: Go to 2nd line and start with the new context $(\widehat{s}, \widehat{\mathcal{N}}^*, \widehat{i})$.
17: **end if**
18: **end loop**

Specifically, observations 1 and 2 enable us to decrease the size of action set according to the traffic type and user preferences. That is, some network choices can be removed in the Q learning action set if they are not suitable. This is realized by selecting the traffic type-dependent action set, using the refined action set $\mathcal{N}_s \subseteq \mathcal{N}^*$ and the Q vector $\boldsymbol{Q} = [Q(1), Q(2), \ldots, Q(|\mathcal{N}_s|)]$ as shown in the 8th line of the algorithm. Observation 3 actually indicates that the load statistical distribution at the same time period and the same location across different weekdays are approximately the same. Thus, we can reuse the past learned experiences. Specifically, the past learning experience that shares the same context $(s, \mathcal{N}^*, i)$ with the current learning task is the source-task and the current learning task is the target-task. Hence, the starting-point of the current learning process could be initiated by the results derived from the corresponding source-

task, as shown in Fig. 3. In the algorithm, the context-specific learning experience in terms of Q tables $\boldsymbol{Q}_{(s,\mathcal{N}^*,i)}$ is stored in a database. Once it is found that there is already some learning record for current context $(s, \mathcal{N}^*, i)$, the learned Q table will be used for initiation. Otherwise, the Q table is initiated with a 0 vector, as shown in the 1st to 5th lines of the algorithm. Accordingly, the initial exploration probability is $\varepsilon' < \varepsilon''$. In the loop, the algorithm updates the Q table and also detects the context change. Once the context varies due to the change of traffic type, available network set or time period, the learned experience in terms of the Q table will be stored and followed by a restart of the algorithm with the new context $(\widehat{s}, \widehat{\mathcal{N}}^*, \widehat{i})$. This process realizes the context-dependent learning and knowledge transfer.

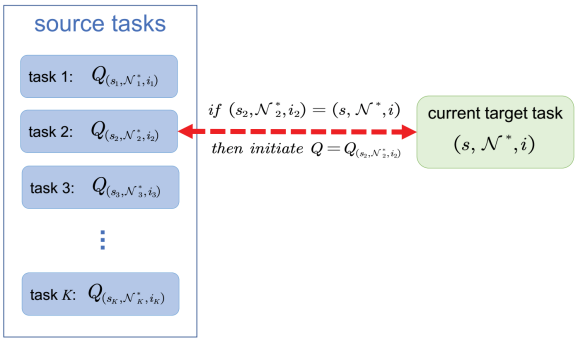


FIGURE 3. The transfer learning in the proposed algorithm. The source-task and the target-task are mapped according to the context vector.

We make several remarks on the algorithm. Firstly, the introduction of knowledge transfer mainly modifies the Q table according to contexts and has no change to the learning framework, thus, the convergence of the transferred reinforcement learning still holds [25]. Secondly, there is a concern about the division of time periods in the algorithm. Given the fixed traffic type location variation pattern, the resolution of time periods affects the performance and convergence of the proposed algorithm. Apparently, a larger time period length indicates a smaller context vector space and a longer learning experience length T for each context vector. However, this may experience varying load statistical distributions and thus incur negative learning experiences. A shorter time period length could provide more fine-grained contextual differentiation and a larger context vector space, which indirectly reduces the sample number in reusing experience for each context vector. Therefore, the division of time period should be carefully evaluated according to the evolution law of network load statistical distributions. Thirdly, although the size of saved Q tables in the database will grows linearly as the increase of experienced contexts or situations, the storage complexity of the algorithm is very limited due to: the number of typical context is small, *e.g.*, home, office and playground, and the stored information (context vector and Q table) is very limited for each context. Fourthly, thanks to the learning experience reuse, the proposed algorithm can greatly cut down the exploration frequency, that is, the visit of random selection behavior in 9th line of the algorithm is restrained. Thus, the effect of "ping-pong" and associated handoff cost is greatly reduced, compared with the standard reinforcement learning. The negative effects can be further alleviated by carefully choosing the slot duration and resorting to some multi-path concurrent transmission protocols, such as the stream control transmission protocol that is able to provide multi-homing and redundant paths facilitating smooth network handoff with low-cost. Finally, the context vector can be easily extended to include other factors if a finer context resolution is needed, such as user's activity description, age, preference and some other user profiles. In addition, different users may share their learning experience to further improve the learning efficiency if their have common context vectors.

VI. SIMULATION RESULTS

A. SIMULATION SETUP

We consider an indoor scenario composed of one LTE small cell, two WLAN access points and one VLC access point. In the LTE, WLAN and VLC standards, the user achieved instantaneous rate is discrete, which is determined by the user's location and varies with the fading effect over time. Following a similar idea in reference [26], we make a set of discrete achievable peak rates $R_{1,k} < R_{2,k} < \cdots < R_{M_k,k}$ for each network k, where M_k is the maximum number of achievable rates in network k. The data rate dynamic ranges of the LTE small cell and WLAN are set by referring to some measured data from the "Speedtest" app. Specifically, the dynamic ranges of downlink date rates of the LTE, WLAN and VLC are [4000 kbps, 7000 kbps], [3000 kbps, 10000kbps] and [8000 kbps, 13000 kbps], and their dynamic ranges of uplink date rates are [500 kbps, 6000 kbps], [3000 kbps, 9000 kbps], and [80 kbps, 120 kbps], respectively. The maximal number of active users in a network is 8. The slot duration is assumed to be 1 minute. We adjust the actual numbers of active users and the data rate dynamic ranges of the four networks to create network performance diversity. Some other parameters are listed in Tab. 1. The simulation listed below is based on the Monte-Carlo method averaged over 500 times.

TABLE 1. Parameter set.

parameter	value	parameter	value
l	30s	Θ_0	50kbps
λ	1	c	10
β	2	γ	0.3
Θ_1	100kbps	ε'	0.1/0.3
Θ_2	8000kbps	ε''	0.3

B. RESULTS

In this subsection, we first run the proposed algorithm to assess its convergence and compare it with several existing algorithms in static contexts. Then, we consider another scenario where the learning algorithms may not converge in time

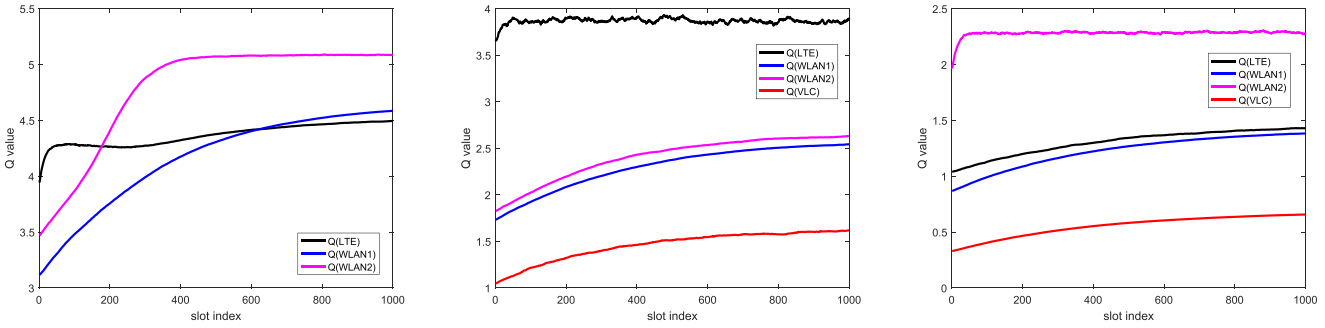


FIGURE 4. Q values in different traffic types. (a) Uplink-dominated traffic. (b) Downlink-dominated traffic. (c) Uplink-downlink symmetric traffic.

due to the context evolution and check the advantage of the knowledge transfer.

1) CONVERGENCE AND PERFORMANCE COMPARISON WITH STATIC CONTEXT

Since Q-value is a key parameter indicating the convergence of Q learning, Fig.4 shows the convergence behavior of average Q-values of the proposed algorithm for the three traffic types. The algorithm initiates with the average Q values learned after the 200-th slot iteration of the standard Q learning algorithm and the exploration probability $\varepsilon' = 0.1$. We can see that after a period of learning, all Q-values finally converge to some stable and diverse values. Importantly, for the uplink-dominated traffic (VLC is not considered due to its low uplink data rate), the Q-values have experienced a dramatic change in which the largest Q-value shifted from LTE to WLAN2. Nevertheless, we found that the phenomenon actually reflects the slower convergence, partly because the log utility leads to quite small gaps among different data rate samples. The following average reward result shows that it could also converge to a stable state.

Fig. 5 to Fig. 7 show the performance comparisons of several algorithms for the three traffic types. Since observations 1 and 2 have revealed that the uplink of VLC could hardly support the high uplink performance requirement, we can remove VLC to obtain the Q learning algorithm with the refined action set. The reuse of learning experience revealed by observation 3 is called Q learning with experience. The last one is the proposed Q learning with knowledge transfer algorithm (i.e., Q learning with refined action set, learning experience and $\varepsilon' = 0.1$). We can observe the following: i) The proposed Q learning with knowledge transfer algorithm converges much faster than the other algorithms (It converges even at the beginning, except for the uplink-dominated traffic type.) Furthermore, it achieves the largest average reward in all cases; ii) The Q learning with experience and the standard Q learning converge to nearly the same average reward, but the former converges much faster owing to the reusing of learned Q-values; iii) Compared with the standard Q learning, the Q learning with refined action set for uplink-dominated traffic could obtain better performance on both

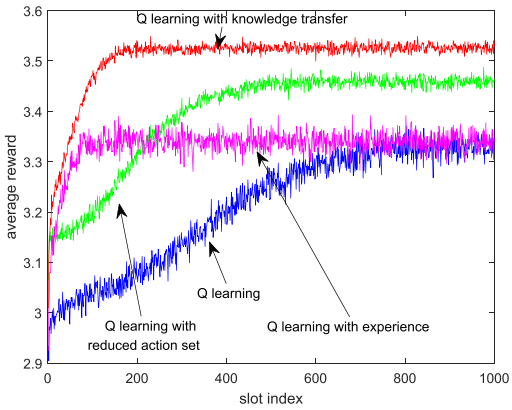


FIGURE 5. Performance comparison of different algorithms with uplink-dominated traffic.

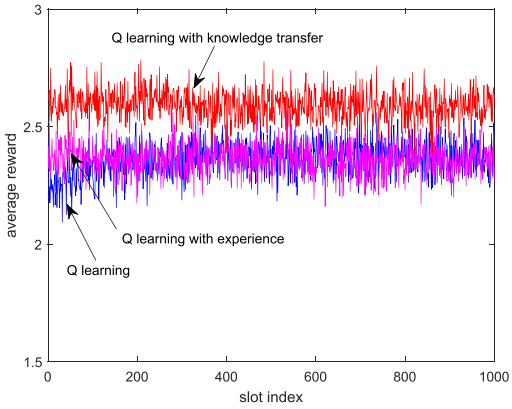


FIGURE 6. Performance comparison of different algorithms with downlink-dominated traffic.

the convergence and final average reward. However, it has a slower convergence speed than the Q learning with experience. These results indicate that the reuse of learned Q-value and the action set reduction could improve the algorithm's convergence speed and achieved performance, respectively. The relatively small exploration probability ε' could further increase the average rewards in the proposed algorithm.

We also present the convergence results of the proposed algorithm with different learning experience lengths.

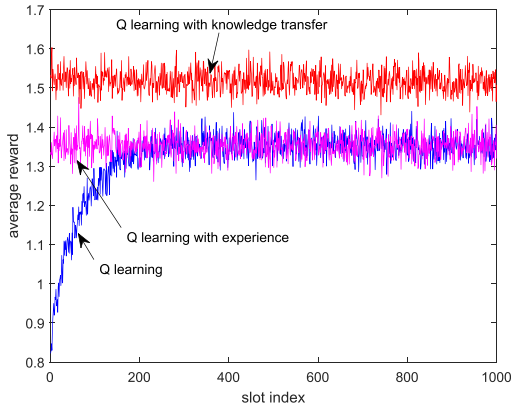


FIGURE 7. Performance comparison of different algorithms with uplink-downlink symmetric traffic.

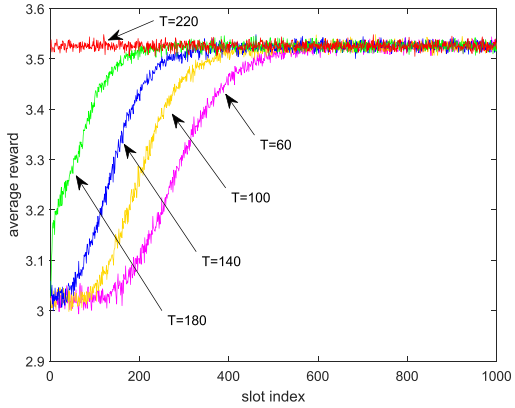


FIGURE 8. Convergence behaviors with different learning experience lengths.

As shown in Fig. 8, when the learning experience length T (the number of slots to derive the averaged Q-values for reusing, grows from 60 to 220, the algorithm's convergence speed becomes faster.

2) PERFORMANCE COMPARISON WITH CONTEXT EVOLUTION

We consider a practical scenario where the context may evolve and thus the learning algorithms may have insufficient time to converge. We assume that the traffic type s in $(s, \mathcal{N}^*, i)$ evolves in the order "uplink-dominated $\rightarrow$ downlink-dominated $\rightarrow$ uplink-downlink symmetric $\rightarrow$ uplink-dominated $\rightarrow$ downlink-dominated $\rightarrow$ uplink-downlink symmetric" and each traffic type lasts for 50 slots. Moreover, the time periods i in $(s, \mathcal{N}^*, i)$ are different for the first 150 slots and the second 150 slots. In other words, we generate different performances (data rates and user number distributions) of all networks for the two ranges. In addition to the proposed algorithm, the standard Q learning algorithm is not sensitive to the context change and thus does not restart itself. The dynamic Q learning will restart with all-0 Q-values once the context changes.

Fig. 9 and Fig. 10 show the related results with different exploration probabilities. We can see in Fig. 9 that the

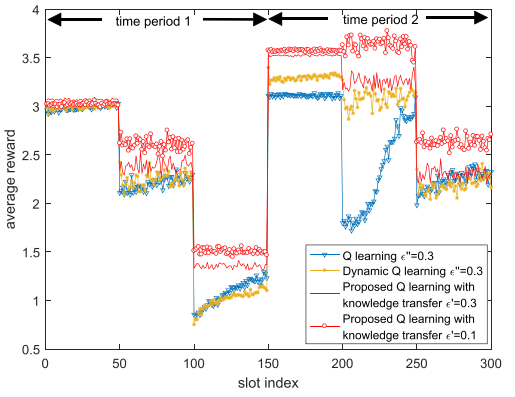


FIGURE 9. Performance comparison with context evolution. ($\varepsilon'' = 0.3$ and $\varepsilon'' = 0.1$).

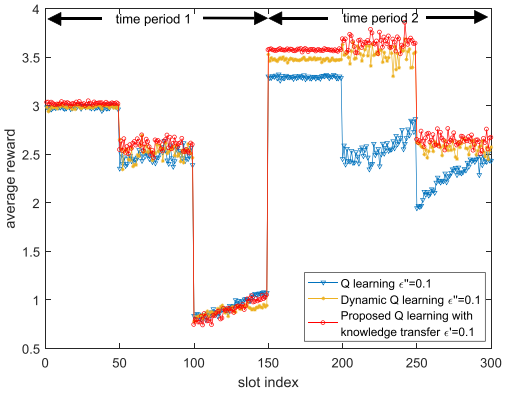


FIGURE 10. Performance comparison with context evolution. ($\varepsilon'' = 0.1$).

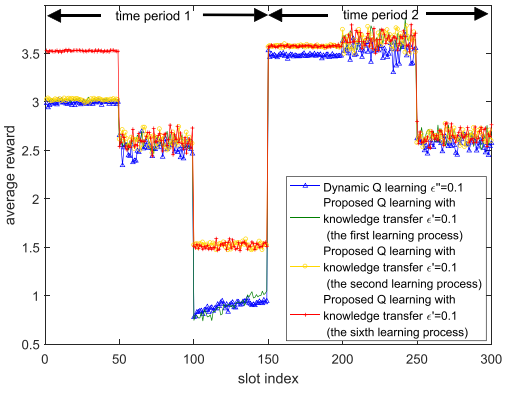


FIGURE 11. The advantage of learning experience accumulation.

proposed algorithm with $\varepsilon' = 0.3$ and $\varepsilon' = 0.1$ are both better than the other two algorithms in all contexts. However, when the exploration probabilities are 0.1 for all cases, the proposed algorithm seems to have the same performance with the other algorithms during the first time period (1st to 50th slots and 100th to 150th slots). We infer that the algorithm did not converge, which constrained its performance improvement. To verify this point, we run the proposed algorithm to further accumulate learning experience. That is, as the algorithm progresses, the latest Q-values will be updated in

the learning record database $\boldsymbol{D}$ (as stated in line 15 of the algorithm pseudocode). Fig. 11 shows that after the second learning process, the proposed algorithm achieves significant performance improvement during the range of 100th to 150th slots. It also shows performance gains during the range of the 1st to the 50th slots in the sixth learning process. The result confirms that the accumulation and reuse of learning experience in terms of Q-values could provide satisfactory outcomes, even with short learning time in scenarios with long convergence durations.

VII. CONCLUSION

In this paper, we studied the context-aware indoor network selection problem. We first formulated the network selection by differentiating the asymmetric downlink-uplink features of traffic requirements and network performance as a learning problem. On this basis, we exploited the time-location dependent load distribution to propose a reinforcement learning with knowledge transfer based algorithm. The simulation results revealed that the introduction of transfer learning could significantly improve both the convergence speed and performance of reinforcement learning based network algorithms.

REFERENCES

[1] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[2] M. Ayyash *et al.*, "Coexistence of WiFi and LiFi toward 5G: Concepts, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 64–71, Feb. 2016.

[3] A. Gupta and E. R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[4] H. Burchardt, N. Serafimovski, D. Tsonev, S. Videv, and H. Haas, "VLC: Beyond point-to-point communication," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 98–105, Jul. 2014.

[5] Z. Du, Q. Wu, P. Yang, Y. Xu, J. Wang, and Y.-D. Yao, "Exploiting user demand diversity in heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4142–4155, Aug. 2015.

[6] L. Torrey and J. Shavlik, "Transfer learning," *Handbook Of Research On Machine Learning Applications and Trends*. Hershey, PA, USA: IGI Global, 2009, ch. 11.

[7] F. Pantisano, M. Bennis, W. Saad, S. Valentin, and M. Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 4483–4488.

[8] F. Pantisano, M. Bennis, W. Saad, S. Valentin, M. Debbah, and A. Zappone, "Proactive user association in wireless small cell networks via collaborative filtering," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 1601–1605.

[9] Z. Du, Q. Wu, and P. Yang, "Dynamic user demand driven online network selection," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 419–422, Mar. 2014.

[10] Q. Wu, Z. Du, P. Yang, Y.-D. Yao, and J. Wang, "Traffic-aware online network selection in heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 381–397, Jan. 2016.

[11] M. Haddad, S. E. Elayoubi, E. Altman, and Z. Altman, "A hybrid approach for radio resource management in heterogeneous cognitive networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 831–842, Apr. 2011.

[12] Y. Xu, J. Wang, Q. Wu, Z. Du, L. Shen, and A. Anpalagan, "A game-theoretic perspective on self-organizing optimization for cognitive small cells," *IEEE Commun. Mag.*, vol. 53, no. 7, pp. 100–108, Jul. 2015.

[13] X. Bao *et al.*, "Protocol design and capacity analysis in hybrid network of visible light communication and OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1770–1778, May 2014.

[14] M. B. Rahaim, A. M. Vegni, and T. D. C. Little, "A hybrid radio frequency and broadcast visible light communication system," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Dec. 2011, pp. 792–796.

[15] X. Wu, D. Basnayaka, M. Safari, and H. Haas, "Two-stage access point selection for hybrid VLC and RF networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.

[16] F. Wang, Z. Wang, C. Qian, L. Dai, and Z. Yang, "MDP-based vertical handover scheme for indoor VLC-WiFi systems," in *Proc. Opto-Electron. Commun. Conf. (OECC)*, Shanghai, China, Jun./Jul. 2015, pp. 1–3.

[17] S. Liang, Y. Zhang, B. Fan, and H. Tian, "Multi-attribute vertical handover decision-making algorithm in a hybrid VLC-femto system," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1521–1524, Jul. 2017.

[18] *GigaIR, Infrared Data Association Standards*. Accessed: May 2018. [Online]. Available: http://irda.org

[19] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[20] D. A. Basnayaka and H. Haas, "Hybrid RF and VLC systems: Improving user data rate performance of VLC systems," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, May 2015, pp. 1–5.

[21] M. Kavehrad, "Sustainable energy-efficient wireless applications using light," *IEEE Commun. Mag.*, vol. 48, no. 12, pp. 66–73, Dec. 2010.

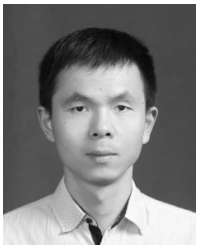
[22] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, May 1996.

[23] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

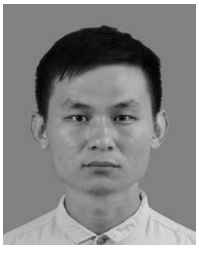
[24] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.

[25] E. Talvitie and S. Singh, "An experts algorithm for transfer learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 1065–1070.

[26] M. Ibrahim, K. Khawam, and S. Tohme, "Congestion games for distributed radio access selection in broadband networks," in *Proc. IEEE Global Telecommun. Conf. (GlobeComm)*, Dec. 2010, pp. 1–5.



ZHIYONG DU received the B.S. degree in electronic information engineering from the Wuhan University of Technology, Wuhan, China, in 2009, and the Ph.D. degree in communications and information systems from the College of Communications Engineering, Nanjing, China, in 2015. Since 2015, he has been an Assistant Professor with the National University of Defense Technology. His research interests include 5G, quality of experience, learning theory, and game theory.



CHUNXI WANG is currently pursuing the M.S. degree in signal and information processing with the National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China. His research interests include visible light communication and radio resource allocation.

YOUMING SUN received the B.S. degree in electronic and information engineering from Yanshan University, Qinhuangdao, China, in 2010, and the M.S. degree from the National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China, in 2013, where he is currently pursuing the Ph.D. degree in communications and information system. His research interests include resource allocation in small cell networks, cognitive radio networks, game theory, and statistical learning. He currently serves as a regular reviewer for many journals, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE SYSTEMS JOURNAL, *Wireless Networks*, IET Communications, and *KSII Transaction on Internet and Information Systems*. He has acted as a Technical Program Committees Member for the IEEE International Conference on Wireless Communications and Signal Processing 2015.

GUOFENG WU received the B.S., M.S., and Ph.D. degrees in signal and information processing from the National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China, in 2004, 2006, and 2011, respectively. He is currently an Associate Professor with the National Digital Switching System Engineering and Technological Research Center. His research interests include signal processing, visible light communication, and radio resource allocation.