Q-learning based Handover Algorithm for High-Speed Rail Wireless Communications

Siling Wang, Li Zhang
School of Electronic and Electrical Engineering
University of Leeds
Leeds, United Kingdom
Email: elswa@leeds.ac.uk, l.x.zhang@leeds.ac.uk

Abstract—**High-speed railways (HSRs) has become one of the most preferable modes of transportation. In the evolution of the railway wireless communication system from Long Term Evolution for Railway (LTE-R) to the 5th Generation Wireless System (5G), the rapid increase in the train speed and number of base stations along the railway track led to challenging handover (HO) problems, such as high failure rate and frequent HOs. In order to address this challenge, an improved handover decision strategy is proposed based on Q-learning algorithm. The simulation results demonstrate that our proposed scheme is capable of reducing the number of unnecessary handover and improving the network performance remarkably.**

Index Terms—**Handover, Mobility, High-speed Railway, Q-learning, 5G**

I. Introduction

In recent years, HSRs offers transport that runs considerably fast in excess of 300km/h for the new lines and 200km/h for the existing ones. HSRs can immensely increase mobility and improve customer satisfaction [1]. With the booming development of HSRs, the evolution of high speed rail wireless communications system from LTE-R to 5G is essential for the reliable safety-related communications and seamless mobile connectivity of passengers, but also face new challenges, one of which is the handover problem. The handover mechanism involves transferring an ongoing call or data session from one cell to another cell while maintaining the quality of service (QoS). In the next generation wireless cellular networks, the HO process has even more significant impact on the network performance. For instance, the number of HOs between different cells will increase in the 5G network because of the smaller coverage area of the cells. Frequent HOs causes disruption of data service, higher signalling overhead and decreased network throughput. Moreover, most of the high-speed trains have metal bodies with large windows of a single layer or multi-layer glasses, therefore, with the increase of frequency, signals which go through the compartments will suffer greater attenuation. Additionally, handover failure occurs if the user equipments (UEs) fail to connect to the target cell. The probability of handover failures increase with the higher moving speed. Thus, the objective of this paper is to propose a novel HO scheme that can reduce the number of HOs and enhance the HO success probability without compromising the performance of the communications in HSR.

A. Related Work

There already exist several different schemes which solve the HO problems mentioned above. Work in [2] adopts mobile relays (MRs) to decrease the signalling load caused by the large number of HO procedures triggered by the UEs in the same carriage at the same time. Further, authors of [3] adopt a distributed antenna system (DAS) which combines with the two-hop architecture, and proposed a fast predictive HO algorithm to reduce the HO latency and handover command failure probability. In [4], type-2 fuzzy-based HO optimization for advanced long term evolution (LTE-A) network was proposed, with the aim to reduce call drops and number of handover. In the future, machine learning will play a crucial role in many innovative 5G technology and architectures [5]. It is an algorithm which has a self-learning ability to the systems and improve the performance from its training experience. Machine learning based algorithms are required to accurately apply HO decision which enable the system to automatically adjust parameters according to the users' demand and requirements. As a result, frequent HO can be reduced efficiently [6]. The authors in [7] introduce a HO optimization algorithm based on reinforcement learning (RL) for 5G system. They model the HO problem as a contextual multi-armed bandit problem and then solve it by using Q-learning algorithm, in order to improve the performance of the link beam. In [8], the authors present a solution based on RL to select the optimal base station (BS) for proactive decision HO in Millimeter-wave (mmWave) wireless communication. Their results illustrate that intelligent self-learning agent can decrease the number of HOs. A self-learning framework is proposed in [9], which jointly optimizes HO problem and beamforming for mmWave network. RL algorithm is exploited to determine the optimal backup BSs along user trajectory which can help reduce the overhead signalling during channel estimation and minimize the number of HOs. In [10], the authors introduce a RL based mobility model for drone UEs. This model provides a flexible HO decision making for a given flight trajectory while trading off the number of HOs and received signal power to decrease the frequent HOs. Optimization of HO by jointly considering AHP-TOPSIS and RL algorithm is discussed in [11], where optimal BS is selected by using AHP-TOPSIS, then RL approach is utilized to get a proper setting of triggering point such as time to trigger (TTT) and hysteresis according to the different speed of users for LTE-Advance network. The results show that the proposed scheme can minimize the handover failure rate and ping pong effect. The algorithms in [12] and [13] present methods to decrease the link failure during HO process. An intelligent handover scheme based on Elman neural network in HSR scenario is proposed in [12], where the authors correlate past measurement parameters such as reference signal received power (RSRP) and reference signal received quality (RSRQ) with future handover decisions, which can accelerate the HO execution and optimize the HO process. In [13], fuzzy Q-learning is used to optimize two contradictory handover problems, which are radio link failure (RLF) and ping-pong effect for LTE network, the aim is to minimize RLF for real time users and reduce ping pong effect for non-real time users while keeping RLF within acceptable limits. Although the HO problems have been widely investigated, the problem of HO optimization using machine learning algorithm for HSR still remains an open problem. Some of them consider that RSRP measurements, HO success probability, etc. as rewards. In fact, different reward configurations such as signal to interference plus noise ratio (SINR), train speed and downlink-throughput, etc. can also be considered in order to achieve the optimised HO decision in 5G HSR system.

B. Contribution

In this paper, we develop an efficient HO scheme based on RL for 5G high-speed rail wireless communication system. In a dense 5G high-speed rail deployment, HO decisions can be optimized by utilizing Q-learning algorithm in order to reduce the unnecessary HOs and improve the network performance. In addition, the criteria is selected considering the high-speed rail movement. The proposed scheme aims to reduce the HO rate and achieve better throughput by jointly considering four parameters including HO cost, SINR, RSRP and time of stay (ToS) as rewards. By using Q-learning algorithm, the users' current decision is linked to the long-term benefit to improve the HO decision, and the effectiveness of this method is demonstrated using simulation results.

The rest of this paper is organized as follows. In section II, the network architecture model and usage scenarios are introduced. Section III presents a brief introduction of the RL and the proposed Q-learning based HO scheme. Simulation results are presented and discussed in section IV. Finally, section V concludes this paper.

II. System model

A. HSR Network Architecture

In [14], the challenges of high speed railway communications are to maintain consistent user experience, and reliable and seamless communication. In high-speed rail system, dedicated BS deployment along railway line is considered, which is illustrated in Fig. 1. The network is covered with macro BS with carrier frequency 4GHz. The distance between BS and railway track is 100m. Passenger UEs are located in train carriages. For the passenger UEs, all information is collected by the train relay station (TRS) which is placed on the top of the train. The TRS acts as a single big user to execute HO from the serving BS to the target BS to reduce the HO signalling overhead and avoid the penetration loss caused by train. The passengers can get access to the internet via WiFi or access points which are connected to the train relay by wired link.

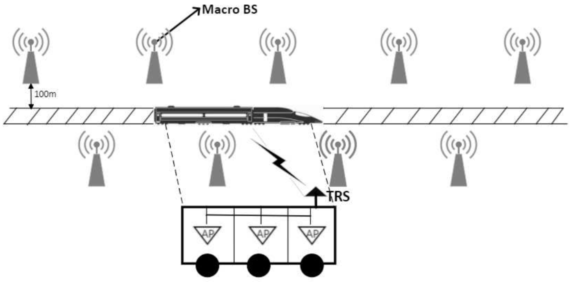


Fig. 1. System architecture

B. Channel Model

In this paper, only the main path signal is considered since high-speed train runs through rural or viaduct areas most of the time, so multi-path effect can be ignored. At time *t*, let a high-speed train be located at position *x* in the *i*-th cell. The received signal power as well as RSRP from the cell *i* is:

$$RSRP(i,x)[\text{dBm}] = Pt[\text{dBm}] - PL\left(i,d_i\right)[\text{dB}] - \zeta\left(0,\sigma_i\right)[\text{dB}] \tag{1}$$

where *Pt* is the transmission power of the *i*-th cell. $PL(d) = A \cdot d^{-\gamma}$ is the path loss, where A is a constant, d is the distance, γ is the path loss exponent. $\zeta\left(0,\sigma_i\right)$ is a Gaussian distributed random variable with a zero mean and a standard deviation to describe the shadow fading.

The pathloss model is proposed in [15][16], which is applicable to HSR scenario. Thus, the path loss (in decibels) of the *i*-th cell can be expressed as

$$PL\left(i,d_i\right)[\text{dB}] = 10\theta\log\left(d_i\right) + PL\left(d_0\right), \tag{2}$$

where θ is the path loss exponent, d_i is the distance between the BS and mobile train, $PL(d_0)$ is the intercept. Path loss depends on the distance between the transmitter and receiver.

We consider inter-cell interference caused by the two nearest neighboring cells and express it as

$$I_i[mW] = \sum_{n=1,n\neq i}^{N_{nei.i}} 10^{RSRP(n,x)/10}, \tag{3}$$

where $N_{nei.i}$ is the number of the co-channel cells. $RSRP(n,x)$ obtained from (1) is the received interference signal power from the co-channel cells.

Then, we can obtain the SINR of the *i*-th cell as

$$SINR_i[dB] = 10\log\left[\frac{10^{RSRP(i,x)/10}}{I_i + N_0}\right], \tag{4}$$

where N_0 is the noise power.

III. Proposed Handover Scheme

In this section, we first introduce the Q-learning algorithm in the RL framework briefly, and then a Q-learning based HO algorithm is designed to solve the frequent HO problem while maintaining the reliable connection.

A. Background of Reinforcement Learning

RL is a type of machine learning which can make an agent learn in an interactive environment by a great number of trials and errors utilizing feedback from its own actions and experiences. It is about taking suitable action to maximize the reward in a particular situation.

RL is different from the supervised learning. Specifically, in the supervised learning, the training data has the answer key with it so the model is trained with the correct answer itself whereas in RL, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience. As compared to unsupervised learning, RL differs from it in terms of their respective goals. The goal in the unsupervised learning is to find similarities and differences between data points. While in the RL, the goal is to find a suitable action model which can maximize the total cumulative reward of the agent. The relationship between agent, action and environment is shown in Fig. 2. The agent learns the best policy by multiple interactions with the environment.

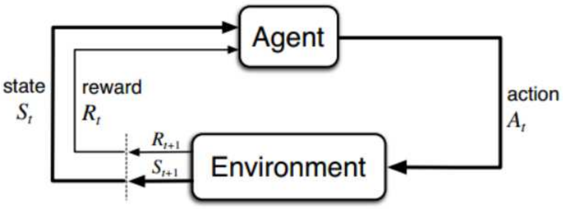


Fig. 2. Framework of RL

Markov Decision Processes (MDPs) are mathematical frameworks which are often used to describe an environment in RL and almost all RL problems can be formulated using MDPs. An MDP usually consists of a set of finite environment states *S*, a set of possible actions *A*, a real valued reward function *R* and a transition model *P*. The learning procedure is detailed as follows: we first define the main elements of RL. At time *t*, the agent observes the state of the environment, $s_t \in S$, where *S* is the set of possible states. After observing state s_t, agent will take an action, $a_t \in A(s_t)$ where $A(s_t)$ is the set of possible actions at state s_t. After selecting the action a_t from state s_t, agent will receive the immediate reward R_{t+1} from the state-action pair (s_t, a_t). The selected action in state s_t moves the agent to state s_{t+1} at time *t*+1. It is crucial for the environment to have state dynamics such that $P(s_{t+1} \mid s_t, a_t)$ exists. This is the probability that the action a_t in state s_t at time *t* will lead to new state s_{t+1} at time *t*+1. Specifically, the probability that the process moves into its new state s_{t+1} is impacted by the chosen action.

However, real world environment is more likely to lack any prior knowledge of environment dynamics. Model-free RL method such as Q-learning, is more suitable. Since Q-learning is a model-free RL algorithm to learn the value of an action in a particular state. It does not require a model of the environment. For any finite MDP, Q-learning can find an optimal policy to maximise the expected value of the total reward over all successive steps, starting from the current state, and it can identify an optimal action in a given situation based on trial and error. The model is composed of a set of states, agent and set of actions per state. An action is performed by an agent by moving from one state to another state. The reward is provided to the agent after executing the action in the specific state. The maximum total reward is the main goal of the agent by learning the optimal action at each state.

B. Q-learning Based Handover Algorithm

In this paper, we regard the high-speed railway 5G wireless communication system as a reinforcement agent and consider the physical network environment of the HO area of two adjacent BSs as the external environment, so the interaction process between the RL agent and the external environment can be taken as a MDP. The MDP can be represented by the following four tuples: $\langle S, A, R, \pi\rangle$, which are defined as follows:

1) State: The state explains the current condition of the network environment and decides what will occur next. For our method, the state is represented by *S* consisting of the train's location and current serving base station.

2) Action: In the conventional HO events such as A3, the users will always select the target BS with the highest RSRP or SINR according to the HO trigger condition resulting in the sub-optimal decision and frequent HO problem. Hence, we define the action $a \in A(s)$ as the scalar representation of the serving BS at current state *s*. The action space $A(s)$ includes all BSs along with the train railway track. For instance, choosing an action is analogous to select an appropriate BS to handover or stay with the current BS. The aim is to select an action in a given state that maximizes the expected reward.

3) Reward Design: The reward *R* is a measurement of the rationality of the agent's behavior in a specific environment state. The selection of the reward will impact on the learning efficiency of the agent and reflect the environment feedback in RL. The significance of reward is to motivate the agent to learn to reach the target through reward maximization, and our objective is to reduce the HO number while maximizing the throughput. In a handover process, the reward is closely relevant to handover performance and network performance. Thus, we design the immediate reward so that multiples parameters which affect the HO performance can be considered and it can be expressed as follows:

$$R = -\omega 1 \times HOcost + \omega 2 \times SINR + \omega 3 \times RSRP + \omega 4 \times ToS \tag{5}$$

where $\omega 1$, $\omega 2$, $\omega 3$ and $\omega 4$ represent the weights for the HO cost, SINR, RSRP and ToS respectively. In addition, their range is from [0,1] and $\omega 1 + \omega 2 + \omega 3 + \omega 4 = 1$. Based on RL, we can find out the best state-action pair with the maximum *R*.

The reward function is a weighted combination of four parameters: HO cost, RSRP, ToS and SINR. The definition of RSRP and SINR can be found in the section II. HO cost is used to describe whether a HO occurs between the current and the next state. For example, HO cost=0 if there is no HO occurring from the current state to the next state, and vice versa. In terms of ToS, it is the expected time that the users will stay in the coverage area of a BS. When it combines with the train speed, this parameter can be used to decrease the unnecessary HO number. If the train is always connected to the BS with the best received signal without considering how long the users stay in that cell, it will result in frequent HOs, and consequently also increase overheads. Instead, if the time of stay in one cell is too short, then the handover to this cell can be avoided.

4) Policy: In a given HO route, the policy π can be expressed as $Q(s, a)$ which is a value function corresponding to different actions *A* in all states *S*. In the proposed scheme, we select the ε-greedy algorithm as the policy π, where ε means the probability that the agent chooses the next state *s'*. Its range is from 0 to 1. This algorithm means that the agent will select the state with the maximum reward *R* by the probability of 1-ε, and will randomly select the next state by the probability of ε. The update rule of the policy is given as

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q\left(s', a'\right) - Q(s,a) \right] \tag{6}$$

where, α is the learning rate, *r* is the reward and γ is the discount factor which is used to indicate the influence of the next action and state on the current state. Its value is [0,1]. The ε-greedy algorithm for Q-learning based HO is described in Algorithm 1. A Q-table is created by trying random actions for the received states and recording the reward observed. After a great number of iterations, the RL converges and the state with the largest *R* value selected by the agent will remain stable. In other words, the largest *Q* value stored in Q-table describes the optimal selection. Therefore, we can find the optimal HO decision in a given railway trajectory by selecting the largest *Q* value at each state.

IV. Results and Discussion

This section evaluates the performance of the proposed Q-learning based algorithm. We first describe the experiment set-up and then present the simulation results with discussions.

A. Experiment Setup

In order to evaluate the handover performance of the proposed algorithm in the HSR system, we set up a HSR communication network model in MATLAB. In this model, the overall length of the railway track is 10km and ten BSs are distributed alternately along both sides of the track. Based on analysis in section III, we build a *Q* table during the training phase in the Q-learning algorithm (Algorithm 1). In this phase, agent will take random actions (HO to certain BS) based on ε-greedy exploration and then update the *Q* table. After the training phase, the table we obtained can be used in HO decision making phase.

In addition, the whole track is divided into 100 waypoints evenly and the train will take steps in a straight direction along the railway track. In each step, the measurement reports will be sent to the agent which exploits the *Q* table to take an action that maximizes the *Q* values for each state.

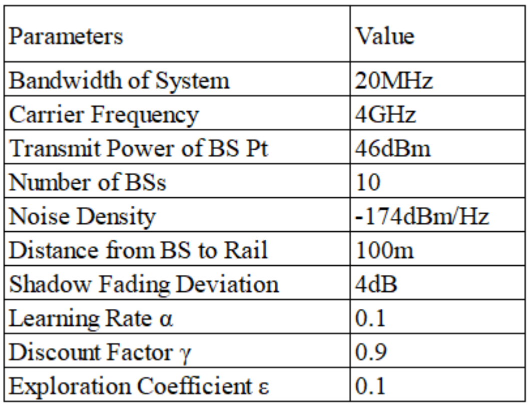
Algorithm 1 Q-learning with ε-greedy algorithm for HO situation

```
Initialization: An initial value $Q(s,a)$ with random rewards
for $n \leftarrow 0$ to number of episodes do
    $t \leftarrow 0$, Observe $s_0$, $s_{initial}$=$s_0$ //At time_step 0
    for $t < max\_step$ do
        Take a random variable g uniformly from [0,1]
        if $g < \varepsilon$ then
            Take a random action $a_t$ uniformly from set $A(s)$
        else
            Take action $a_t = \text{argmax}_{a \in A(s)} Q^*(s,a)$
        end if
        Observe $s_{t+1}$ and $r$
        Update the $Q$ value function as:
        $Q(s_t, a_t) \leftarrow Q(s_t, a_t) +$
        $\alpha \left[ r + \gamma \max_{a \in A(s)} Q(s_{t+1}, a) - Q(s_t, a_t) \right]$
    end for
end for
```

B. Simulation Results

The performance of the proposed HO scheme is evaluated in terms of the average number of HOs, HO success probability, throughput and HO latency, and the performance is the average of 1000 runs. The parameters used in the simulation are listed in Table I.

TABLE I
Simulation Parameters



Parameters	Value
Bandwidth of System	20MHz
Carrier Frequency	4GHz
Transmit Power of BS Pt	46dBm
Number of BSs	10
Noise Density	-174dBm/Hz
Distance from BS to Rail	100m
Shadow Fading Deviation	4dB
Learning Rate α	0.1
Discount Factor γ	0.9
Exploration Coefficient ε	0.1

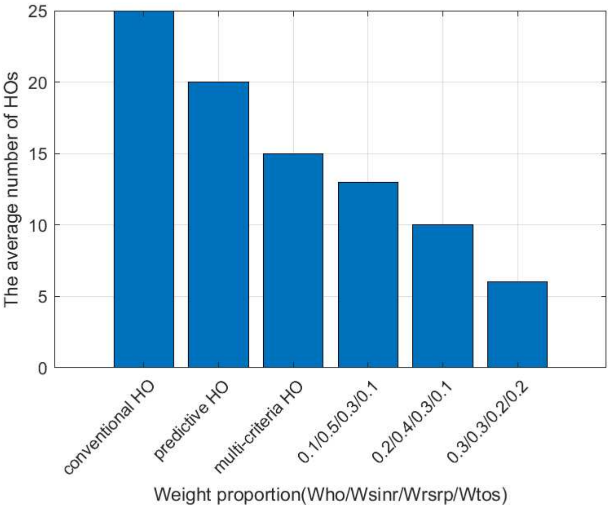


Fig. 3. Average number of HOs for different weight proportion

In Fig. 3, we compare the average number of HOs of the proposed scheme with different weight combinations considering four criteria including HO cost, SINR, RSRP and ToS with the conventional HO method and the existing algorithms [3] [17]. The conventional HO scheme considers that the train users always connect to the BS with the strongest RSRP. From the figure, it can be clearly observed that the proposed scheme outperforms the existing algorithms as well as the conventional method and achieves the lowest HO rate. The proposed scheme can be decreased by 48% when weight proportion $\omega 1/\omega 2/\omega 3/\omega 4$=0.1/0.5/0.3/0.1, compared to the conventional scheme. Overall, the simulation result clearly indicates that the proposed scheme is effective in reducing the number of unnecessary handovers.

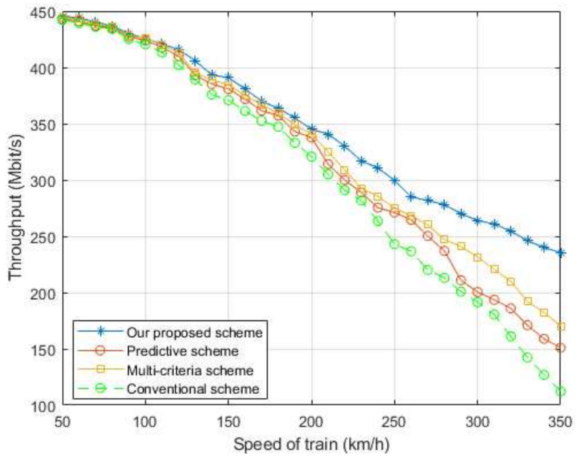


Fig. 4. Throughput for HSR system

The throughput vs speed of train is shown in fig. 4. The proposed method clearly outperforms the conventional method and other algorithms [3][17] in the literature. Specifically, when the speed of train is 300 km/h, the throughput of the proposed scheme is increased by about 12.2%, 23.9% and 27.4% respectively in comparison to multi-criteria, predictive HO scheme and conventional HO scheme. The reason is that the proposed method can effectively reduce the HO number, especially when the train moves faster.

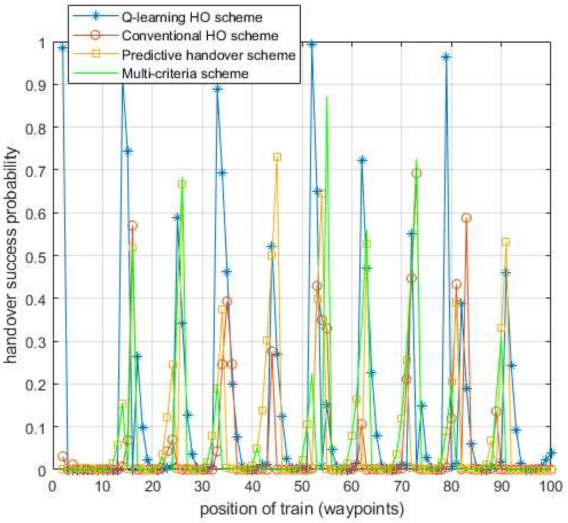


Fig. 5. Handover success probability

Fig. 5 illustrates the handover success probability at different train location (waypoints) for different HO methods. The proposed scheme achieves the best performance at most of the train locations when compared with the conventional HO method and the existing HO algorithms [3][17]. For those locations with a success probability of 0, that means no HO occurring.

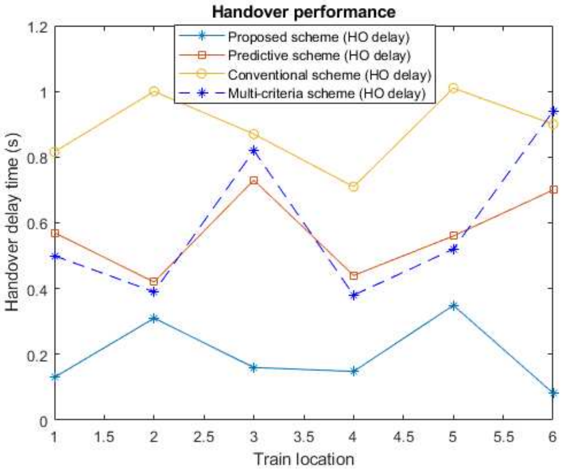


Fig. 6. Handover latency

Handover latency is another decisive performance metric when it comes to QoS. The handover latency time can be defined as

$$T = Psuccess \times Ts + (1 - Psuccess) \times Trec \qquad (7)$$

where *Ts* and *Trec* are the handover latency in case of a successful and failure situations respectively. Note that the calculation of *Ts* and *Trec* is based on [18].

In fig. 6, the abscissa is the train locations which we pick along the whole train railway track and we use the numerical values to represent for convenience. It is clear that the proposed scheme has the lowest HO latency time and also achieves the best user QoS when compared with the conventional HO scheme and other existing schemes [3][17].

V. Conclusion

This paper has investigated the HO problems in HSR wireless communication system. A novel HO scheme based on Q-learning algorithm is proposed to achieve better HO performance. This paper establishes a Q-learning framework for the HSR HO, and provides an effective method for optimal HO decision making which considers multiple criterias as weighted reward along the railway track. As demonstrated by the simulation results, the proposed scheme outperforms the conventional approach and existing algorithms in terms of throughput, HO number, HO success probability and HO latency. In the future, we plan to further design an advanced HO strategy using deep learning in 5G ultra-dense mobile heterogeneous network over dual connectivity for high-speed railway communication.

References

[1] W. Gheth, K. Rabie, B. Adebisi, M. Ijaz and G. Harris, ”Communication systems of high-speed railway: A survey”, Transactions on Emerging Telecommunications Technologies, vol. 32, no. 4, 2021.

[2] M. Pan, T. Lin and W. Chen, ”An Enhanced Handover Scheme for Mobile Relays in LTE-A High-Speed Rail Networks”, IEEE Transactions on Vehicular Technology, vol. 64, no. 2, pp. 743-756, 2015.

[3] W. Ali, J. Wang, H. Zhu and J. Wang, ”An Expedited Predictive Distributed Antenna System Based Handover Scheme for High-Speed Railway,” GLOBECOM 2017 - 2017 IEEE Global Communications Conference, 2017, pp. 1-6.

[4] M. Saeed, H. Kamal and M. El-Ghoneimy, ”Novel type-2 fuzzy logic technique for handover problems in a heterogeneous network”, Engineering Optimization, vol. 50, no. 9, pp. 1533-1543, 2017.

[5] Sönmez, Ş., Shayea, I., Khan, S. A., Alhammadi, A, ”Handover management for next-generation wireless networks: A brief overview”, 2020 IEEE Microwave Theory and Techniques in Wireless Communications (MTTW). Vol. 1, IEEE, 2020.

[6] S. Khan, I. Shayea, M. Ergen and H. Mohamad, ”Handover management over dual connectivity in 5G technology with future ultra-dense mobile heterogeneous networks: A review”, Engineering Science and Technology, an International Journal, p. 101172, 2022. Available: 10.1016/j.jestch.2022.101172.

[7] V. Yajnanarayana, H. Rydén and L. Hévizi, ”5G Handover using Reinforcement Learning,” 2020 IEEE 3rd 5G World Forum (5GWF), 2020, pp. 349-354.

[8] M. S. Mollel, S. Kaijage, M. Kisangiri, M. A. Imran and Q. H. Abbasi, ”Multi-User Position Based on Trajectories-Aware Handover Strategy for Base Station Selection with Multi-Agent Learning,” 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 2020, pp. 1-6.

[9] S. Khosravi, H. Shokri-Ghadikolaei and M. Petrova, ”Learning-Based Handover in Mobile Millimeter-Wave Networks”, IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 2, pp. 663-674, 2021.

[10] Y. Chen, X. Lin, T. Khan and M. Mozaffari, ”Efficient Drone Mobility Support Using Reinforcement Learning,” 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020, pp. 1-6.

[11] T. Goyal and S. Kaushal, ”Handover optimization scheme for LTE-Advance networks based on AHP-TOPSIS and Q-learning”, Computer Communications, vol. 133, pp. 67-76, 2019.

[12] D. Li, D. Li and Y. Xu, ”Machine Learning Based Handover Performance Improvement for LTE-R,” 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), 2019, pp. 1-2.

[13] R. Hegazy, O. Nasr and H. Kamal, ”Optimization of user behavior based handover using fuzzy Q-learning for LTE networks”, Wireless Networks, vol. 24, no. 2, pp. 481-495, 2016.

[14] F. Hasegawa et al., ”High-Speed Train Communications Standardization in 3GPP 5G NR”, IEEE Communications Standards Magazine, vol. 2, no. 1, pp. 44-52, 2018.

[15] B. Ai, A. Molisch, M. Rupp and Z. Zhong, ”5G Key Technologies for Smart Railways”, Proceedings of the IEEE, vol. 108, no. 6, pp. 856-893, 2020.

[16] C. Wu, X. Cai, J. Sheng, Z. Tang, B. Ai and Y. Wang, ”Parameter Adaptation and Situation Awareness of LTE-R Handover for High-Speed Railway Communication”, IEEE Transactions on Intelligent Transportation Systems, pp. 1-15, 2020.

[17] S. Wang and L. Zhang, ”A Multi-Criteria Handover Scheme for Distributed Antenna System Based High-Speed Rail Wireless Communications,” 2021 24th International Symposium on Wireless Personal Multimedia Communications (WPMC). IEEE, 2021.

[18] 3GPP TR 36.881, “Evolved Universal Terrestrial Radio Access (E-UTRA); Study on latency reduction techniques for LTE,” 3rd Generation Partnership Project (3GPP), Sophia-Antipolis, France, June. 2016.