Decentralized Learning Control for Multi-UAV Swarm Simultaneous Coverage and Tracking

Runfeng Chen, Ning Xu, Yuchong Huang, Yiting Chen, and Jie Li(✉)

College of Intelligence Science and Technology, National University of Defense Technology,
Changsha 410073, China
lijie09@nudt.edu.cn

Abstract. Environment coverage and target tracking are two important tasks in many research and applications about multi-UAV Swarm Systems. However, finding an optimal solution for maximizing the system's performance of coverage and tracking can be challenging, especially in limited resources scenarios, where swarms of UAVs need to cooperate with each other to cover areas and track multiple targets simultaneously in denied environment. The paper presents a decentralized control for UAV swarm simultaneous coverage and tracking with deep reinforcement learning, which also supports a collision-avoiding assembly. The proposed method uses a value network to evaluate the actions and tasks, which can offload the online computation to an offline leaning network meeting the demand of real-time applications. In particular, reciprocal simultaneous coverage and tracking method is used to generate mass data, which contribute to the initialization and convergence of value network. The effectiveness of the proposed method is corroborated by the numerical test.

Keywords: Decentralized control · Deep reinforcement learning · Simultaneous coverage and tracking · UAV swarm

1 Introduction

UAV swarm simultaneous coverage and tracking is a very challenging problem, especially in denied environment when the intention of the neighboring UAV is difficult to know, such as the target location is unknown [1, 2]. In addition, in the process of finding efficient paths, UAVs usually need to interact with neighboring UAVs in the expected state, which often requires a lot of calculation and consumes a lot of time.

At present, in the environment where there is no interaction between UAVs, simultaneous coverage and tracking methods can be roughly divided into two categories: one is the path method [3], the other is the reaction method [4]. Path-based optimization method is a kind of rolling optimization method, in which UAV makes decision for its actions by predicting the future state information of neighboring UAV. However, in a crowded environment, a large part of the predicted path set is usually unsolvable, which can easily lead to planning dead zones [5]. One solution is to introduce interaction, in

Z. Ren et al. (Eds.): *Proceedings of 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control*, LNEE 934, pp. 1705–1715, 2023.
https://doi.org/10.1007/978-981-19-3998-3_159

which the movements of each UAV can inspire and guide the movements of others, and each UAV can infer the intentions of the others and plan possible paths for them [6]. However, planning a path for all UAVs is time-consuming and computationally expensive [7]. In addition, due to the uncertainty of modeling and measurement, it is difficult to ensure that the actual path of other UAVs is consistent with the planned or predicted path, especially after a few seconds in the future [8]. Therefore, the path type needs to run at a high perception update frequency, which aggravates the problem of large calculation [9]. On the contrary, the reaction method may take longer time to finish tasks and be less effective than the path method, but this is not enough to cover its advantages of fast computing speed [10]. The reaction method is a single step rule decision based on the current geometric information, and the simultaneous covering and tracking method based on reciprocity mechanism is a kind of reaction method [10, 11]. It achieves the task requirements of simultaneous area coverage and target tracking by adjusting the speed of UAV, but it does not consider the future state of neighboring UAVs. This short-term optimization method is easy to cause unnatural behaviors such as motion oscillation in special cases [12]. Although the simultaneous coverage and tracking method based on reciprocity mechanism has the advantages of faster operation speed, shorter coverage time and higher coverage than the traditional algorithm, it cannot estimate the time required for regional coverage, so it cannot optimize the estimated regional coverage time.

Therefore, based on the reciprocal simultaneous coverage and tracking method [10, 11], this paper proposes a new distributed multi-UAV swarm coverage and tracking decision method combined with the deep reinforcement learning method, which can effectively reduce the online computation of interaction prediction to the offline learning process. Different from other traditional Q learning methods, the proposed method encodes the time required for area coverage, the location and speed information of UAV itself, as well as the location and speed information of its neighboring UAVs and adds them into the value network, so as to obtain continuously accessible speed action decisions through training the value network. This value network takes into account the uncertainties of other UAVs' movements and quickly provides effective speed decisions in real time.

2 Preliminaries

As a method of machine learning, reinforcement learning is mainly used to solve continuous action decision problem under unknown state transition model. In general, continuous state decision problems can be described as Markov decision processes.

Markov decision process is usually denoted as a quintuple $M = \langle S, A, P, R, \gamma \rangle$, of which, S denotes States, A denotes Actions, P denotes state transfer function, R denotes Reward function, and γ denotes decay factor within the value function V. By defining and describing these elements within a tuple, the two-UAV coverage problem is formally described as follows:

States S: in a two-UAV simultaneous coverage and tracking system, the system state of a single UAV $\mathrm{s^c}$ is usually composed of two parts, one is its own completely observable state s, the other is the external observation state $\tilde{\mathrm{s}}^{\mathrm{o}}$. The external observation state

$\tilde{s}^o = [s^{o,neighbor}, s^{o,target}, s^{o,sensor}]$ includes the observable state of another neighboring UAV $s^{o,neighbor}$, the optimal target tracking observation state $s^{o,target}$ and the perceived boundary/obstacle state $s^{o,sensor}$.Therefore, the system state of single-UAV simultaneous area coverage and target tracking can be formally described as $s^c = \left[s, \tilde{s}^o\right] \in \mathbb{R}^{26}$.

Actions A: Actions contain a series of feasible speed vectors. In this case, it is assumed that the UAV is a quadrotor type, that is, it can fly in any direction at any time and is only limited by the optimal movement speed, that is $\mathbf{a}(s) = \left\{\mathbf{v} | \|\mathbf{v}\|_2 < v_{pref}\right\}$.

Reward function R: The reward function is an important guide for UAV to complete the task, and is the key for UAV to carry out two-UAV cooperative coverage and target tracking in the designated boundary area under the obstacle environment. The main idea of the algorithm is to punish the wrong behaviors in two-UAV cooperative coverage and tracking, such as collision behavior, too close to the boundary or obstacle behavior, and too far from the best tracking target location behavior. In this section, the setting of the reward function $R(s^c, \mathbf{a})$ is shown in Formula (1). If an action results in a collision between two aircraft or collision with an obstacle (boundary), the penalty will be given a reward value of −1. If a certain behavior results in a distance from an obstacle or boundary less than the covering radius, the corresponding penalty shall be imposed, as shown in the second line of Formula (1); If a certain behavior causes the UAV to be too far away from the optimal tracking position, that is, the distance between the UAV and the optimal tracking position is less than a certain set threshold (set as half of the coverage radius in this paper), the corresponding punishment will be carried out, as shown in the third line of Formula (1). If the coverage task is successfully completed, that is, both machines exert their maximum coverage ability to cover more non-obstacle areas in the area as far as possible, the reward is 1. In other cases, there is no penalty and the reward value is 0.

$$R(s^c, \mathbf{a}) = \begin{cases} -1 & \text{if } d^a_{\min} < r + \tilde{r}_a \text{ or } d^\theta_{\min} < r \\ -0.5 + d^\theta_{\min}/2R & \textit{else} \text{ if } d^\theta_{\min} < R \text{ (coverage)} \\ -0.25 - d_t/2R & \textit{else} \text{ if } d_t > 0.5R \text{ (tracking)} \\ 1 & \text{else } \textit{if } \left| d^a_{final} - R - \tilde{R}_a \right| < \varepsilon \;\; \& \;\; d^\theta_{\min} >= R \\ 0 & \text{other} \end{cases} \tag{1}$$

Of which, $d^a_{\min}$ denotes two-UAV's least distance of distinct vision in the time period Δt, and $d^\theta_{\min}$ denotes the least value of eight directions of the UAV's sensor; d_t is the distance between the UAV and the optimal tracking position, and d^a_{final} is the distance between the two UAVs after the UAV performs the decision action in the time period Δt. It is assumed that the speed of action taken by the UAV remains constant during the time period Δt, i.e. $\mathbf{v} = \mathbf{a}, \forall t \in \Delta t$, other UAVs also fly at the observed speed $\tilde{\mathbf{v}}$ during the time period.

State transfer model P: State transfer model refers to how the state of the system changes after the UAV performs the decision action, i.e. $P(s^c_{t+1}, s^c_t | \mathbf{a}_t)$. One possible system transfer model is the update of UAV position state as shown in Formula (2) and Formula (3). However, UAVs only know their own strategies and intentions, while the strategies and intentions of other UAVs are unknown, so the transfer model of the whole system is also unknown. Assuming that other UAVs also use the same action strategy as UAVs themselves, i.e. $\pi = \tilde{\pi}$, the state transition model determines the strategy with UAVs.

$$\mathbf{p}_t = \mathbf{p}_{t-1} + \Delta t \cdot \pi : \left(\mathbf{s}_{0:t}, \tilde{\mathbf{s}}^o_{0:t}\right) \tag{2}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1} + \Delta t \cdot \pi : \left(\tilde{\mathbf{s}}_{0:t}, \mathbf{s}^o_{0:t}\right) \tag{3}$$

Value function V: A value function is a way to find the best decision for a period of time.

$$V * (s_0^c) = \sum_{t=0}^{T} \gamma^{t \cdot v_{pref}} \cdot R(s_t^c, \pi * (s_t^c)) \tag{4}$$

where, γ is a discount factor whose range is $\gamma \in [0, 1)$. v_{pref} is the optimal speed of UAV coverage flight, which is here regarded as a constant and does not change with time. Because the optimal speed of UAV varies with aircraft performance, the value function value of UAV with poor performance may be very small because of the low optimal speed. Therefore, in order to meet the numerical training learning under a large number of samples, this paper normalized it into a unit factor.

Thus, the optimal coverage strategy can be derived from the above value function:

$$\pi * (s_0^c) = \arg\max_{\mathbf{a}} R(s_0, \mathbf{a}) + \gamma^{\Delta t \cdot v_{pref}} \cdot \int_{s_1^c} P(s_0^c, s_1^c | \mathbf{a}) \cdot V * (s_1^c) \mathrm{d}\, s_1^c \tag{5}$$

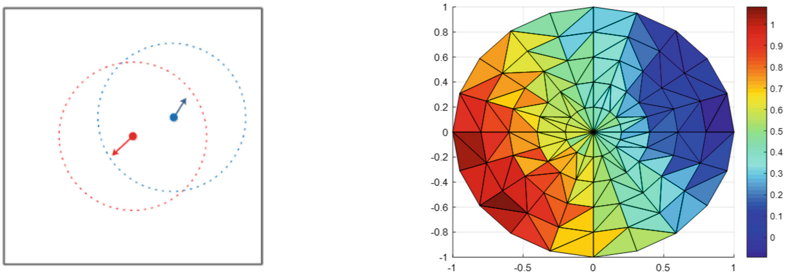
Different from the traditional Q learning method for discrete and limited action space decision making $Q(s^c, \mathbf{a})$, value network is more suitable for continuous feasible speed action space decision optimization. Therefore, this paper chooses the optimized value network function $V * (s^c)$ to solve the swarm simultaneous coverage and tracking decision problem.

3 Approach

3.1 Action Decision Driven by Value Network

Based on the receiving status value, an appropriate coverage network or simultaneous coverage and tracking value network is selected by discriminator D, thus a value network (v-net) V is given. UAV can make periodic simultaneous coverage and tracking action decisions by adopting the maximized value of value network each time. The specific algorithm is shown in Algorithm 1.

Each time the UAV chooses the decision that makes the action the most valuable, it performs that decision to update the state. Nevertheless, the numerical integration of Formula (5) make it hard to assess, resulting from the fact that UAVs cannot observe other UAVs' intentions, so the next state of other UAVs $\tilde{s}^o_{t+1}$ is an unknown distribution. Therefore, assuming that other UAVs move at a constant speed in a short period of time (lines 6–7 in Algorithm 1), this assumption estimation is not applicable to the next moment position estimation of other UAVs in the time $t > \Delta t$ of nonlinear motion model. This uncertainty about the next action of other UAVs can be reflected in the value of the next state V in the value network $V(\hat{s}_{t+1}, \hat{\tilde{s}}^o_{t+1})$. Then, the UAV selects the behavior with the highest value from the set of feasible velocity vectors, that is, the best action decision. Figure 1 shows the action strategy of reinforcement learning, in which Fig. 1(a) shows the state of the red agent (left), and Fig. 1(b) shows various state values with different actions (velocity vectors).



(a) input condition

(b) output value diagram

Fig. 1. Reinforcement learning action strategies

```
Algorithm 1  SCAT_DRL (Simultaneous Coverage And Tracking)
Input: Discriminator C, v-net $\mathbf{V}_c(\cdot,\mathbf{w})$ and $\mathbf{V}_{scat}(\cdot,\mathbf{w})$
Output: Paths $s_{0:t_f}$
while task unfinished do
  refresh time $t$, get latest observed values $s_t$, $\tilde{s}^o_t$
  Discriminator C Select the network $\mathbf{V}(\cdot,\mathbf{w}) \overset{C}{\leftarrow} \text{Select}(\mathbf{V}_c(\cdot,\mathbf{w}), \mathbf{V}_{scat}(\cdot,\mathbf{w}))$
  $\hat{\tilde{v}}_t \leftarrow \text{filter}(\tilde{v}_{0:t})$
  $\tilde{s}^o_{t+1} \leftarrow \text{propagation}(\tilde{s}^o_t, \Delta t \cdot \hat{\tilde{v}}_t)$
  $A \leftarrow \text{ActionSampler}()$
  $\mathbf{a}_t \leftarrow \underset{\mathbf{a}_t \in A}{\text{argmax}}\ R(s^c_t, \mathbf{a}_t) + \bar{\gamma} \cdot V(\hat{s}_{t+1}, \hat{\tilde{s}}^o_{t+1})$
      of which $\bar{\gamma} \leftarrow \gamma^{\Delta t \cdot v_{pref}}$,  $\hat{s}_{t+1} \leftarrow \text{propagation}(s_t, \Delta t \cdot \mathbf{a}_t)$
return $s_{0:t_f}$
```

3.2 Value Network Training

The training process of value network is shown as Algorithm 2, which contains v-net initialization and v-net training.

The first step is initialization of the value network. Using the reciprocal simultaneous coverage and tracking method, the path data of two-UAV cooperative coverage and tracking is generated, and a basic coverage and tracking strategy is created. Such positive sample data training can be regarded as supervised value network training (Algorithm 2, line 3). The paths of each training is a generated "state-value" pair $\{(\mathrm{s^c}, \mathrm{y})_\mathrm{k}\}_{\mathrm{k}=1}^{\mathrm{N}}$, of which $\mathrm{y} = \gamma^{\mathrm{t_c} \cdot \mathrm{v_{pref}}}$ and $\mathrm{t_c}$ denotes remaining-time between the current state and completion of the coverage task. The value network is trained in the way of backpropagation to reduce the quadratic regression error, namely $\arg\min\limits_{\mathbf{w}} \sum_{\mathrm{k}=1}^{\mathrm{N}} (\mathrm{y_k} - \mathrm{V}(\mathrm{s_k^c}; \mathbf{w}))^2$. Specifically, this paper generates 500 sets of two-UAV cooperative coverage and tracking path data, about 20,000 state-value pairs. The generated data set of the reciprocal simultaneous coverage and tracking method for training not only can help the value network quickly learn the good coverage and tracking strategy and accelerate network convergence, but also contribute to learn how to estimate the time needed to complete the coverage task, which is beneficial to generate new and better coverage trajectories using Algorithm 1.

Algorithm 2 Value network training

```
1 Input: paths set $D$
2 Output: v-net $V(\cdot,\mathbf{w})$
3 $V(\cdot,\mathbf{w}) \leftarrow \mathrm{init}(D)$          // procedure-1: initializer
4 copy v-net $V' \leftarrow V$          // procedure-2: learner
5 experience-pool initialization $E \leftarrow D$
6 for round $= 1,\dots,\mathrm{N_{eps}}$ do
7   for count=1,...,n do
8     $\mathrm{s_0}, \tilde{\mathrm{s}}_0 \leftarrow$ Randomization()
9     $s_{0:t_f} \leftarrow \mathrm{SCDRL}(V), \tilde{s}_{0:\tilde{t}_f} \leftarrow \mathrm{SCDRL}(V)$
10     $y_{0:T}, \tilde{y}_{0:\tilde{t}_f} \leftarrow \mathrm{get}(V', s_{0:t_f}, \tilde{s}_{0:\tilde{t}_f})$
11     $E \leftarrow \mathrm{add}(E, (y, s^c)_{0:t_f}, (\tilde{y}, \tilde{s}^c)_{0:\tilde{t}_f})$
12   $e \leftarrow \mathrm{Randomization}(E)$
13   w ← backpropagation(e)
14 return $V$
```

The second step is to train the value network through reinforcement learning to further optimize the coverage strategy. In each round, a small number of random test cases are generated and two-UAV coverage is carried out through the ε-greedy algorithm, that is, random action is selected with probability (Line 8) and the maximum action in the value network is selected at other times (line 9). These simulation paths then also create numerous "state-value" pairs. In order to speed up the network convergence, the value network is not updated immediately, but the newly generated "state-value" pair replaces the old record and is absorbed into the experience- pool E. Next, some data are chosen at random in experience-pool E to obtain the "state-value" pairs of many different simulation paths (line 12). Finally, v-net come out through propagation of adopted subdata set. Note, v-net is checked through predetermined periodic evaluation test cases for inspection.

3.3 Swarm Simultaneous Coverage and Tracking Driven by Learning

When performing swarm simultaneous coverage and tracking task, the value network of two-UAV can be extended and applied to single UAV control in swarm system.

The symbol $\tilde{\mathrm{s}}_{\mathrm{i}}^{\mathrm{o}}$ is represented as the observation state data of the current UAV to the i-th neighboring UAV, and the system state obtained by the current UAV including the observation state of the i-th neighboring UAV and its own state. Therefore, Algorithm 1 can be extended to achieve learning-driven swarm simultaneous coverage and tracking control. In other words, cooperative coverage and tracking based on reinforcement learning can update the status of each other neighboring UAVs through line 6–7 of Algorithm 1, select the best action for each neighboring UAVs from the action set. The formula is as follows:

$$\underset{\mathbf{a}_t \in A}{\arg\max} \ \min_{i} \quad R(s_{i,t}^{c}, \mathbf{a}_t) + \gamma^{\Delta t \cdot v_{pref}} \cdot V(\hat{s}_{t+1}, \hat{\tilde{\mathrm{s}}}_{\mathrm{i,t+1}}^{\mathrm{o}}) \tag{6}$$

where the estimation of the next state of the current UAV $\hat{\mathrm{s}}_{\mathrm{t+1}}$ is up to $\mathbf{a}_{\mathrm{t}}$. Though the used v-net is for two-UAV, the generated swarm simultaneous coverage and tracking trajectory shows that the current reinforcement learning-based approach can present complex interaction patterns.

However, Formula (6) is still only an estimate of the reinforcement learning value network of UAV swarm, which will be discussed in future researches.

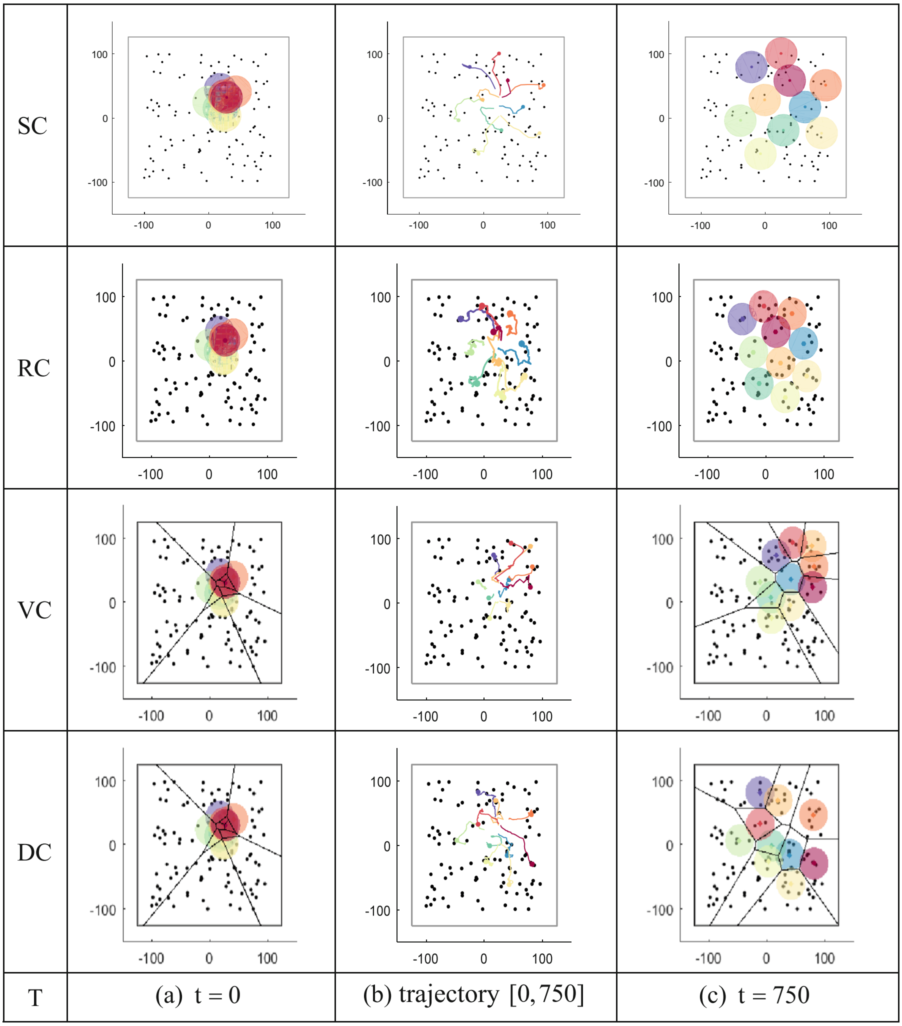
4 Experimental Simulation and Analysis

In this experimental environment, all kinds of conditions remain the same as in example 1, and swarm simultaneous coverage and tracking experiments are carried out in the same environment [10].

In this experiment, all the targets were in a static state, and each UAV performed simultaneous coverage and tracking tasks, covering more areas to find more targets and tracking the detected targets. Under the same conditions, the proposed method is compared with the three different representative algorithms, one is the region partition based Voronoi control method [8] (referred to as VC method), one is the local region based Density control method [12] (DC method), the other is reciprocal control method [10] (RC method). Under the same convergence threshold, if the algorithm meets the conditions $\sup \left\| \mathbf{p}_{\mathrm{i+1}}^{*} - \mathbf{p}_{\mathrm{i}}^{*} \right\|_2 \leq \zeta$, the operation will be terminated.

Figure 2 is a visual representation of the proposed method. As can be seen from Fig. 2(a), when all targets are randomly distributed in the square region, while UAVs are randomly distributed in the region. The distance between UAVs is relatively close and the number of detected targets is small. Subsequently, the UAV began to attempt to keep track of the detected target while searching for more potential potential targets by expanding its coverage area. The swarm trajectory is shown in Fig. 2(b). Finally, the

UAV swarm reaches a stable state, as shown in Fig. 2(c). Through the comparison, it can be intuitively found that the swarm of the proposed method covers more area (colorful circles), detects more targets (black points) with smoother trajectory.



(a) t = 0

(b) trajectory [0, 750]

(c) t = 750

Fig. 2. Comparison of the coverage and target detection numbers of four methods.

In order to quantify the effect of the proposed algorithm, statistical analysis is conducted on the number of detected targets for the four types of simultaneous coverage and tracking methods every 100 s, as shown in Fig. 3. Compared with VC and DC methods, the RC method can detect more targets at the same time due to the consideration of neighborhood reciprocity performance, while the SC method can further improve the coverage ability and tracking ability and detect more targets at the same time by learning and training on the basis of RC data sets.

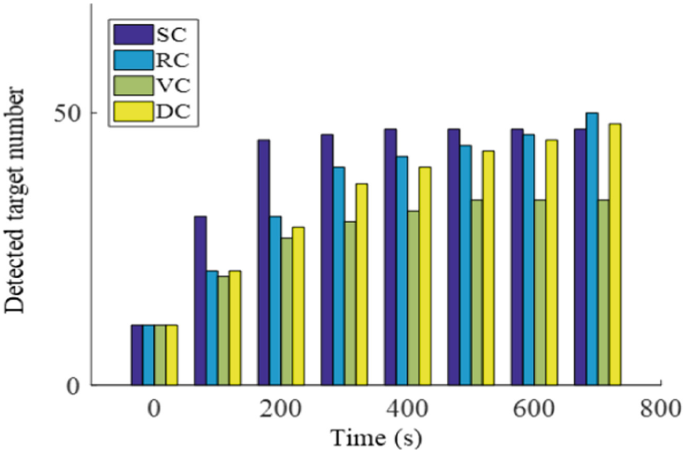


Fig. 3. Comparison of target detection numbers of four methods.

Coverage rate is the ratio of the total coverage of all UAVs to the area of the designated area. In this experiment, the maximum coverage of ten UAVs to the designated area is 31.4%, that is the coverage of all UAVs without overlap has reached the maximum coverage capacity. Figure 4 shows the comparison of swarm coverage of the four methods. It can be found that the proposed SC method after learning RC method has higher coverage than the other three methods.

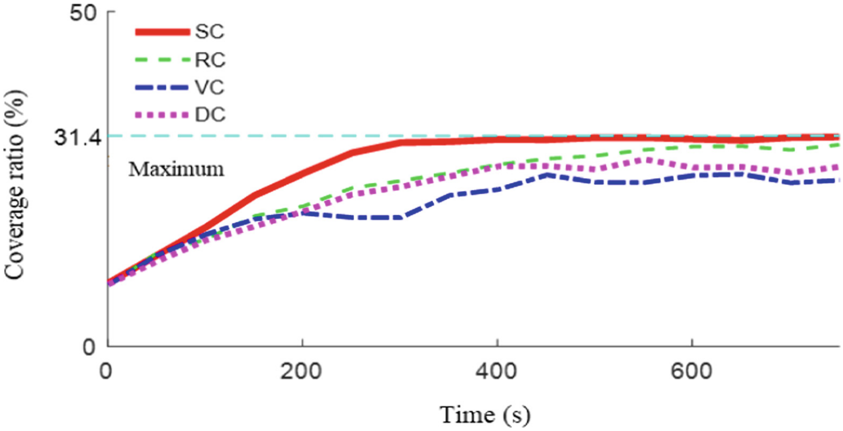


Fig. 4. Comparison of coverage rate among four methods.

5 Conclusion

In this paper, a simultaneous coverage and tracking method based on deep reinforcement learning is proposed. In particular, it uses the reciprocal simultaneous coverage and tracking method to perform a large number of two-UAV coverage simulation and obtain its trajectory data set to initialize the value network. This method can not only learn the reciprocal method strategy and accelerate convergence, but also enable the value network to learn the estimation method of the time required to complete the coverage task. In addition, the proposed two-UAV simultaneous coverage and tracking value network can not only achieve good results in the two-UAV cooperative problem, but also can be extended to the swarm cooperative control problem through testing. This method has strong real-time performance and can be used in swarm distributed system. The feasibility and effectiveness of the proposed algorithm are verified by a lot of simulation experiments.

Acknowledgement. This work was supported in part by Science and Technology Innovation 2030-Key Project of "New Generation Artificial Intelligence" under Grant 2020AAA0108200.

References

1. Khaledyan, M., Vinod, A.P., Oishi, M., Richards, J.A.: Optimal coverage control and stochastic multi-target tracking. In: 2019 IEEE 58th Conference on Decision and Control (CDC). IEEE (2019)
2. Pereira, N., Simonetto, A., De Visser, C.: Distributed asynchronous algorithm for collborative multi-UAV multi-target tracking, pp. 1–12 (2013)
3. Lin, D., Shen, B., Liu, Y., Alsaadi, F.E., Alsaedi, A.: Genetic algorithm-based compliant robot path planning: an improved Bi-RRT-based initialization method. Assem. Autom. **37**(3), 261–270 (2017)
4. Sugimoto, C., Natsu, S.: Self-organizing node deployment based on virtual spring mesh for mobile wireless sensor network (2014)
5. Morgan, D., Chung, S.: Swarm assignment and trajectory optimization using variable-swarm, distributed auction assignment and model predictive control, no. January, pp. 1–23 (2015)
6. Abbasi, F., Mesbahi, A., Mohammadpour, J.: Team-based coverage control of moving sensor networks, pp. 5691–5696 (2016)
7. Thanou, M., Stergiopoulos, Y., Tzes, A.: Distributed coverage using geodesic metric for non-convex environments. In: Proceedings of IEEE International Conference on Robotics and Automation, no. February (2015)
8. Moon, S., Frew, E.W.: Distributed cooperative control for joint optimization of sensor coverage and target tracking. In: 2017 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 759–766 (2017)
9. Skoglar, P., Orguner, U., Ornqvist, D.T., Gustafsson, F.: Road target search and tracking with gimballed vision sensor on an unmanned aerial vehicle. Remote Sens. **4**(7), 2076–2111 (2012)
10. Chen, R., Li, J., Shen, L.: A self-organized reciprocal control method for multi-robot simultaneous coverage and tracking. Assem. Autom. (2018)

11. Chen, R., Xu, N., Li, J.: A self-organized reciprocal decision approach for sensing coverage with multi-UAV swarms. Sensors **18**(6), 1864 (2018)
12. Pimenta, L.C.A., et al.: Simultaneous coverage and tracking (SCAT) of moving targets with robot networks. In: Chirikjian, G.S., Choset, H., Morales, M., Murphey, T. (eds.) Algorithmic Foundation of Robotics VIII. STAR, vol. 57, pp. 85–99. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-00312-7_6