Small sample vehicle target recognition method for unmanned aerial vehicle system based on deep learning

Tingping Zhang *, Di Wan
School of Information Science & Engineering, Chongqing Jiaotong University, Chongqing 400074, China
*Corresponding author e-mail: ztp@cqjtu.edu.cn

Abstract. Vehicle target detection technology refers to the process of vehicle detection and recognition in the image data set of different backgrounds by means of feature extraction. The vehicle target detection technology based on deep learning shows obvious advantages in the accuracy and speed of target detection. With the development of science and technology, the detection and recognition of vehicles in UAV aerial images has become an important applied research direction. This paper studies the detection and recognition of UAV aerial vehicle based on deep learning, and proposes a new deep learning-based algorithm to solve the problem that incomplete vehicle targets in the UAV aerial vehicle based on YOLOV3 algorithm cannot be recognized, and vehicles close to the UAV aerial vehicle are missed. Experimental verification results show that, compared with the existing algorithms, the proposed algorithm can significantly improve the detection accuracy of UAV aerial vehicle based on deep learning while ensuring real-time performance.

Keywords: vehicle, detection, unmanned aerial vehicle system.

1. Introduction

The UAV aerial photography system consists of two parts: unmanned aerial vehicle and airborne photoelectric reconnaissance platform. Using UAVs to carry photoelectric reconnaissance equipment, the ground image information can be obtained stably and efficiently [1-2]. The collected image information is transmitted to the information processing center on the ground through the wireless link, and the aerial data processing is finally completed. In practical application, the contradiction between a large number of aerial data processing requirements and the limited processing capacity of the ground information processing center is increasingly prominent [3]. Especially for some occasions that require multiple aircraft to collect information and provide analysis results in a short time, the processing pressure on the ground end will be very large, and the system is difficult to meet the demand of real-time processing [4]. The realization of target recognition in airborne reconnaissance equipment can promote the discretization and marginalization of the entire system information processing, thus improving the timeliness of system information analysis. In the face of massive aerial images, how to quickly and efficiently realize the detection and recognition of objects of interest and eliminate the interference of redundant background information has become the top priority of research [5]. At present, computer vision technology has made great progress, and has made important research and
application achievements in face detection, license plate recognition and other general scenes. However, for aerial scenes, the detection and recognition technology of typical targets remains to be further studied. Vehicle is a typical target in aerial photography scene. Vehicle target recognition has important application space in both military and civil fields [6-9].

Using computer vision to accurately detect other vehicles on the road is a challenging task, and it is also a research hotspot in the past two decades. Traditional vehicle detection methods extract features from images with the help of machine learning theory [10]. The three most common methods are HOG, HAAR and LBP algorithms, which are the first in the field of vehicle detection. In 2017, Daniel et al. combined the above three common methods and proposed an image classification method based on feature combination. This method uses stereo model for target detection of moving vehicles, but this method is more complex, and the detection efficiency is low, the robustness is poor, can not be suitable for multi-type environment. In addition, the scale invariant feature transformation (SIFT) method is also used in vehicle detection, which is often used to detect the rear surface of traffic vehicles [11]. This method has a good solution to the problems of local occlusion and scale change as well as the constantly changing view point of the detection camera. However, SIFT feature extraction involves a large amount of computation, which requires a long time for recognition test, and has poor real-time effect on traffic detection [12]. In essence, the above methods are all traditional feature extraction methods based on manual design and lack of robustness. In the face of different image processing problems, the processing effect is also very different, so it is difficult to design a universal and robust method in practical application problems. In recent years, deep learning has achieved good results in the field of image recognition. Compared with traditional image processing methods, deep learning method combines multi-layer convolutional neural network to directly convolve images, so as to realize image feature extraction. The intelligence of deep learning makes it have obvious advantages in the field of image recognition, and has achieved rapid development, bringing a pioneering solution for the field of image feature extraction.

Various vehicle re-recognition models based on deep learning have also been put forward continuously. Liu et al. proposed a "PROVID" model to carry out progressive vehicle re-identification, and realized the search from coarse to fine in terms of features and from near to far in monitoring. Yan et al. proposed a model based on the multi-task framework to model vehicle images into multi-granularity relations, and proposed a generalized binary "similar/dissimilar" relationship of paired sorting and a multi-granularity based list sorting method, and progressively used multi-granularity sorting constraints to alleviate the problem of accurate vehicle search. Liu et al. proposed a "RAM" model composed of four branches, which combined the overall features and regional features to extract more detailed and more discriminative features.

2. Model framework
In this paper, deep learning is divided into two parts: classified network training; the knowledge of classification network is transferred to the detection model, and then the detection model is trained. For the classification network part, the classification network structure is established according to the detection model structure. As shown in table 1, based on the convolution layer of the detection model feature extraction part, by adding a layer of 13 convolution kernels with the number of 1 000 × 13 convolution layer, the characteristic matrix is converted to 1 000 × 1 dimension eigenvector, and then add a softmax layer to transform the eigenvector into probability output.
| Type       | Num | Filter | Channel | Stride |
|------------|-----|--------|---------|--------|
| Convolutional | 3×3 | 32     |         | 1      |
| Convolutional | 3×3 | 64     |         | 2      |
| Convolutional | 1×1 | 32     |         | 1      |
| Convolutional | 1   | 3×3   | 64      | 1      |
| Residual    |     |        |         |        |
| Convolutional | 3×3 |        | 128     | 2      |
| Convolutional | 1×1 | 64     |         | 1      |
| Convolutional | 4   | 3×3   | 128     | 1      |
| Residual    |     |        |         |        |
| Convolutional | 3×3 |        | 256     | 2      |
| Convolutional | 1×1 | 128    |         | 1      |
| Convolutional | 8   | 3×3   | 256     | 1      |
| Residual    |     |        |         |        |
| Convolutional | 3×3 |        | 512     | 2      |
| Convolutional | 1×1 | 256    |         | 1      |
| Convolutional | 8   | 3×3   | 512     | 1      |
| Residual    |     |        |         |        |
| Convolutional | 3×3 |        | 1024    | 2      |
| Convolutional | 1×1 | 512    |         | 1      |
| Convolutional | 4   | 3×3   | 1024    | 1      |
| Residual    |     |        |         |        |
| Convolutional | 1   | 13×13 | 1000    | 1      |
| SoftMax     |     |        |         |        |

After the classification network is established, Imagenet data set is used for training. After training the classification network, the knowledge of classification network training can be transferred to the inside of the commutator to expand the feature extraction part of the image defect detection model, and the migration process can be completed. After training the classification network with Imagenet dataset, the weights of convolution kernel can be trained sufficiently to extract generalized features, and these convolution kernels can be used as the knowledge learned by the classification network. The knowledge is transferred to the inside of the commutator to expand the feature extraction part of the image defect detection model structure, and the learning rate of the feature extraction part is set to 0. The convolution kernel of the feature extraction part is locked. The convolution kernel of the regression prediction part is assigned by random initialization, and then the convolution kernel of the regression prediction part is trained by using the sample set. The specific training process is as follows:

Step 1: construct classification network and determine the structure of classification network.
Step 2: use Imagenet data set to train the classification network, and get the trained classification network model.
Step 3: apply the principle of transfer learning to transfer the classification network parameters to the detection network, and set the learning rate of the feature extraction network part to 0, that is, the fixed learning rate.
Step 4: training the convolution kernel parameters of regression prediction part by using data set.

3. Experimental results and analysis

3.1. Data sets and evaluation indicators
Selection of experimental evaluation criteria: This paper uses ResNet-50 network as the reference network, and comprehensively evaluates the methods proposed in the paper from local to global on the MSCOCO data set. The standard COCO data set indicators are selected here to evaluate the experiment, which mainly include Average Precision (AP), Average Recall (AR), Training Time (Train Time), Test Rate (Test Rate) and Variant standard measurement of different indicators.
Experimental data set description: The experiment mainly relies on the MSCOCO data set. The data set has 80 object categories, among which there are 80k images in the training set, and 40k images in the test set and validation set.

The verification experiment in this article will divide the 40k verification set into 35K and 5k data sets. Then, the 80k training set and the 35k validation set are combined to obtain a 115k training set and a 5K small validation set. Due to the choice of GPU1080 configuration, the number of iterations should be set as small as possible, but a lot of time will not be wasted on verification, which makes the update frequency of Tensorboard higher. Trainval 115k is used for training in this article. The initial training sets the learning rate to $10^{-5}$ and carries out 160k iterations. The subsequent training has a learning rate of $10^{-4}$ for 40k iterations and a learning rate of $10^{-5}$ for 40k iterations.

Due to the processor limitation, each GPU in the article only processes 1 image at a time, and each image will set 2000 regions of interest for training and 1000 regions of interest for testing. Secondly, the number of verifications has been fine-tuned many times to prevent the training from being slowed down if it is too large and the verification accuracy is reduced if it is too small. The image input size is 1024×1024, but the input size of FPN operation is 256×256 feature map. Secondly, in the expanded network module, ReLU, which is simple and can avoid the disappearance of the gradient, is still selected, and BN is added. Finally, the backbone network introduces the method of migration learning, and respectively migrates the pre-trained Mask R-CNN and Inception models to their own models. The size of the pooling layer is set to 7, and the training parameters and BN parameters are further modified through migration learning. As a result, the training speed is improved.

3.2. Deep extended convolutional network experiment

Due to the inner side orientation of the industrial commutator, it is difficult to observe the surrounding image of the inner hole area with naked eyes and determine the defect. Therefore, professional equipment is used to take the cylindrical expansion image of the inner hole of the commutator for research. Since automatic shooting is difficult to maintain data unification and may produce data redundancy, the data is preprocessed first, and the excessively similar images in the data images are first eliminated to form the original data. Due to the large manpower and material resources needed to shoot the original data and the limited amount of data collected, this section uses some data enhancement methods to expand the scale of the data set and increase the diversity of images in the data set.

Data enhancement technology refers to processing the original image through the image processing method and adding the processed image to the original data set. Data enhancement technology can simulate many changing factors and increase the amount of input image data through spatial, geometric, morphological and other aspects.

The pre-processing process of the original data is as follows:

\[\text{Input: raw data } \text{RawData} \{i\mid i=1, 2, 3\}\]

\[\text{Step1: Put the raw data } \text{RawData} \{i\mid i=1, 2, 3\} \text{ image standardization;}\]

\[\text{Step2: The image data will be normalized } \text{NRawData} \{i\mid i=1, 2, 3\} \text{ image data enhancement is carried out through horizontal flipping, scaling, vertical flipping and other operations.}\]

\[\text{Output: Enhanced data } \text{AugmData} \{i\mid i=1, 2, 3\}.\]

The specific process is shown in Table 2.

3.3. Deep extended convolutional network experiment

Considering that FPN can be a good trade-off between speed and accuracy. In order to further verify the effectiveness of the proposed D_dNet-65 network in FPN, the D_dNet-65 network is compared with the ResNet-50 network. First of all, the D_dNet-65 network will be classified and trained in the article.

In addition, analyzing the experimental results of Table 1, Table 2 and Table 3, it is found that the experimental results of the DetNet-59 network model are always higher than the D_dNet-65 network model based on deep expansion convolution in this paper. Considering that the starting point of this model is to ensure accuracy Under the premise of increasing the rate, improve the target detection speed. The model in this paper adds BN and ReLU layers to the core module, which greatly accelerates the
training speed, the convergence speed, and reduces the computational complexity. It is better than the DetNet-59 network model under these indicators. Although the final accuracy rate is 0.3 to 0.7% lower than that of the DetNet-59 network model, it is still within an acceptable range.

3.4. Lightweight network experiment
Considering that the use of lightweight network to compress the feature map channel will affect the subsequent feature map extraction. In this paper, the original 3969 (81 × 7 × 7) channels are modified to 490 (10 × 7 × 7) channels, and a simple fully connected layer is added at the end. The ResNet-50 network is used as the backbone of the comparison experiment, and a series of comparison experiments are performed on the MSCOCO small data set to verify the effectiveness of the method. The lightweight network is embedded into the R-FCN and Mask R-CNN-50 models respectively, and compared with the experimental results of the four original network structures, as shown in Table 2.

| Table 2. The training results of each model embedded in the lightweight network on MSCOCO (%) |
|-----------------------------------------------|
| Model            | mAP  | AP_s | AP_m | AP_1 |
|------------------|------|------|------|------|
| R-FCN            | 32.11| 19.68| 37.77| 48.02|
| D_59dNet-65 R-CNN| 39.59| 22.19| 42.58| 52.69|

After embedding the lightweight network, the accuracy of Mask R-CNN-50 on the COCO mini-validation set has been reduced, but the corresponding training speed has been significantly improved, and it can be seen from the table that the accuracy of R-FCN is similar. Increased more than 1% from the original. Although compared with the current more advanced Light-Head R-CNN method, it is indeed inferior, but compared with Light-Head R-CNN this local speed-up method, the method in this paper is relatively stable in target detection. And the regression loss in this paper is obviously smaller than the classification loss.

4. Conclusion
The model consists of three parts: backbone network, head network and a new transfer learning method for experimental verification. In the head network, the weight of the head network is reduced by compressing the feature map, thereby improving the speed of network training while ensuring accuracy. This paper verifies that the model has achieved very effective results in terms of accuracy, whether it is a single training or a comprehensive experiment.

Acknowledgements
The work is partially supported by Ministry of Education of Humanities and Social Science project (Grant No.20YJAZH132) and Natural Science Foundation of Chongqing (Grant No.cstc2020jcyj- msmtpX0047).

References
[1] Niu Yifeng, Zhu Yuting, Li Hongnan, Wang Chang, Wu Lizhen. Journal of National University of Defense Technology, 201, 43(01): 117-126.
[2] Qi Zhen, Liu Huwei, He Peng, Feng Weijia. Information Technology and Information Technology, 2020(06): 205-207.
[3] Chu Zenan, Zhao Kai, Ren Junjun. Road Traffic Monitoring Based on UAV Mode [J]. Henan Science and Technology, 2020(04): 100-102.
[4] Peng Bo, Xiaoyu Cai, Jiming Xie, Yuanyuan Zhang. Vehicle Recognition in High-altitude Video Based on Morphological Detection and Deep Learning [J]. Journal of Transportation Systems Engineering and Information, 2019, 19(06): 45-51.
[5] Xiao Jian, Liang Dingkang, Xu Wei, Guo Yufeng. A Autonomous Patrol Inspection System Based on UAVs [J]. Computer Technology and Development, 2019, 29(12): 153-157.
[6] Hu Jihua, Cheng Zhifeng, Zhong Hongzhen, Jing Zehao, Zhang Liyue. Vehicle Detection and
Tracking at Road Intersections Based on UAV Video [J]. Laboratory Research and Exploration, 2019, 38(05): 28-31+106.

[7] Wu Jinting, Zhao Xiaoguang, Yuan Decai. Control Engineering of China, 2019, 26(02): 246-250.

[8] Li Xiaolong, Chen Guoliang, Ge Kaikai. Vehicle Autonomous Plate Reading Based on Quadrotor UAV [J]. Computer and Modernization, 2018(10): 63-67.

[9] Li Shiyao, Lu Yuanbo, Pei Yulong. Research on investigation and identification method of road traffic flow parameters by unmanned aerial vehicle [J]. Forest Engineering, 2018, 34(05): 72-77.

[10] Bo Bo, Xiaoyu Cai, Tao Zhou, Shaobo Li, Youjie Zhang, Lianfei Duan. Automatic recognition algorithm of UAV video vehicle based on morphological analysis [J]. Journal of Chongqing Jiaotong University (Natural Science Edition), 2019, 38(04): 15-22.

[11] Liang Dingkang, Qian Rui, Chen Yihao, Xiao Jian. Design and implementation of UAV patrol inspection and illegal stop system based on vision [J]. Electronic Science and Technology, 2018, 31(05): 76-77+92.

[12] XIANG Yang. UAV image vehicle recognition based on deep learning [J]. Information & Computer (Theoretical Edition), 2018(09): 71-73.