UAV Maneuvering Target Tracking based on Deep Reinforcement Learning

Z Gan\textsuperscript{1}, B Li\textsuperscript{1}, E S Neretin\textsuperscript{2} and S A Dyachenko\textsuperscript{2}

\textsuperscript{1} Northwestern Polytechnical University, Xi’an, 710072, China
\textsuperscript{2} Moscow Aviation Institute (National Research University), 4, Volokolamske shosse, Moscow, 125993, Russia

E-mail: libo803@nwpu.edu.cn

Abstract. Maneuvering target tracking means that UAV observes the target through sensors, follows the target and maintains the tracking range with the target. Aiming at the target tracking problem of UAV, we built UAV motion model and UAV tracking model based on reinforcement learning, and verified the feasibility and effectiveness of the method in the simulation environment.

1. Introduction
UAVs have the advantages of small size, light weight, fast response. In recent years, UAV have been extensively applied for different types of tasks, such as object tracking [1], wildlife protection [2], disaster rescue [3]. Maneuvering target tracking means that UAV observes the target through sensors, follows the target and maintains the tracking range with the target.

Traditional target tracking algorithms have poor performance and cannot realize real-time target tracking. To overcome the shortcomings, researchers propose methods based on reinforcement learning (RL). RL is an effective approach to train agent to make decisions. Mnih [4] developed a novel artificial agent, termed a deep Q-network (DQN), that can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning. Since then, deep reinforcement learning algorithms have been more widely used in maneuvering target tracking. Huang [5] combined DQN with UAV navigation to enable the agent to make decisions based on the received signal strength. Wu [6] proposed the search algorithm based on DQN, which can realize automatic search and path finding of UAV. However, the value based DRL algorithms such as Q-learning, Sarsa, DQN etc. can only achieve the control of discrete action spaces, so they can only realize the discrete direction control of UAV.

In order to achieve continuous control, Peters [7] proposed policy gradient algorithm, which expressed the policy as a continuous function containing the parameters, and optimized the policy to find the maximum reward through gradient ascent. Silver [8] demonstrated an off-policy actor-critic algorithm deterministic policy gradient (DPG) which can significantly outperform stochastic counterparts in high-dimensional action spaces. Lillicrap [9] introduced DQN on the basis of DPG, proposed DDPG algorithm combining the training characteristics of DQN neural network. Because of the excellent performance of DDPG, many researchers have applied DDPG to UAV navigation and target tracking, and achieved excellent results. Wang [10] developed a DRL framework for UAV navigation and realized UAV navigation based on DDPG.
However, these progresses discuss more about the navigation of UAV to fixed targets. If the navigation of stationary target is applied to the tracking of moving target directly, the tracking effect will not be guaranteed. We will have a separate discussion on UAV maneuvering target tracking and propose a DRL algorithm for UAV target tracking.

In this paper, we constructed the UAV maneuvering target tracking method based on deep reinforcement learning. We evaluate the effect of the training of reinforcement learning through the average reward and the tracking effect. Through experiments, we verified the feasibility and effectiveness of the method.

2. Problem Formulation
2.1. UAV Motion Model
In target tracking mission, we assume that UAV is flying at a fixed altitude.

The UAV motion model in inertial coordinates is shown in (1):

\[
\begin{align*}
\varphi_{t+1} &= \varphi_t + \omega_t \, dt, \\
v_{t+1} &= v_t + n_t \, dt, \\
x_{t+1} &= x_t + v_{t+1} \sin(\varphi_{t+1}) \, dt, \\
y_{t+1} &= y_t + v_{t+1} \cos(\varphi_{t+1}) \, dt,
\end{align*}
\]

(1)

In (1), \(x_t, y_t\) denote the position coordinates of the UAV; \(v_t\) denotes the velocity of the UAV, \(n_t\) is the acceleration of the velocity of the UAV; \(\varphi_t\) denotes the heading angle, \(\omega_t\) denotes the angular velocity of the heading angle (see figure 1).

![UAV motion model](image)

**Figure 1.** UAV motion model.

The control of UAV is realized through the acceleration and the angular velocity of the heading angle, where acceleration represents the control of throttle and the angular velocity represents the control of steering.

2.2. Target Tracking Model Based on Reinforcement Learning
Reinforcement learning is a method in which an agent learns a near-optimal behavior by interacting with the environment [7]. Reinforcement learning can be described as Markov decision process (MDP) by a five-tuple \((S, A, P, R, \gamma)\). State space \(S\) is the collection of all states of the UAV. Action space \(A\) is the collection of all the actions that the UAV can perform. \(P(s' | s, a)\) indicates the transition probability. \(R\) is denoted as \(R(s, a)\), indicating the reward that UAV can be obtained by taking action \(a \in A\) in state \(s \in S\). \(\gamma\) represents the discount factor. The specific definition is as follows.
2.3. State space
The state space describes the information that the UAV can obtain from the environment.

The state space consists of the following variables, \( x, y, \varphi, v, \psi, d, x, y \) denote the coordinate of UAV, \( \varphi \) denotes the heading angle, \( v \) is the linear velocity of the UAV, \( \psi \) denotes the angle between the linear velocity of the UAV and the target line and \( d \) is the Euclidean distance between the UAV and the target.

Every state \( s \) is defined as
\[
s = [x, y, \varphi, v, \psi, d]^T.
\] (2)

2.4. Action space
We set the acceleration and the angular velocity as actions. Action of the model is defined as (3):
\[
a = [\omega, n]^T.
\] (3)

2.5. Reward function
The reward function is composed of three parts: distance reward, angle reward and extra distance penalty. The normalized distance reward function is defined as
\[
r_d = -\delta \frac{d_t}{d_{max}},
\] (4)

where \( \delta \) is a positive constant, \( d_{max} \) is the max distance that UAV can detect. To ensure the UAV flying in the direction of the target, we set an angle reward. The normalized angle reward function \( r_\psi \) is defined as
\[
r_\psi = -\varepsilon |\psi_t - \psi_{t-1}|,
\] (5)

where \( \varepsilon \) is a positive constant and \( \psi_t \) denotes the angle between the linear velocity of the UAV and the target line at time step \( t \). In order to achieve faster tracking of the target, we set extra distance penalty. The extra distance penalty is defined as
\[
r_p = p, \quad \text{if} \quad d_t > L.
\] (6)

To summarize, the final reward function can be formulated as
\[
r = r_d + r_\psi + r_p.
\] (7)

2.6. Training process
The framework for reinforcement learning includes Actor and Critic. Figure 2 shows the UAV target tracking reinforcement learning framework.

The loss function of actor module is defined as (8):
\[
J = E\{Q_{\theta_1}(s, a)|s=s_j, a=\pi_\varphi(s_j)\}.
\] (8)

The agent can get action according to the current state through the actor. The input of the critic module is the state and the action of the UAV, the output is the Q-value. In order to obtain the maximum Q-value and reduce the over-estimation of the Q-value, according to the Bellman equations and [4], the target optimal Q-value function \( y_t \) is defined as (9):
\[
y_t = r(s_j, a_j) + \gamma \min_{i=1,2} Q_{\hat{\theta}_i}(s_{j+1}, \pi_\varphi(s_{j+1})).
\] (9)

The loss function of critic module is defined as (10). In equation (10), \( \theta_1, i = 1, 2 \) denote the neural networks parameters of online Q-value Networks
\[
L = \frac{1}{N} \sum_{j=1,2, \ldots, i=1,2} \sum_{j=1,2} (y_j - Q_{\theta_i}(s_j, a_j))^2.
\] (10)
Agent makes decisions according to the state of the UAV and the state of the target. Agent outputs the action to the environment to change the state of the UAV, and the target changes the state according to its own movement mode. The environment feedbacks the corresponding reward to the agent. The policy is dynamically updated in the process, so that the action tends to be optimal.

3. Simulation and Analysis

We set up comparative experiments to verify the implementation effect.

Assuming that the target is moving on the ground, the motion mode of target will change in real time. At time $t$, if the distance $d_t$ exceeds $L = 500$ m, the UAV will get a penalty $p = -1$. The maximum speed of UAV is $v_{\text{max}} = 15$ m/s. The maximum speed of target is $10$ m/s. The acceleration of the UAV is $n_t \in [-1.5, 1.5]$, the angular acceleration is $\omega_t \in [-2, 2]$. The coefficients in reward function are instantiated as $d_{\text{max}} = 4000$, $\delta = 0.6$, $\varepsilon = 0.3$. We evaluate the effect of the training of reinforcement learning through the average reward and the tracking effect. The experimental results are as follows (figure 3).

![Figure 2. UAV target tracking reinforcement learning framework.](image)

![Figure 3. Average reward trends during training.](image)
Figure 3 shows the average reward trends during training. During the training process, the average reward increases within 400 episodes, and reaches the maximum around 500 episodes. After 650 episodes, the average reward converges to the maximum reward. It means the drone has been able to track its target steadily.

Figure 4 shows the tracking effect of the UAV and the change of tracking distance over time.

![Tracking trajectory and tracking distance over time.](image)

**Figure 4.** The tracking trajectory and tracking distance over time.

In figure 4(a), the red star indicates the initial position of the UAV and the blue star indicates the initial position of the target. In figure 4(b), it shows the change of tracking distance over time. According to the results shown in figure 4, the UAV can approach the target quickly and stay within a set range. Within 50 steps, the distance between the UAV and the target decreases rapidly, which means that the drone is rapidly approaching the target. After about 120 steps, the UAV stays within 50 m of the target, and the range fluctuation is small, which means that the UAV has been able to achieve stable tracking of the moving target.

In summary, UAV Maneuvering Target Tracking based on Deep Reinforcement Learning can quickly reach the vicinity of the target, maintain the distance to the target, and have high effect and stability.

4. Conclusion
In this paper, we construct UAV motion controlled by acceleration and angular acceleration on a two-dimensional plane, and define the UAV maneuvering target tracking as MDP. We constructed the UAV maneuvering target tracking method based on deep reinforcement learning and realized the UAV maneuvering target tracking. Through experiments, we evaluate the effect of the training of reinforcement learning through the average reward and the tracking effect. Last, we verified the feasibility and effectiveness of the proposed method in the simulation environment. Nevertheless, there are still some work to do in the future. For example, we intend to extend the control of the UAV to 3D space and achieve target tracking in more complex environments.

References
[1] Fu C, Carrio A, Olivares-Mendez M A, Suarez-Fernandez R and Campoy P 2014 Robust real-time vision-based aircraft tracking from Unmanned Aerial Vehicles IEEE International Conference on Robotics and Automation (ICRA) pp 5441–5446
[2] Olivares-Mendez M A, Fu C, Ludivig P, Bissyandé T F, Kannan S, Zurad M, Voos H and Campoy P 2015 Towards an autonomous vision-based unmanned aerial system against wildlife poachers Sensors 15(12) 31362–91

[3] Birk A, Wiggerich B, Bülow H, Pfingsthorn M and Schwertfeger S 2011 Safety, Security, and Rescue missions with an Unmanned Aerial Vehicle Journal of Intelligent & Robotic Systems 64(1) 57–76

[4] Mnih V et al 2015 Human-level control through deep reinforcement learning Nature 518(7540) 529–533

[5] Huang H, Yang Y, Wang H, Ding Z, Sari H and Adachi F 2020 Deep Reinforcement Learning for UAV Navigation Through Massive MIMO Technique IEEE Trans. Veh. Technol. 69(1) 1117–21

[6] Wu C et al 2019 UAV Autonomous Target Search Based on Deep Reinforcement Learning in Complex Disaster Scene IEEE Access 7 117227–45

[7] Peters J and Schaal S 2006 Policy Gradient Methods for Robotics IEEE/RSJ International Conference on Intelligent Robots and Systems pp 2219–25

[8] Silver D, Lever G, Heess N, Degris T, Wierstra D and Riedmiller M 2014 Deterministic Policy Gradient Algorithms Proceedings of the 31st International Conference on Machine Learning (ICML-14) (Beijing, 21–26 June 2014) pp 387–395

[9] Lillicrap T P et al 2019 Continuous control with deep reinforcement learning [Online] (accessed: 2020-09-02) URL http://arxiv.org/abs/1509.02971

[10] Wang C, Wang J, Shen Y and Zhang X 2019 Autonomous Navigation of UAVs in Large-Scale Complex Environments: A Deep Reinforcement Learning Approach IEEE Trans. Veh. Technol. 68(3) 2124–36