A Scene-Assisted Air-to-Ground Object Detection Method

Xing Liu, Dongfang Yang, Jian Chen, Dawei Sun and Hao He
Xi’an Research Institute of High Technology, Xi’an, China
Email:549645043@qq.com

Abstract. Object detection based on deep learning has achieved outstanding performance in conventional field of view. However, under the air-to-ground field, coupled with the influence of background environment and camera shaking, the application of this kind of detection algorithm greatly decreases its performance, which poses a great risk for performing detection tasks on air platforms such as UAVs. This article discussed the difficulty of deep learning in air-ground detection and put forward a method of using Bayesian inference decision in the later stage of detection. Using the scene information to filter the detection results which made the detection process have the decision-making thinking greatly reduces the false detection rate.

1. Introduction
Air-to-ground object detection technology is a specific application scenario of the visual multi-target detection technology which is the premise to achieve the object tracking. Air-to-ground object detection technology obtains the ground image information by using the visual sensor mounted on the airborne mobile platform (UAV), and then extracts a plurality of ground targets being detected (vehicles, airplanes, etc.) which in the military and civilian areas have important application value and broad technical needs.

There have been many studies on this issue before [1]-[6]. With the development of deep learning, object detection achieves a very good performance under the application of convolution neural network. A regional-based two-stage object detection algorithm and a regression-based one-stage detection algorithm have emerged. R-CNN is the beginning of two-stage object detection algorithm. After that, the Fast R-CNN and Faster R-CNN [7] improved algorithms were proposed respectively, which proposed multi-tasking coupling loss optimization and region proposal network (RPN) respectively. Both of them have greatly improved the accuracy and speed. After that, YOLO (You Only Look Once [8]), a one-stage object detection algorithm, divides grids into 7x7 on the last layer of feature maps. Each grid directly uses regression to get the object category and box. However, this method speeds up at the expense of precision. Later, SSD [9] algorithm was proposed. SSD combined YOLO with Faster R-CNN, which utilized Faster R-CNN's anchor mechanism to extract feature at multiple scales. The accuracy of object detection was greatly improved without excessive loss of speed.

However, most of the current research is based on the conventional environment, which is close to the target. The target occupies a large proportion in the image field of view and the edge contour is clearly distinguishable. In practical applications, especially in the context of military aerial reconnaissance applications, the target is too far away from the field of view so that the edge information is too small or directly unavailable which will lead to serious misjudgment. The current air-to-ground object detection mainly has the following problems:

(1) The object size is small under the air-to-ground detection with the single detection angle. Due to the perspective of the top view, and the long distance from the target, the available feature information of the object detection is small.

(2) Air-to-ground target detection environment is harsh. Occlusion of the background environment...
and changing of the lighting environment will weaken the target characteristics. During the relative movement, due to the camera shaking, the size and shape of the target will change significantly, which increase the environmental noise, seriously affecting the robustness of the detection.

Aiming at the problems existing in the object detection task of air-to-ground, this article is devoted to the research. In order to solve the first problem, we re-collect the air-to-ground scene dataset and use the SSD one-stage detection algorithm to generate different size default boxes on different feature maps to make up for the poor detection accuracy of the YOLO algorithm (especially for small object detection). In order to solve the second problem, our paper proposed a scene-assisted air-ground object detection method. By introducing scene classification information, the target detection process of the current image is assisted (figure 1), which improves the robustness of the object detection from air to ground.

Figure 1. Scene-Assisted network model.

2. Air-To-Ground Datasets
Usually we get the image data with cameras, cell phones and other electronic devices. These devices are very common today which produced a large number of image data such as ImageNet, PASCAL VOC and COCO. Under the contribution of such data, R-CNN series of two-stage detection algorithm and one-stage detection algorithm (like YOLO) can achieve better detection results. Air-to-ground scene is a relatively special kind of scenario. The data acquisition methods are relatively difficult and often require unmanned aerial vehicles and satellites to collect such unusual data. Therefore, the data are scarce. In order to solve the problem of insufficient datasets, we collected data from Google Maps and the existing datasets (such as the Munich vehicle dataset [10]) while utilizing the laboratory drone for aerial photography to reorganize the air-to-ground dataset (figure 2), And finally manually annotate it in the PASCAL VOC format. Categories are mainly included cars, buses, trucks, ships, aircraft and pedestrians.

At the same time, we also produced a separate scene classification datasets. In order to accurately classify the scene, we collected a large number of scene types covering 13 categories including cities, airports, ports, deserts, farmland, lakes, and mountains and so on. Each kind of scene picture number is more than 1000(figure 3).

Figure 2. Examples of object detection datasets
3. Scene Classification and Object Detection

Due to the excellent feature extraction capability of convolutional neural networks, scene classifiers based on deep learning currently have very good classification performance. Taking into account both the real-time and accuracy, we use the most common VGG16 [12] model for scene classification. There are 16 weight layers (13 convolutional layers and 3 fully connected layers) as the basic convolutional structure of VGG16. The convolution kernel size is fixed with 10 convolution kernels (3x3) and 3 convolution kernels (1x1). The activation function uses ReLU and the input size of the picture is 224x224, which takes 13 categories of category scores as output.

The SSD (Single Shot MultiBox Detector) object detection model is inspired by the YOLO model and the faster r-cnn model. The model obtains the final detection result by generating some fixed-size bounding boxes and the probability scores of the target categories in these boxes and then applying non-maximal suppression. The backbone network for SSD feature extraction uses some of the standard image classification networks (e.g VGG16). On this basis, an additional network structure has been added to obtain richer feature information. The model has two main characteristics that are different from the YOLO model:

1. Multi-scale information to do the object detection. Generate more scale information by adding more feature layers, and at the same time predict in different layers;

2. Convolution prediction using anchor mechanism. Different feature layers are divided into m x n grids, and a 3x3 convolution kernel is used for convolution prediction. Each center of a convolution kernel corresponds to a grid, which generates different proportion of detection boxes respectively. It predicts the difference of Category probability of these convolution boxes and the offset from the ground truth to detect the picture.

The loss function of the SSD model is defined as:

\[
L(x, c, l, g) = \frac{1}{N} (L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g))
\]

The objective confidence loss optimization function is defined as:

\[
L_{\text{conf}}(x, p) = - \sum_{i \in \text{Pos}} x_i^c \log(\hat{p}_i^c) - \sum_{i \in \text{Neg}} \log(\hat{p}_i^c) \quad \text{where} \quad \hat{p}_i^c = \frac{\exp(p_i^c)}{\sum_{c} \exp(p_i^c)}
\]

The bounding box regression loss optimization function is defined as:

\[
L_{\text{loc}}(x, l, g) = \sum_{i \in \text{Pos}} \sum_{m \in \{cx, cy, w, h\}} x_j^m \text{smooth}_{l_1}(l^m - \hat{g}_j^m)
\]

Where \(x\) is the predict target. \(p\) is the category confidence level. \(l\) is the predict box parameter, \(g\) is the ground truth box parameter. \(cx, cy, w, h\) are the coordinates and length and width of the prediction box respectively. \(x_i^c\) is an indicator function, which to be 1 when the ith default box and the jth ground truth box match the category, 0 otherwise. \(\hat{g}_j^{cx}, \hat{g}_j^{cy}, \hat{g}_j^{w}\) and \(\hat{g}_j^{h}\) are the offsets relative to the default bounding box.
4. Scene Inference Network

As we know, category detection in deep learning object detection utilizes the softmax function, getting the probability of belonging to a class $i$ and using the maximum a posteriori probability model to derive the class result.

Softmax function:

$$P(c_i | x) = \hat{p}_i^c = \frac{\exp(p_i^c)}{\sum_c \exp(p_c^c)}$$

(4)

Maximum a posteriori probability:

$$h(x) = \arg\max_c P(c | x)$$

(5)

When we take into account the scene information, the output of the entire function is:

$$h^*(x) = \arg\max_{c \in \mathcal{Y}} P(c | s, x)$$

(6)

Where $s \in \mathcal{Z} = \{s_1, s_2, ..., s_n\}$ represents the target category. Based on naive Bayesian derivation, we have come up with algorithms that use scene-assisted detection:

$$h^*(x) = \arg\max_{c \in \mathcal{Y}} P(c | s, x) = \arg\max_{c \in \mathcal{Y}} \frac{P(c | s) P(s | x)}{P(c)}$$

(7)

We have come up with a scene-assisted object detection algorithm. In order to get the probability of different objects in different scenes, we set up an object-scene information inference table:

|       | city  | ocean | desert | ... |
|-------|-------|-------|--------|-----|
| car   | 0.99  | 0.11  | 0.22   | ... |
| bus   | 0.99  | 0.09  | 0.21   | ... |
| truck | 0.99  | 0.10  | 0.26   | ... |
| boat  | 0.21  | 0.99  | 0.02   | ... |
| airplane | 0.94 | 0.43  | 0.01   | ... |
| ...   | ...   | ...   | ...    | ... |

Figure 4. Object-scene information inference table.

Figure 5. No scene-assisted classification results.
The prior class probability $P(c \mid s)$ in the table expresses the proportion of all kinds of samples in the sample space. According to the law of large numbers, when the training samples are sufficient independent and identically distributed samples, the frequencies of all kinds of samples can be used for estimation (figure 4). We obtain this by summarizing the dataset annotation information.

With small size, less features and coupled with the light and background interference in natural scenes, small object detection is particularly prone to misjudgment. For example, when we use a deep learning object detection model to detect small vehicles that are subject to environmental information, we conclude that the probability of suspicion target of 41% for car and 51% for boat. Without any information inference process, using the maximum a posteriori probability model will get the wrong target category: boat. (figure 5)

When using the scene information as an auxiliary inference, we know that the scene of the object is a city. Therefore, using the naive Bayesian inference, the probability of the ship can be reduced immediately and finally we get the correct decision information (figure 6).

$$S_{\text{car}} = 0.41$$
$$S_{\text{boat}} = 0.51$$
$$S_{\text{airplane}} = 0.08$$

**Figure 6.** Vehicle Detection Model Example.

5. **Scene-Assisted Inference**

We tested the SSD target detection model and naive Bayesian inference model on air-to-ground video. Firstly, a scene inference experiment was carried out. Based on the prior information of the scene type detected by VGG16, Naive Bayesian inference was used to determine the SSD target detection results.

**Figure 7.** Combined with scene information. Left shows SSD results, right shows scene-assisted results.
6. Conclusion
At present, the object detection algorithm based on deep learning has achieved good results in the conventional field of view which gradually plays an increasingly important role in practical application. Air-to-ground field of view is a special kind of view whose object size is small, less features, full of interference information. Especially in real-time detection process, the camera shaking, light interference and other factors make the object detection has been difficult to overcome, which cannot guarantee the reliable practical application. By combining SSD target detection network and VGG16 classification network, this paper proposes a scene-assisted air-to-ground object detection algorithm. Adding decision-making in the end of the object detection process, using of scene information, the test results to judge and effectively reduce the rate of false positives.

The scene network and object detection network consists of two independent networks. The next step intends to couple two networks into one network to realize multi-task coupling model. At the same time with the small object characteristics, we intend to further improve the SSD network to make it more suitable for the target detection tasks of the air-to-ground field of view.

7. References
[1] Cheng H Y, Weng C C, Chen Y Y. Vehicle detection in aerial surveillance using dynamic Bayesian networks [J]. IEEE Transactions on Image Processing a Publication of the IEEE Signal Processing Society, 2012, 21(4):2152.
[2] W. Shao, W. Yang, G. Liu, and J. Liu, “Car detection from high-resolution aerial imagery using multiple features,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2012, pp. 4379–4382.
[3] A. Kembhavi, D. Harwood, and L. S. Davis, “Vehicle detection using partial least squares,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 6, pp. 1250–1265, Jun. 2011.
[4] Liu H, Sun F, Fang B, et al. Multimodal Measurements Fusion for Surface Material Categorization [J]. IEEE Transactions on Instrumentation & Measurement, PP (99):1-11.
[5] Liu H, Wu Y, Sun F, et al. Weakly Paired Multimodal Fusion for Object Recognition [J]. IEEE Transactions on Automation Science & Engineering, 2017, PP (99):1-12.
[6] Z. Chen et al., “Vehicle detection in high-resolution aerial images via sparse representation and superpixels,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 1, pp. 103–116, Jan. 2016.
[7] Ren S, Girshick R, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
[8] Joseph Redmon, Santosh Divvala, Ross Girshick, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. 2015:779-788.
[9] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2015:21-37.
[10] Liu K, Mattyus G. Fast Multiclass Vehicle Detection on Aerial Images [J]. IEEE Geoscience & Remote Sensing Letters, 2015, 12(9):1938-1942.
[11] Goldszmidt, Moises. Bayesian Network Classifiers [J]. Machine Learning, 1997, 29(2-3):131-163.
[12] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.