UAV flight control method based on deep reinforcement learning

Shuangxia Bai
School of Electronics and Information, Northwestern Polytechnical University
Xi'an, China
nwpu18734760639@163.com

Bo Li
School of Electronics and Information, Northwestern Polytechnical University
Xi'an, China
libo803@nwpu.edu.cn

Zhigang Gan
School of Electronics and Information, Northwestern Polytechnical University
Xi'an, China
ganzhigang@mail.nwpu.edu.cn

Daqing Chen
School of Engineering, London South Bank University
London, UK
chend@lsbu.ac.uk

***Abstract*—Aiming at the intelligent perception and obstacle avoidance of UAV for the environment, an obstacle-avoidance flight decision method of UAV based on image information is proposed in this paper. Add Gate Recurrent Unit (GRU) to the neural network, and use the deep reinforcement learning algorithm DDPG to train the model. The special gates structure of GRU is utilized to memorize historical information, and acquire the variation law of the environment of UAV from the time sequential data including image information and UAV position and speed information to realize the dynamic perception of obstacles. Moreover, the basic framework and training method of the model are introduced, and the generalization ability of the model is tested. The experimental results show that the proposed method has better generalization ability and better adaptability to the environment.**

Keywords—GRU, DDPG, image information.

I. INTRODUCTION

With the progress of the science and technology and the advent of the information age, artificial intelligence is developing rapidly. Deep reinforcement learning provides a new idea for UAV flight control with its excellent performance in data processing and decision making. In the future, the UAV needs to grasp the state of environment constantly, perceive and analyze changes of the environment, at the same time , timely feedback to improve their own behavior strategy. Therefore, it is an inevitable requirement for development to adopt deep reinforcement learning algorithm to enable UAV to have the ability of perception, analysis and understanding when dealing with unfamiliar environment, and to make dynamic and autonomous decisions according to the state of environment. At present, method of flight control of UAV at home and abroad [1-3] are mostly based on the state of a single moment. Actually, the flight process of UAV has strong temporal dependence, so mining sequential feature of environment changes in the flight process of UAV is important to control the flight of UAV.

This paper uses the deep reinforcement learning algorithm DDPG, which combines image information and UAV state information as input, and continuously outputs UAV linear velocity to control the movement of the UAV to realize dynamic obstacle avoidance and flight. In the image processing part of this paper, GRU is added to control the input and memory information to make a prediction for the UAV linear velocity in the current time step. Therefore, based on GRU, an UAV flight control model is designed in this paper. By using four consecutive images as input, mining the sequential variation features. This model is superior to other tradition algorithms in point of sequential features extraction. Finally, the effectiveness of the proposed model for UAV flight control and obstacle avoidance is verified by experiments.

II. RELATED THEORY

A. Deep Deterministic Policy Gradient Algorithm

Deep Deterministic Policy Gradient algorithm (DDPG)[4]is an actor-critic, model-free algorithm based on the deterministic policy gradient that can operate over continuous action spaces. DDPG algorithm is able to find policies whose performance is competitive with those find by planning algorithm. For many of the tasks, the algorithm can learn policies "end-to-end": directly from raw pixel inputs. DDPG adopts network simulation policy function and Q function, and introduces replay memory to update network parameters. This paper uses the deep reinforcement learning algorithm DDPG to train the model. The training process of DDPG is shown in Figure 2.

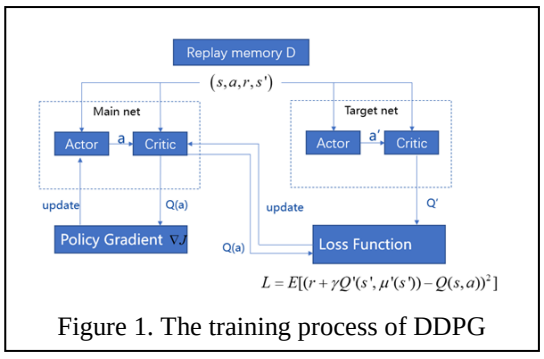

Figure 1. The training process of DDPG

B. Gate Recurrent Unit

As far as the network structure is concerned, the Recurrent Neural Network (RNN) is same as the traditional neural network, but the hidden layer neurons in the RNN are interconnected, so RNN can memorize the previous information and use this information to influence the output of the following nodes.

Gate Recurrent Unit (GRU)[5] is a type of RNN, it is also proposed to solve the problems of long-term memory and gradients in back propagation in RNN, same as Long-Short Term Memory(LSTM[6]. The principle of GRU is that the gating mechanism is used to control input, memory and other information to make predictions at the current time step, so that the information can selectively affect the state of each moment in the RNN. The structure of GRU is shown in Figure 2.

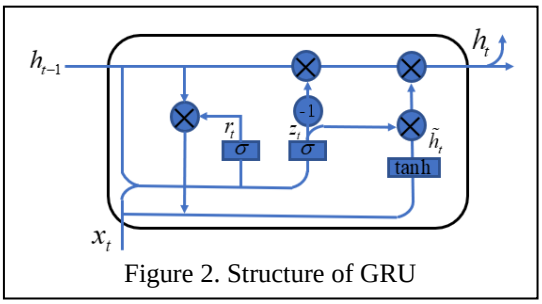

Figure 2. Structure of GRU

The calculation formula in GRU is as follows:

$$\begin{aligned} z_t &= \sigma\left(W_z \bullet \left[h_{t-1}, x_t\right]\right) \\ r_t &= \sigma\left(W_r \bullet \left[h_{t-1}, x_t\right]\right) \\ \tilde{h}_t &= \tanh\left(W \bullet \left[r_t, h_{t-1}, x_t\right]\right) \\ h_t &= \left(1 - z_t\right) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \tag{1}$$

GRU has two gates, namely reset gate and update gate. The reset gate determines the fusion of input information and memory information. The update gate controls the amount of data that memory information is saved to the current time.

III. Obstacle-avoidance flight decision method of UAV based on image information

A. Task Specification

Set the target point, the UAV judges the obstacle information and the relative position between the UAV and the target point based on the image information, its own position and speed information, and makes decision to bypass the obstacle and reach the designated target point.

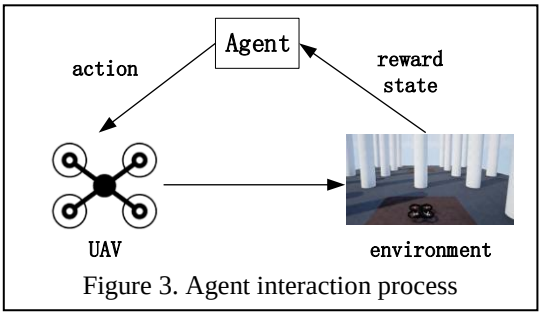

Figure 3. Agent interaction process

In this paper, the agent outputs the linear velocity of the UAV as a command based on the data obtained during the interaction between the UAV and the environment, including image information from the front camera of the UAV and the position and velocity of the UAV. After receiving the command, the UAV executes the action, obtains the corresponding reward in the environment, and updates the state. The interaction process of the agent is shown in Figure 3.

B. Data Processing

This paper takes the UAV to the designated destination through obstacle avoidance flight as the background, and obtains data from the simulation system that simulates the real flight process of the UAV. The data obtained includes image information from the front camera of the UAV and the position and velocity of the UAV. Process the image into a grayscale image, and stack the grayscale values of four consecutive frames into a tensor of size $(1,4,72,128)$ as input.

In the process of flight, the data of image, speed and position are generated according to the time sequence. These data describe the state and trend of the UAV at a certain time and were collected to form the historical dataset. The size of data collected for different attributes varies greatly, so the data must be normalized. In this paper, the data is normalized by the max-min normalization method, which can map the data values to [0,1].

C. Network Structure

According to the GRU and DDPG algorithm principle mentioned above, this paper constructs the neural network of obstacle-avoidance flight decision method. By inputting image tensor, the speed and position of UAV into the neural network, the UAV flight control variable is obtained: linear velocity.

The essence of the neural network learning process is to learn the data distribution. Once the distribution of the training data and the test data are different, the generalization

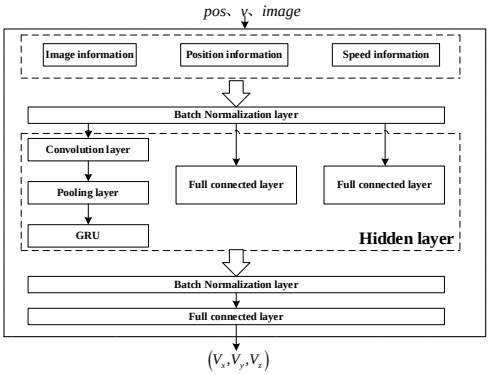

Figure 4. The network structure of obstacle-avoidance flight decision method

ability of the network is also greatly reduced. On the other hand, once the distribution of each batch of training data is different (batch gradient descent), the network must learn to adapt to different distributions in each iteration, which will greatly reduce the training speed of the network. Therefore, add Batch Normalization layer to the network to speed up the network training and convergence speed, control the gradient explosion to prevent the gradient from disappearing and prevent overfitting

Convolutional neural network (CNN) performs well in image processing, adding a convolutional layer to the network to process image information. The convolutional layer conducts in-depth analysis of each small block in the neural network to obtain more abstract features. The pooling layer is used to reduce the dimensionality of features, quickly reduce the size of the matrix, and reduce network parameters, thereby speeding up the calculation process and preventing overfitting. This paper selects the elu function as the

activation function for calculating the state value of GRU and the output value of convolutional layer in the hidden layer. And the output of the remaining layers is activated by the tanh function. In this paper, the optimum structure of the hidden layer is determined by several experiments with different numbers of hidden nodes and layers. The network structure of obstacle-avoidance flight decision method is shown in Figure 4.

D. Reward Function

(1) If the UAV flies too far from the existing scene, or a collision occurs, the UAV is judged to have crashed and the round ends. The reward value at this point is shown in equation (2).

$$r = -2 \tag{2}$$

(2) If the UAV reaches near the target point, the reward value is:

$$r = 2 \tag{3}$$

(3) The reward function during UAV flight is defined as follows:

- A positive reward is given if the UAV is closer to the target point than it was at the previous moment, and a negative reward is given if it is not. Define the distance reward as:

$$r_1 = \sqrt{(x_0 - x_{aim})^2 + (y_0 - y_{aim})^2 + (z_0 - z_{aim})^2} - \sqrt{(x - x_{aim})^2 + (y - y_{aim})^2 + (z - z_{aim})^2} \tag{4}$$

(x_0, y_0, z_0) is the last position of the UAV, (x, y, z) is the current position of the UAV, $(x_{aim}, y_{aim}, z_{aim})$ is the position of target point.

- The UAV should always be flying towards the target point. Define the UAV's flight direction reward:

$$\mathbf{s} = (x_{aim}, y_{aim}, z_{aim}) - (x, y, z) \tag{5}$$

$$\mathbf{v} = (v_x, v_y, v_z) \tag{6}$$

$$\cos\theta = \frac{\mathbf{s} \bullet \mathbf{v}}{|\mathbf{s}| \bullet |\mathbf{v}|} \tag{7}$$

$$r_2 = 2 * \cos\theta \tag{8}$$

The reward function during UAV flight:

$$r = \alpha r_1 + \beta r_2 \tag{9}$$

α and β are weight coefficients, it is adjusted according to the influence of various factors on the control effect in the experiment.

IV. Experiment and Analysis

This paper is based on the AirSim simulation platform for experiments. AirSim is a high-fidelity simulation platform with realistic vision, which adds many models consistent with the real world, such as weather, gravity, etc., to simulate the real environment. The UAV simulation environment built in this paper is a three-dimensional space.

A. Experiment Details.

This paper uses Adam for learning the neural network parameters with a learning rate of 10^{-4} and 10^{-3} for the actor and critic respectively. For *Q*, this paper includes L_2 weight decay of 10^{-2} and uses a discount factor of $\gamma = 0.99$. For the soft target updates, this paper uses $\tau = 0.001$. The final output layer of the actor was a tanh layer, to bound the actions. The final layer weights and biases of both the actor and critic are initialized from a uniform distribution $\left[-3\times10^{-3}, 3\times10^{-3}\right]$ and $\left[-3\times10^{-4}, 3\times10^{-4}\right]$ for the low dimensional and pixel cases respectively. This is to ensure the initial outputs for the policy and value estimates are near zero. The other layers are initialized from uniform distributions $\left[-\frac{1}{\sqrt{f}}, \frac{1}{\sqrt{f}}\right]$ where *f* is the fan-in of the layer.

B. Obstacle-Avoidance Flight Experiment of UAV Based on Image

This paper completes the construction and trains the flight decision model based on Torch module. The changes in the loss of the flight decision model training process are shown in Figure 5. It can be seen that the obstacle-avoidance flight result reaches the expected value After the network has been

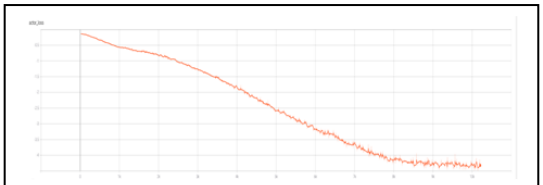

Figure 5. The loss function during training

updated 10,000 times.

Choose a set of parameters to test. During the test, the UAV can successfully identify obstacles and avoid them, and finally reach the designated destination. The flight trajectory of UAV is shown in the Figure 6.

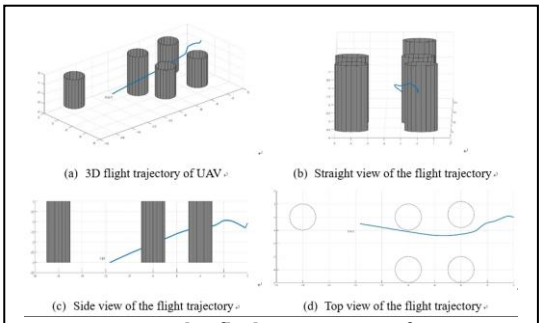

Figure 6. The flight trajectory of UAV

C. Generalization Ability Test

The generalization ability of reinforcement learning (RL) refers to the performance of the RL model in the test environment after training in the training environment. The generalization of RL models between different tasks is the focus of current deep reinforcement learning algorithm research. RL models can solve complex tasks after repeated training, but it is difficult to directly apply the learning experience to new environments. In the training process, the agent learns the details of the environment through continuous interaction with the environment, rather than learning the general performance in various environments. RL model training in a specific environment can achieve good training results, which may cause the model to overfit.

The model performs well in the training environment, but it is difficult to migrate to other environments, so it lacks generalization ability.

In this paper, the decision is made based on the image information of consecutive frames fused with sensor information. Taking image information as input, the model can perceive obstacles through image information. The distance change between the UAV and the obstacle is perceived through the change of continuous frame images on the time scale. The sensor information (the UAV's position and speed) can reflect the relative position relationship between the UAV and the target point to determine the direction of the UAV's movement. According to the above information, the UAV makes an obstacle avoidance action to realize the vision-based autonomous obstacle avoidance of UAVs.

The generalization ability of reinforcement learning model is tested from two aspects :(1) Verify the validity of image information in the decision-making process. Change the initial flight position and speed of the UAV in the environment, and place a column obstacle in front of the UAV at the same position for testing. (2) Test the adaptability of the model to different types of image information. During the testing process, change the shape of the obstacle in the environment for testing.

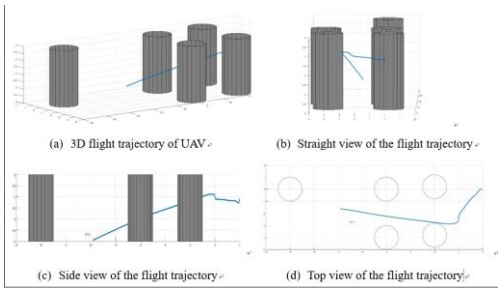

Figure 7. The flight trajectory of UAV after changing the starting point

Change the starting point to modify the position and speed information in the state information, the UAV can still bypass obstacles and reach the target point. This indicates the effectiveness of the image information in the decision-making process. The flight trajectory is shown in Figure 7.

Change the shape of the obstacle, the UAV can still identify the obstacle and avoid the obstacle, and finally reach the designated target point. The flight trajectory of UAV is shown in Figure 8.

The model generalization ability test results show that obstacle-avoidance flight decision method of UAV based on image information has a strong generalization ability, which can identify the unknown obstacles and make decisions according to the distance changes between the UAV and the obstacles, and successfully avoid obstacles.

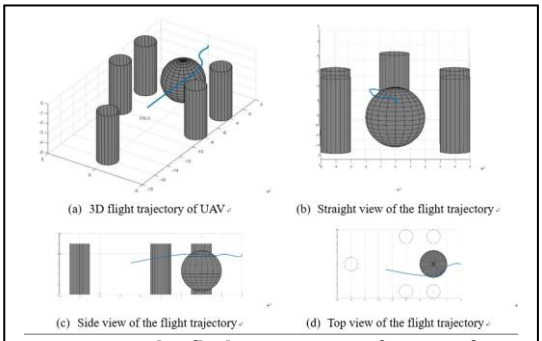

Figure 8. The flight trajectory of UAV after changing obstacles

V. CONCLUSION

In this paper, the deep reinforcement learning algorithm DDPG is used to study the flight decision method of autonomous obstacle avoidance of UAV based on image information under unknown environment. This article is based on the traditional neural network, adding GRU to build the network. The network takes image information and UAV position and speed information as input, and outputs UAV linear velocity. Finally, the generalization ability test of the model is carried out. The results show that the autonomous obstacle avoidance method of UAV based on image information has good scalability and improved adaptability to the environment.

REFERENCES

[1] Xue X. Indoor UAV Obstacle Avoidance Based on Deep Reinforcement Learning [D]. Harbin Institute of Technology, 2020.

[2] Xu G, Zong X, Yu G Su H. Research on Intelligent Obstacle Avoidance Method for Unmanned Vehicles Based on DDPG[J]. Automotive Engineering, 2019, 41(02): 206-212.

[3] Arnab M, Leonhard H C, Florian H. Time-Varying Parameter Model Reference Adaptive Control and Its Application to Aircraft[J]. European Journal of Control,2019.

[4] Lillicrap, T P, Hunt, J J, Pritzel A, Heess N, Erez T, & Tassa Y, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[5] Cho, K. , Merrienboer, B. V. , Gulcehre, C. , Ba Hdanau, D. , Bougares, F. , & Schwenk, H. , et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[6] Xingjian S H I, Chen Z, Wang H, Yeung D Y, Wong W K, & Woo W C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems , 2015(pp. 802-810).

[7] Volodymyr M, Koray K, David S, Andrei A R, Joel V, Marc G B, et al. Human-level control through deep reinforcement learning[J]. Nature, 2019, 518(7540):529-533.