Policy-based monocular vision autonomous quadrotor obstacle avoidance method

Baojin Zheng¹, Xiao Guo² and Jiajun Ou², *
¹ School of Aeronautic Science and Engineering, Beihang University, Beijing100191, China
² Research Institute for Frontier Science Beihang University Beijing 100191, China

*Corresponding author e-mail: oujiajun9042@163.com

Abstract. Aiming at the obstacle avoidance control problem of small quadrotor, a method of quadrotor obstacle avoidance based on reinforcement learning is proposed. The proposed method can make training converge quickly and has good environmental robustness. The proposed methods include: (1) a framework adopts perception module and decision module to improve the generalization ability of the obstacle avoidance model; (2) An Actor-Critic framework-based Proximal Policy Optimization (PPO) algorithm to provide quadrotor with policy-based decision-making capabilities; The experimental simulation results show that the strategy-based framework converges quickly and has a high success rate, the training time is much lower than that of the value-based framework. The monocular vision observation ability is limited, which leads to deviations between local observation and global state, So LSTM layer is usually added to increase model performance. Policy-based decision can have a good obstacle avoidance effect without adding the LSTM layer, and have good generalization ability after short relearning after changing.

Keywords: obstacle avoidance, deep reinforcement learning, policy-based, PPO.

1. Introduction
Quadrotor is a type of unmanned aerial vehicle. Due to its low cost and reusability, it is widely used in military and civilian fields. In 2002, the U.S. Air Force Research Office established 10 categories of autonomous control levels for UAV systems [1]. From the current development situation, autonomous control technology, as one of the core key technologies of unmanned aerial vehicles, has solved the problem of flight. The problem of automatic control, but the ability of intelligent autonomous control still needs to be improved. With the advancement of perception technology and computer computing ability, the development of intelligent control and intelligent decision-making is progressing.

Deep Reinforcement Learning is an emerging cross-research hotspot in the field of artificial intelligence [2]. Compared with classic control methods, deep reinforcement learning does not rely on expert knowledge. It can learn complex strategies only by interacting with the environment. It has been widely used in the fields of strategy games, autonomous driving and robot control, and can provide good support. Intelligent control and decision-making of human aircraft. Deep reinforcement learning integrates the powerful representation and abstract features of Deep Learning for strategies and states.
At the same time, it uses reinforcement learning to give the agent the ability to supervise learning so that it can autonomously interact with the environment. The evolution strategy in interaction has a very good application effect, especially for complex intelligent decision-making problems where image sequences are used as state inputs.

In the problem of quadrotor obstacle avoidance, the quadrotor can only afford a fixed monocular camera to provide limited environmental observation. Therefore, the autonomous obstacle avoidance of quadrotors is still a challenging task.

The decision module can adopt two types: value-based and policy-based. The value-based method has insufficient ability to deal with problems in a restricted state. When using features to describe a certain state in the state space, it is possible that due to the limitations of individual observations or the limitations of modeling, two states that are originally different in the real environment have the same feature description after we model them. It is very likely that our value-based method cannot get the optimal solution. Compared with value-based methods, our policy-based approach has better performance for the local observable of quadrotor monocular obstacle avoidance problem without adding the LSTM layer.

In order to solve the decision problem of limited monocular vision acquisition status, in this paper, a policy-based algorithm is proposed. This method is used for monocular vision quadrotor to provide a stable obstacle avoidance control strategy in an environment with limited global observation. At the same time, in order to prepare the improved model of continuous action space output, the main contributions of this research are as follows.

1. A two-stage architecture consisting of a perception module and a decision-making module is built, and the obstacle avoidance model is divided into two parts for training to improve the generalization ability of the obstacle avoidance model;
2. Established a monocular vision quadrotor obstacle avoidance decision module based on the PPO algorithm to improve the model's learning ability in a limited monocular vision acquisition environment;

2. Related worked

With the continuous maturity of deep learning technology, more efficient convolutional neural networks have been proposed, and the use of deep learning for deep estimation research has also received more and more attention, and more and more algorithms have emerged to complete images through deep neural networks. Estimate of depth. Loquercio et al. [3] design a fast 8-layers residual network to output the steering angle and a collision probability for each input image. The network is trained by the dataset manually annotated by the authors. In the paper [4], the authors use a multi-scale deep neural network to solve the problem of depth prediction, and successfully estimates the scene depth based on a single picture. Long et al. [5] build a fully convolutional neural network model, which is different from the previous network using a fixed-size fully connected layer, which cancels the previous model’s limitation on the input image size to achieve image-to-image conversion, and makes better use of GPU. On this basis, Zeng et al. [6] tried a deeper network, and the test results showed that the depth estimation ability has been greatly improved.

The supervised learning method needs to know in advance the reference standard of the depth value corresponding to a large number of input pictures as a training constraint, and the trained neural network is used to perform depth prediction for similar scenes. However, in reality, obtaining the depth value corresponding to the scene not only requires a specific sensor to synchronously collect image data and depth information, but also has very high requirements for data accuracy, and the acquisition cost is very high. Compared with the data preparation of supervised learning, the data preparation is difficult, and the training cost is high. In the paper [7] the authors use the monocular visual perception research based on unsupervised learning, and use the unsupervised learning method to estimate the depth, get the disparity map, and then calculate the depth map according to the obtained disparity.

Learning-based avoidance methods can be divided into end-to-end architecture and hierarchical architecture. The end-to-end architecture goes directly from sensor data to obstacle avoidance actions. Ou et al. [8] propose a quadrotor obstacle avoidance method which based on unsupervised learning
divides the framework into two modules: perception and decision. The perception module trained in an unsupervised manner can extract depth information from the on-board camera image. Moreover, the decision module uses dueling double deep recurrent Q-learning to eliminate the adverse effects of the on-board monocular camera’s limited observation capacity while choosing practical obstacle avoidance action.

The learning-based autonomous perception obstacle avoidance model is based on deep learning and reinforcement learning methods, and requires a certain amount of data for training to improve its capabilities. Training through actual flight is not only time-consuming and laborious. When the model training is not completed, the obstacle avoidance function cannot be realized, and the UAV is very likely to collide and be damaged. Therefore, it is an ideal way to build a virtual simulation environment, train the model through the simulation environment and then migrate it to the actual UAV platform for application.

In order to reduce the difficulty of training and improve the feasibility of application, make the training process converge quickly, and make the quadrotor have certain predictability and environmental robustness. The framework consists of two modules, and its training does not require annotated data sets. The first module is used to perceive the environment. It uses depth estimation based on unsupervised learning to generate a depth map of the environment. The second module uses a strategy-based algorithm to respond to obstacle avoidance decisions, and its strategy is obtained through deep reinforcement learning. The former can be trained by the original RGB monocular image sequence, and the latter can be trained in a simulated environment. This method is efficient and relatively easy to train. Facing new scenes, our framework only needs to re-use the pictures of the new scene to train the depth map, and it can adapt to the new work scene.

3. Proposed method

3.1. Simulation environment construction

Using the distributed architecture of ROS simulation, the GAZEBO module in ROS is run in the Ubuntu system environment, and the learning model framework based on PYTORCH is built and runs in the virtual environment of CONDA. The two interact through ROSMSG. Set up a simulation flight environment in the GAZEBO module of ROS to simulate the actual flight conditions of the quadrotor. The GAZEBO environment has been set up as shown in Figure 1. The interaction between the code and the quadrotor is realized through the ROS platform, and the flight status of the quadrotor is obtained and controlled in real time.

![Figure 1. GAZEBO simulation environment](image)
Before training the depth map, we must first obtain the training set, set the speed of the drone, and randomly control the yaw angle. The control frequency is 3.3 Hz. The control commands are shown in Table 1.

### Table 1. Control instruction.

| Action num | Linear velocity(m/s) | Angular velocity(rad/s) |
|------------|----------------------|-------------------------|
|            | (x, y, z)            | (x, y, z)               |
| 1          | (2, 0, 0)            | (0, 0, 0)               |
| 2          | (2, 0, 0)            | (0, 0, -0.5)            |
| 3          | (2, 0, 0)            | (0, 0, 0.5)             |
| 4          | (2, 0, 0)            | (0, 0, -0.25)           |
| 5          | (2, 0, 0)            | (0, 0, 0.25)            |

Through the control command, the drone is allowed to fly randomly in the simulation environment, and continuous pictures are taken directly in front of the drone. After the laser sensor detects that the drone collides with an obstacle, the drone will reset to (0, 0, 0) Point to fly again.

#### 3.2. Depth perception algorithm

The actual camera model is shown in Figure 1.

![Actual camera model](image)

According to this actual camera model, the transformation relationship formula between pixel coordinates and camera coordinates can be obtained as follows:

\[
D \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = K \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix},
\]

(1)

Where K represents the camera intrinsic matrix, f is the distance from the optical center of the camera to the picture, which is determined by the wide-angle parameters of the monocular camera, and thus the internal parameter matrix K of the camera is determined.

Calculate the pixel coordinates after the pose change according to the camera internal parameter K as follows:

\[
p_{s} \sim K_{\tilde{T}_{t-1}D_{t}}(p_{t}) K^{-1} p_{t},
\]

(2)

The differentiable bilinear sampling mechanism [9] is used to obtain Is ($p_{s}$) for populating the value of I’s ($p_{t}$).

By reconstructing the neighboring frame pictures of the target image into the target image, the image similarity calculation is performed on the target image and the reconstructed target image. By optimizing the calculation results of image similarity, the network parameters are back-propagated and updated, and the unsupervised learning of the depth estimation model and the pose change estimation model is realized.

The training pipeline is shown in Figure 2.
3.3. Policy-based Obstacle avoidance decision making

In the paper, the policy-based algorithm PPO based on the Actor-Critic framework is adopted to realize the obstacle avoidance of the quadrotor. The Actor network outputs the action distribution of the UAV after inputting the depth map, and the Critic network outputs the current after outputting the depth map. The value of state. In order to solve the data waste caused by the data of the AC framework can only be updated once, we adopted importance sampling to change the AC framework from an online strategy to an offline strategy.

For solving the conundrum that the behavior strategy is too divergent from the target strategy, we cut the scope of the update and limit the importance weight. Final loss of clipped objective PPO can be defined as follows

$$
loss = -min(Rto \times adv, clip.Rto \times adv) + 0.5 \times \sum_{t=0}^{n} (V_t - G_t)^2 - 0.01 \times De ,
$$

(3)

Where Rto is importance weight, adv is TD-error, $V_t$ is output of Critic network , $G_t$ is sum of subsequent rewards , De is entropy of distribution which is output of Actor network.

The training pipeline is shown in Figure 3

**Figure 3.** Perception learning based on view reconstruction.

4. Training and test

The learning curves of the two models are shown in Figure 4. The results are shown in Table 2.

**Figure 4.** Decision learning based on AC framework.
Figure 5. Learning curves of the DDQN and our PPO

Table 2. Test results of 2 different models

| Model | Obstacle avoidance path success rate |
|-------|-------------------------------------|
| Random | 0.02 |
| DDQN  | 0.13 |
| PPO   | 0.095 |

5. Conclusion
In this article, a policy-based decision algorithm is proposed. It allows quadrotors with only monocular cameras to make good obstacle avoidance decisions through fast training. Training and test results show that PPO can converge faster than DDQN method and has better environmental robustness. In the next plan, the proposed decision-making algorithm will have a more complex logical structure to control the quadrotor to perform continuous action space decision making to make the control model close to the real scene.

References
[1] Gupta S G, Ghonge M M, Jawandhiya P M. Review of Unmanned Aircraft System (UAS)[J]. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2013, 2(4): 2278–1323.
[2] Metalearning | Workshop on Meta-Learning (MetaLearn 2019)[EB/OL]. [2020-02-19]. http://metalearning.ml/2019/.
[3] A. Loquercio, A. I. Maqueda, C. R. Del-Blanco, D. Scaramuzza, Dronet: Learning to fly by driving, IEEE Robotics and Automation Letters 3 (2) (2018) 1088–1095.
[4] David Eigen, Christian Puhrsch, Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In NIPS, 2014.
[5] Long, Jonathan, Shelhamer, Evan, Darrell, Trevor. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[6] Laina I , Rupprecht C , Belagiannis V , et al. Deeper Depth Prediction with Fully Convolutional Residual Networks[J]. 2016.
[7] Godard, Clément, Mac Aodha O , Brostow G J . Unsupervised Monocular Depth Estimation with Left-Right Consistency[C]// In CVPR, 2017.
[8] Ou, J., et al., Autonomous quadrotor obstacle avoidance based on dueling double deep recurrent Q-learning with monocular vision. Neurocomputing, 2021.
[9] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, Advances in Neural Information Processing Systems 28 (NIPS 2015) (06 2015).