Publication metadata

Title: Trajectory Planning of Multiple Dronecells in Vehicular Networks: A Reinforcement Learning Approach

Author(s): Moataz Samir, Dariush Ebrahimi, Chadi Assi, Sanaa Sharafeddine, Ali Ghrayeb

Journal: IEEE Networking Letters

DOI/Link: https://doi.org/10.1109/LNET.2020.2966976

How to cite this post-print from LAUR:


© Year 2020

“© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

This Open Access post-print is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives (CC-BY-NC-ND 4.0)

This paper is posted at LAU Repository

For more information, please contact: archives@lau.edu.lb
Trajectory Planning of Multiple Dronecells in Vehicular Networks: A Reinforcement Learning Approach

Moataz Samir¹, Darioush Ebrahimi², Chadi Assi¹, Sanaa Sharafeddine³, and Ali Ghrayeb⁴

¹Concordia University, ²Lakehead University, ³Lebanese American University, ⁴Texas A&M University at Qatar

Abstract—The agility of unmanned aerial vehicles (UAVs) have been recently harnessed in developing potential solutions that provide seamless coverage for vehicles in areas with poor cellular infrastructure. In this paper, multiple UAVs are deployed to provide the needed cellular coverage to vehicles traveling with random speeds over a given highway segment. This work minimizes the number of deployed UAVs and optimizes their trajectories to offer prevalent communication coverage to all vehicles crossing the highway segment while saving energy consumption of the UAVs. Due to varying traffic conditions on the highway, a reinforcement learning approach is utilized to govern the number of needed UAVs and their trajectories to serve the existing and newly arriving vehicles. Numerical results demonstrate the effectiveness of the proposed design and show that during the mission time, a minimum number of UAVs adapt their velocities in order to cover the vehicles.

I. INTRODUCTION

The Internet-of-Vehicles (IoV) paradigm is expected to play a crucial role in the automotive industry by creating innovative services and revenue sources. By 2025, around quarter of a billion of cars navigating along global roads will be equipped with wireless communication capabilities [1]. It is, therefore, indispensable to provide them with continuous and sustainable network connectivity particularly for vehicles running critical services and/or traveling in areas with poor cellular infrastructure. For instance, when the cellular infrastructure is either inadequately provisioned to cover all vehicles or has been subjected to unexpected hardware failure or direct damage. Moreover, in some situations such as in rural areas and in volatile environments, the deployment of terrestrial infrastructures is economically infeasible and challenging due to CAPEX and OPEX. The consequence of unavailable communication services in highways may turn devastating, especially in an emergency situation, or guidance. Mobile base stations such as unmanned aerial vehicles (UAVs), owing to their agility and mobility, are being promoted as a promising paradigm to provide network connectivity when the infrastructure is partially or fully unavailable.

UAVs can be relocated from one zone to another based on the actual traffic demand and can provide network connectivity in situations with damaged terrestrial networks or lack of network infrastructure due to geographical constraints. In vehicular networks which are typically characterized by high mobility and varying vehicle arrival pattern, multiple UAVs with autonomous control are required to cooperatively provide network coverage and adapt to current traffic conditions. Many researchers in the literature have investigated the UAVs’ placement and trajectory. For instance, the authors in [2] jointly optimized the trajectory, multi-user scheduling and power control for UAVs to maximize the minimum rate of ground users. In [3], the authors optimized the UAV’s trajectory to minimize the time to completely disseminate a common file to a number of distributed ground terminals. The energy trade-off in ground-to-UAV communication via trajectory design is studied in [4]. In [5], the authors proposed a matching theory for the allocation of the deployed UAVs to the respective solar-powered charging stations to achieve the highest possible profit in term of energy, communication, and safety. Throughput-access delay tradeoff has been identified for a single UAV communication, however, access delay may be significantly reduced by employing multiple UAVs. Motivated by this, deployment of a swarm of UAVs is required to serve vehicles [6]. However, none of the existing literature provides a practical solution for a highly dynamic environment such as the vehicular network, where the network’s topology frequently changes. Unlike [7] where the dynamic nature of the network is considered but a global knowledge of the environment is assumed to be known; in this work, we assume no knowledge about the environment. Furthermore, we consider the energy consumption of the deployed UAVs, which represents an additional challenge to our problem. Hence, to the best of our knowledge, optimizing the number of UAVs and their trajectories that serve vehicles in a high-mobility network setting remains an unaddressed problem.

To this end, in this paper, we propose a reinforcement learning (RL) framework to explore the unknown environment step by step, and plan trajectories for a minimum number of UAVs, to provide network connectivity for vehicles while minimizing the energy consumption of UAVs. Note that empowering the UAVs with artificial intelligence will play an inevitable role in 5G/B5G. Thus, the existence of intelligent UAVs to provide wireless connectivity will be realized as expected. To achieve that, a central agent in the external network is trained to observe the environment and then control the decision of realizing the minimum number of UAVs and their trajectories to provide effective communication coverage. This task is quite challenging because UAVs have limited energy and cannot fly indefinitely. So, the UAVs should fly in an energy-efficient manner during the coverage process, and at the same time adopt to the aforementioned changes in the network topology.

II. COMMUNICATION SCENARIO

We consider (Fig. 1) a highway segment with a communication infrastructure that is either damaged or non existent. Furthermore, this segment has unidirectional traffic flow of vehicles that depart the coverage of a fixed base-station, where a set $\mathcal{M}$ of $M$ UAVs, indexed by $m = 1,...,M$, is intended to serve as mobile base stations for vehicles crossing the highway segment. The UAVs are assumed to have high
capacity fronthaul links such as free space optics (FSO) or millimeter-wave (mmWave) links with ingress ground base station (GBS), where a central unit with RL-agent resides. The RL-agent observes the dynamic vehicular environment and steadily learn the optimal trajectory policy and manages the cooperation between the deployed UAVs. Therefore, a vehicle that cannot be covered while being within the coverage of one will be covered by other deployed UAVs. We consider a system with multiple time frames, where each frame (of duration $T$) is divided into $N$ equal time slots, each with length $\delta_t$, indexed by $n = 1, \ldots, N$. All UAVs are assumed to fly at a constant altitude $H$ above ground level. We use $\mathcal{V}^n$ to denote the subset of vehicles to be served at time-slot $n$. We use $\mathcal{V}^n \cup V_{\infty}$ to denote the subset of vehicles to be served in time slot $n$, where $V = \mathcal{V}^1 \cup \mathcal{V}^2 \cup \mathcal{V}^n \cup \mathcal{V}^N$. Each vehicle periodically broadcasts on the control channel (CCH) announcement beacon messages, which contain information about the speed and location. The deployed UAVs monitors the CCH, coordinates with the control unit, and then simply switches to a service channel (SCH) to establish a communication link. We assume each UAV can simultaneously communicate with multiple vehicles within its coverage on different spectrums by allocating appropriate orthogonal resources to ensure interference-free communication. The interference-free model is widely used in the literature. Furthermore, the Doppler Effect caused by the UAV and vehicles mobility is assumed to be well compensated. Furthermore, the control-unit manages the synchronization issues to/from the deployed UAVs. In practice the spectrum for distant deployed UAVs can be reused during the covering mission. Thus, we assume the vehicles are served when they lie within the coverage of any UAV without interference from other UAVs, and henceforth our study is concerned with dealing with the UAVs coverage issue. We also assume the UAV communication range is fixed, where vehicle $i$ is considered to be covered if the vehicle is within the UAV communication range, $R_c$, projected on the ground. In fact, a vehicle $i$ is considered to be covered by a UAV with an acceptable (QoS) if the instantaneous achievable rate served by UAV $m$ at time-slot $n$ is greater than a threshold value $r_{min}$. Therefore, vehicle $i$ is considered covered by UAV $m$ with an acceptable QoS, in time-slot $n$ if $|x^n_i - w^n_m| \leq R_c$, where $x^n_i$ and $w^n_m$ are the instantaneous positions of vehicle $i$ and ground level position of UAV $m$, respectively. The communication range of UAVs, $R_c$, to achieve at least a minimum rate $r_{min}$ is given by

$$R_c = \left(\frac{r_{min}}{2 \nu_{\text{max}}} - 1\right)^{-1} \frac{P h_0}{\sigma^2} - H^2 \right)^{0.5},$$

where $h_0$ is the channel power at reference distance $d_0 = 1$ m. $P$ is the vehicle transmit power. $\sigma^2 = B_n N_o$, with $N_o$ denoting the power spectral density of the additive white Gaussian noise at the receivers for the bandwidth each vehicle obtains $B_n$ at time-slot $n$. $\nu_{\text{max}}$ is the variable vector size, we assume the maximum expected number of vehicles within the coverage of highway segment is $U$. The number of vehicles present within the highway segment follows the Poisson distribution [10].

The MDP of the covering problem is defined as follows.

$$P(\omega)_{\text{total}} = \int_0^T \frac{1}{\bar{w}_b} P(\omega) d\omega = P(\omega)_{\text{total}}.$$
At the beginning of each time-slot $n$ (or state $s_n$), the network state is characterized by $(E_n^m, V^n, x_i^n, C_i^n, C_i^m)$ where:

1. $E_n^m$ is a vector of size $M$ containing the remaining energy of each UAV at time-slot $n$, where $0 \leq E_n^m \leq E_{total}$.
2. $V^n$ is the number of vehicles residing within the considered highway segment.
3. $x_i^n$ is a vector of size $U$ containing the instantaneous position of vehicle $i \in (1, 2, .., V^n)$.
4. $w_i^n$ is a vector of size $M$ containing the ground location position of each UAV.
5. $C_i^m$ is a vector of size $U$ containing the coverage indicators of each vehicle. If $C_i^m = 1$, vehicle $i$ lies within the coverage of a UAV; otherwise, $C_i^m = 0$.

At each step, each UAV $m$ carries out an action $a_i^n$ which represents a distance to travel $d_m^n$ in a specific direction $\Phi_m^n$, depending on its current state. The UAVs may travel with arbitrary velocities in different directions, which makes the problem non-trivial to be solved. However, by assuming that the width of the highway lane is ignored as compared to the transmission range of vehicles and UAVs [11], the model may be simplified to as few as two directions (left and right). Furthermore, by assuming each UAV travels at a constant speed, the model may be simplified to a finite travel distance (traveling with a fixed speed in each direction, maintaining hovering and non-serving). Hence, at time-slot $n$, each UAV chooses its action (distance and direction), and accordingly, the vehicular network environment pays an immediate reward; that is, a scalar value that reflects the righteousness of the UAVs’ actions. The immediate reward $r_n$ is the sum of the following normalized quantities: 1) Penalty incurred on network due to the existence of a vehicle within the highway without UAVs’ coverage: the value of this penalty is a normalized quantity proportional to the coverage indicator of each vehicle. As such, the UAVs are encouraged to cover the vehicles within the considered highway segment. Recall that a vehicle communicates its exit point upon its arrival to the highway, and the UAVs should coordinate to continuously cover that vehicle within the highway. The penalty due to non-coverage can be written as $P^n_c = \eta \sum_{i=1}^{V^n} (1 - C_i^n)$, where $\eta$ is weight with a high value to avoid UAVs missing to cover a vehicle. 2) Penalty incurred on network due to the deployment of a new UAV: the network receives this penalty when the current deployed UAVs cannot cover the newly arrived vehicles and a new UAV is required to be deployed. The value of this penalty is proportional to the number of deployed UAVs. As a result, the network learns to optimize the trajectories of the minimum number of UAVs to cover the current and newly arrived vehicles. The penalty due to the deployment of a new UAV can be written as $P^n = \xi \sum_{m=1}^{M} \gamma_m$, where $\gamma_m$ is a binary variable equals to 1 when UAV $m$ is deployed and 0 otherwise and $\xi$ is weight to avoid unnecessary dispatching of UAVs. 3) Cost of energy consumption by each UAV due to the traveling distance $d_m^n$: the RL-agent strives to maximize its rewards (minimize negative rewards), it will learn to minimize the total energy consumption of UAVs. Energy consumed by the deployed UAVs can be written as $E_{n+1}^U = \psi \sum_{m=1}^{M} \gamma_m E_n^m$, where $\psi$ is a weight of energy consumption. The immediate reward $r_n$ is the weighted sum of all three dimensions, where $\eta$, $\xi$, and $\psi$ are weight parameters for coverage, UAV deploying and energy consumption, respectively, and $\eta + \xi + \psi = 1$.

A larger value for $\eta$ will render the coverage the dominant factor, so the solution should deploy a UAV for a small number of vehicles, which economically could be expensive for the operator; a larger value for $\xi$ will render deploying a new UAV the dominant factor, hence, the solution will provide a non-continuous coverage with less deployed UAVs; while a larger value for $\psi$ will render the energy consumption by each UAV the dominant factor, which yields the deployed UAVs to be hovering to minimize the energy consumption. Motivated by this, we propose a RL algorithm for solving the covering problem in the next subsection.

### B. Reinforcement Learning Algorithm

A RL-agent is deployed at the central unit, and interacts with the vehicular environment in a sequence of actions, observations, rewards and penalties. At each time-slot $n$, the agent selects an action $d_m^n$ from the set of feasible actions at that time, where $d_m^n \in [0, d_{min}, 2d_{min}, ..., d_{max}]$ and $d_{min}$ and $d_{max}$ are the minimum and maximum traveling distance within a time slot, respectively. The deployed UAVs will either travel along the highway in a specific direction (left or right) or hovering to serve the vehicles. The agent then observes the dynamic changes in the vehicular environment, modifies the system state representation, and accordingly receives a reward or penalty. In order to achieve the maximum effective coverage on the highway, all UAVs should operate in a consistent, orderly and energy efficient way to provide the vehicles with an acceptable QoS. After each selected action either traveling or hovering, each UAV receives a step reward which is a normalized indicator of how well the selected action accomplishes the previously-mentioned goals. The objective of RL is to construct an optimal action selection policy for each UAV that covers the vehicles along the highway segment in order to achieve an effective coverage. It is worth mentioning that the received reward by each UAV depends on the entire previous sequence of actions and the observations from the vehicular environment. As such, the impact of the action, either moving or hovering, may only be seen after several time steps. In fact, RL is a branch of machine learning, which deals with multi-state decision process of a software agent (a central unit in our case) while interacting with an environment.

In general, RL assumes the system consists of multiple states $S$, where at each state $s_n \in S$, the agent has a finite number of actions $A$ (i.e., the next UAVs’ position). After choosing an action $a_n \in A$, the agent receives a reward $r(s_n, a_n)$, and moves to the next state $s_{n+1}$. The goal of RL is to learn from the transition tuple $(s_n, a_n, r(s_n, a_n), s_{n+1})$, and find an optimal policy $\pi^*$ that will maximize the cumulative sum of all future rewards. Note that the policy $\pi = \{a_1, a_2, ..., a_N\}$ defines which action $a_i$ should be applied at state $s_n$. If we let $r(s_n, \pi(a_n))$ denote the reward obtained by choosing policy $\pi$, the cumulative discount sum of all future rewards using policy $\pi$ is given by [12]:

---

1. In the real applications of UAVs, the deployed UAVs can fly in arbitrary traveling distance without any restrictions, but we constrain their mobility to as few distances. However, the proposed algorithm can accommodate any arbitrary discrete number of traveling distance. We choose few values to strike a tradeoff between the performance and complexity of the system.
\[ R_{\pi} = \sum_{n=1}^{N} \gamma^{n-1} r(s_n, \pi(a_n)) , \]  
\hspace{1cm} (3)

where \( \gamma \in [0, 1] \) is a discount factor, which measures the weight given to the future rewards. Now, let \( \Lambda \) denote the set of all admissible policies, then, the optimal policy is given by \( \pi^* = \arg\max_{\pi \in \Lambda} R_{\pi} \). RL is modeled as MDP, where the tuple \( \langle s_n, a_n, r(s_n, a_n), s_{n+1} \rangle \) is conditionally independent of all previous states and actions. Therefore, the agent does not need to memorize or save all the state-action tuples, just the last one, and subsequently updates it at each cycle or iteration. In this work, we use Q-learning, one of the widely used RL algorithms, which allows the agent to optimally act in an environment represented by an MDP. Q-learning iteratively improves the state-action value function (also known as Q-function or Q-value), and by estimating the future reward if action \( a_n \) is taken, the agent presents the higher probability of going from state \( s_n \) to \( s_{n+1} \) using policy \( \pi \). The Q-value function is usually stored in a table. Starting from an arbitrary Q-value, each time the agent wants to take an action, it approximates the optimal Q-value based on the observations of the environment, updates the Q-value according to equation (4) and stores it into the table. The parameter \( \alpha \in [0, 1] \) denotes the learning rate.

\[ Q(s_n, a_n) \leftarrow (1 - \alpha)Q(s_n, a_n) + \alpha[r(s_n, a_n) + \gamma \max_{a_{n+1}} Q(s_{n+1}, a_{n+1})], \]  
\hspace{1cm} (4)

Obviously, we are dealing with a finite control task since each UAV can carry out finite actions (traveling distance and direction), and hence the use of RL techniques is necessary to explore the effect of the UAVs’ actions on the vehicular environment. Now, it is important to mention that the reward function is not dependent only on one UAV but on a joint actions of all UAVs. It is noteworthy that even if the impact of the occurrence of the above described event is unveiled in a single time-step (i.e., when a vehicle arrives to the considered highway segment or departs from it), the RL agent realizes that the deployed and non-deployed UAVs and their previous trajectories lead to this current system state. This is a clear example that the feedback from an action may sometimes be delayed after many thousands of time slots have elapsed.

### IV. SIMULATION AND NUMERICAL ANALYSIS

The goal of the agent is to interact with the vehicular environment and select the best actions (the travel distance and direction) that maximize cumulative discounted future rewards in the given time \( N \). Recall that, the objective of this paper is to find a control policy that governs the trajectories of the deployed UAVs at each time-slot in order to achieve an effective coverage with a minimum number of UAVs while maximizing the energy efficiency of UAVs. This problem has been formulated as an MDP whose vehicular environment states are modeled as a Markov chain, and after defining the state-action and the cost that incurred at each state. The Q-learning algorithm is presented in Algorithm 1. It is assumed that a highway segment of length 3km is present, on which multiple UAVs are dispatched with communication range \( R_c = 1 \text{Km} \) to ensure a network coverage to vehicles. The flow of vehicles entering the highway segment follows a Poisson distribution. Vehicles velocities are randomly generated using a truncated Gaussian distribution with mean equal 100km/h, variance 16km/h, and velocities can be varied between 80–120km/h, where the vehicles randomly change their velocities within the given highway according to a normal distribution. Without loss of generality, it is assumed that at time \( n = 0 \), all UAVs are located at the beginning of the highway segment. The main input parameters are listed in Table I. For the sake of illustration, we also assume the time duration in the simulations is sampled every 1sec. Two scenarios, namely, RL-25 and RL-10, are considered, where each UAV in the RL-25 has the ability to carry out a traveling distance \( d_{m}^{n} \in [0, 25, 50] \text{m} \) per direction, while in the RL-10, a travel distance of \( d_{m}^{n} \in [0, 10, 20, 30, 40, 50] \text{m} \) is possible per direction. In fact, the UAVs in the RL-10 scenario provide more flexibility compared to those in the RL-25. We compared our proposed approach with two baseline methods to show the efficiency of our proposed approach: 1) Random UAV dispatching approach where, at each time-slot, it randomly selects a travelling distance within \( d_{m}^{n} \in [0, 10, 20, 30, 40, 50] \) as the current action for each UAV. If the new location is beyond the highway edge after executing this action, then the UAV that executed that action abandon this action. We also note that the number of deployed UAVs is selected randomly to study the percentage of coverage in every period \( T \). 2) Maximum Speed, where at each time-slot, the minimum number of UAVs selects the travelling distance \( d_{m}^{n} = 50 \) as the current action for each UAV. If the new location is beyond the highway edge after executing this action, a new UAV is dispatched with travelling distance \( d_{m}^{n} = 50 \). Fig. 2(a) depicts the UAVs trajectories to provide network coverage to vehicles (for vehicular density 5 Veh/Km and RL-10 scenario) over a

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. vehicle speed, ( v_{min} ) [Km/h]</td>
<td>80</td>
<td>Rotor speed, ( \omega_{ch} )</td>
<td>100</td>
</tr>
<tr>
<td>Max. vehicle speed, ( v_{max} ) [Km/h]</td>
<td>120</td>
<td>Air density, ( \rho )</td>
<td>1.225</td>
</tr>
<tr>
<td>UAV max speed, ( \omega_{max} ) [m/s]</td>
<td>50</td>
<td>UAV mass, ( m_U ) [Kg]</td>
<td>5</td>
</tr>
<tr>
<td>UAV's com. range ( R_c ) [Km]</td>
<td>1</td>
<td>UAV surface area ( A[m^2] )</td>
<td>0.25</td>
</tr>
<tr>
<td>Drag and reference area coef., ( F )</td>
<td>0.4</td>
<td>Blade dimension cons., ( R )</td>
<td>0.70</td>
</tr>
</tbody>
</table>

### TABLE I: Simulation Parameters
The vehicles enter the highway segment at different times as depicted in the figure. It also shows that only one UAV is needed to cover all vehicles within the beginning of the considered time period. Clearly, one UAV may not be able to cover all the vehicles within the given highway segment due to the high arrival rate of vehicles. We also observe that, while one UAV covers a set of vehicles on the highway segment, the second UAV is deployed to cover the newly arriving vehicles. After a time period, the deployed UAVs optimize their trajectories by traveling back-and-forth to cover the current and newly arriving vehicles. Due to the flexibility of UAVs, the figure also shows that the UAV starts its trajectory by following the first batch of arriving vehicle(s), then fly back to cover the second subsequent batch and so on.

Next, we study the impact of vehicular density and coverage percentage on the proposed solution (for RL-10 scenario) compared to random deployment of UAVs. As shown in Fig. 2(b), the proposed RL approach covers more vehicles compared to the random deployment. We can also observe from the figure that with lower vehicular density, the proposed RL dispatches fewer number of UAVs compared to random deployment. As expected, when the vehicular density increases, our RL approach dispatches more UAVs. Note that although the random deployment method dispatches fewer number of UAVs when the vehicular density is 6, 8, or 9 veh/km, however, it covers fewer number of vehicles. For instance, when the vehicular density is eight veh/km, the RL covers 12% more vehicles compared to random deployment. This shows the efficiency of the proposed framework in achieving effective coverage for the vehicles within the given highway segment, since the major goal is to maximize the vehicular coverage first and then minimize the number of deployed UAVs. In Fig. 2(c), we compare the performance of the proposed solution (RL-25 and RL-10) in terms of energy consumption versus two other trajectory methods: 1) random UAVs deployment, where the deployed UAVs follow random velocities to cover vehicles. 2) Fixed trajectories with maximum speed. It can be observed that the proposed solution shows a lower energy consumption compared to random and fixed speed deployment. It can be also seen that the RL-10 senario achieves a lower energy consumption compared to RL-25 senario, because the former provides more flexibility for the UAV to adapt its distance.

Fig. 2(d) presents the total traveled distances for the deployed UAVs to follow the batch of vehicles and cover them. From the figure, it is clear that the total traveled distances increases with the system duration time. However, our proposed RL approach significantly reduces the total UAVs’ traveled distances compared to the other methods. Not to mention, the RL-10 senario, as explained above, outperforms the RL-25.

V. Conclusion and Future Work

In this paper, the reinforcement learning framework has been proposed that allows multiple UAVs to autonomously cover vehicles in a mobility environment where communication services are not available. The objective of the proposed framework is to find the trajectories of the minimum number of UAVs with minimum energy consumption to provide effective coverage for the vehicles. It has also shown that the proposed framework guarantees minimization of the number of deployed UAVs with minimum energy consumption. A future extension of this work could be a joint optimization of trajectories and radio resource allocation of multiple UAVs in the presence of co-channel interference. Moreover, studying synchronization issues and multi-agent UAVs in vehicular net-work will be another direction in which the work can be extended.

References