OBJECT DETECTION AND TRCACKING BASED ON CONVOLUTIONAL NEURAL NETWORKS FOR HIGH-RESOLUTION OPTICAL REMOTE SENSING VIDEO

Biao Hou, Jingliang Li, Xiangrong Zhang, Shuang Wang, and Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China,
Xidian University, China (Email: avcodec@hotmail.com)

ABSTRACT

Object detection algorithms, from high-resolution optical remote sensing images, have been booming from the last few years. However, object tracking for high-resolution optical remote sensing video is a challenging task due to the large number and small size of objects. In this paper, we propose an object detection and tracking method based on deep convolutional neural networks for wide swath high-resolution optical remote sensing videos. The proposed method firstly segments each frame of a video into sub-samples using a sliding window of fixed size. In order to detect the objects appearing at the edge of the sliding window efficiently, we use an overlapping sliding window sampling method. Further, we design a network fusing region of interests (RoIs) of the previous and current frames to track the objects occurred in the previous frames of the video. RoIs of previous frame are applied directly to the feature layer of the current frame. Finally, for each frame, we merge the detection and tracking results of sub-samples by non-maximum suppression (NMS) method. The experimental results on our dataset demonstrate the validity and generality of the proposed detection algorithm.

***Index Terms*—**Object detection, object tracking, high resolution optical remote sensing, deep convolutional neural network.

1. INTRODUCTION

Object detection and tracking has been one of the most important applications of remote sensing. Along with the development of sensor technology and aerospace industry, the resolution of remote sensing data has increased dramatically. In high resolution of remote sensing data, the contours and the textures of objects are clearly observed and informative for object detection and tracking, and object architectures are closer to characteristics of human ocular perception than SAR images. Object detection and tracking for optical remote sensing data becomes more significant.

Object detection and tracking is an important branch in the field of computer vision, and many mature algorithms are proposed. Previous works on object detection and tracking can be divided into two categories, for natural videos and remote sensing videos. In the field of remote sensing, remote sensing data brings some new challenges in object detection and tracking using deep convolutional neural networks, e.g. complex background information, small scale and multi-oriented property of objects. A lot of techniques [1-2] have been presented to detect and track remote sensing objects. However, most of the detection and tracking methods for remote sensing data are based on moving objects and the size of the videos is small.

In order to detect and track static and moving objects together in the wide swath high-resolution optical remote sensing videos, we proposed a detection and tracking method based on an object detection algorithm for natural images. The object detection algorithms for natural images can also be divided into two categories, two stage algorithms and one stage algorithms. In the first category, the detection algorithms [3-4] are based on region proposal methods. These algorithms firstly get region of interests by a region proposal method and then get detection results of the proposals by a classifier. For example, Grishick et al proposed a fast and precise detection algorithm called Faster R-CNN [4], which includes a region proposal network (RPN) for generating region proposals and a network using these proposals to detect objects. In the second category, the detection algorithms [5-6] have only one stage without region proposal module. For example, Yolo [6] uses only one network to achieve the region proposal task and classification task. By comparison, one stage algorithms have faster detection speed, and two stage algorithms can obtain better detection results for small objects.

Because the size of objects in high-resolution optical remote sensing videos generally is small, our detection and tracking method is based on Faster R-CNN. We firstly segment each frame of a video into sub-sample in order to adapt videos with any size. Secondly, these sub-samples are fed into the network, where RoIs of the previous and current frames can be fused to get the detection and tracking of results these sub-samples. Finally, we get the detection and tracking results for the frames by merging the results of sub-samples of each frame. The experimental results show the proposed method can obtain better results for wide swath high-resolution optical remote sensing videos.

978-1-5386-9154-0/19/$31.00 ©2019 IEEE

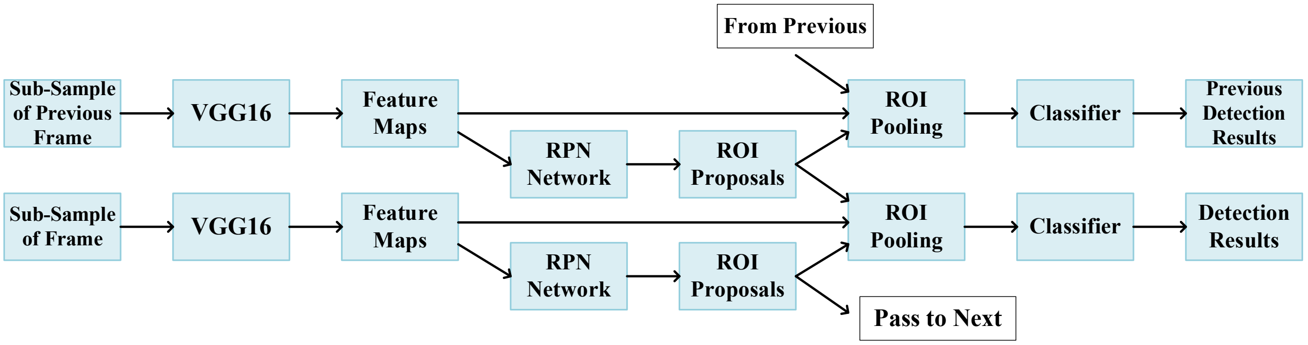

Fig.1. The network structure for our object detection and tracking method

2. THE PROPOSED METHOD

The whole network is derived from the Faster R-CNN which achieves a state-of-the-art performance on object detection from natural images. The proposed method only cares about testing branch. For training, we follow the process of Faster R-CNN.

2.1. Segment each frame of the video

For a wide swath high-resolution optical remote sensing video, we segment each frame of the video into sub-samples using a fixed size sliding window of $K \times K$ (e.g. 300×300), where K is a hyper-parameter that is the size of sliding window. The other hyper-parameters are *H, W* and *S,* where *H* and *W* are height and width of the frame, respectively, and *S* is the distance between two consecutive positions of the sliding window. The overlapping *L* (*L*=*K*-*S*) should be larger than the max size of objects in all frames. Consider the width axis: the window starts on the leftmost part of the input frame and slides by steps of one until it touches the right side of the frame. The same logic applies for the height axis. The number of the sub-samples for each frame *N* will be equal to:

$$N = (\left\lceil \frac{H-K}{S} \right\rceil + 1) \times (\left\lceil \frac{W-K}{S} \right\rceil + 1) . \qquad (1)$$

All of the N sub-samples are fed into the detection and tracking network to get results, and then calculate the detection and tracking results of this frame by merging the results of all sub-samples.

2.2. Detection network design

The proposed method is derived from the Faster R-CNN. The core of the Faster R-CNN is the region proposal network (RPN) that shares features with the detection network. The RPN is a fully convolutional network that simultaneously predicts a set of rectangular object bounds and objectless scores at each position. By sharing features, Faster R-CNN can be trained end-to-end to generate high-quality region proposals. In recently years, algorithms on image classification achieve a great success. Generally, for object detection and tracking, we need to locate the object firstly compared with image classification. So the bottleneck of the object detection and tracking may be location. The detection and tracing results generally are affected by the quality of region proposals.

In a high-resolution optical remote sensing video, for any ground-truth box of moving object in a frame, Intersection-over-Union (IoU) overlap ratio is higher than 0.8 with ground-truth box in two consecutive frames. The static objects have an IoU overlap ratio higher than 0.95. So we can pass the RoIs of the previous frame to current frame in order to track the objects appearing in the previous frame. The structure of the proposed network is shown in Fig.1. The whole network can be divided into four modules which include feature extraction, generating region proposals, fusing RoIs and classification. For fusing RoIs, we fuse RoIs of the previous and current frames. In the video, there are T frames, and each frame has N sub-samples. $x_{i,j}$ refers to the *j*-th sub-sample in *i*-th frame ($i \le T, j \le N$). We append the RoIs of $x_{i-1,j}$ to the RoIs of $x_{i,j}$, and then the fusing RoIs are fed into RoI Pooling layer. Because of fusing the RoIs in the feature layer, we compute the feature maps from the sub-frame only once. This method avoids repeatedly computing the convolutional features.

2.3. Merge detection results

From detection and tracking network, we get the results of all sub-frames for each frame. We need six values to describe each object, including two coordinate values of left top corner, two coordinate values of right bottom corner, the label of the object and a confidence score that estimates probability of object belonging to the category. What we need to do is merging the detection and tracking results of each frame. Because the inputs of the detection network are sub-samples, so the four coordinate values describing object are positions in the sub-sample not in the frame. Before we merge the results, we need to map the four coordinate values in the sub-sample to the coordinate values in the frame.

For each frame, we merge the detection and tracking results of sub-samples by non-maximum suppression (NMS) [7] method. For NMS, it firstly sorts all detection and tracking object results on the basis of their scores. The object M with the maximum score is selected and all other objects with a significant overlap with M are suppressed. This process is recursively applied on the remaining boxes.

3. EXPERIMENTS

3.1. Dataset and base network description

In our experiment, we test only airplane object. Our dataset includes RSOD-Dataset [8-9] and video data from Jilin-1 Video-03 remote sensing satellite. RSOD-Dataset is an open dataset for object detection in remoting sensing images. The dataset includes aircraft, oiltank, playground and overpass. We use only the aircraft dataset, which has 4993 aircrafts in 446 images. Jilin-1 Video-03 is a Chinese commercial remote sensing satellite for high definition video. The main task of this satellite is to acquire visible light video data of high-resolution earth observation in the world. The resolution of the video is 0.92m. The width and height of the video are 12000 and 5000 pixels. We randomly crop 200 images with size 300×300 which include the aircraft objects from the videos. All the 446 aircraft images in RSOD-Dataset and the 200 cropped images are used for training. For testing video, we only detect a cropped region of the video. The width and height of the cropped video are 3200 and 1800 pixels. The training images are not included in the test video.

Our network uses VGG-16 [10] model as block bone. We

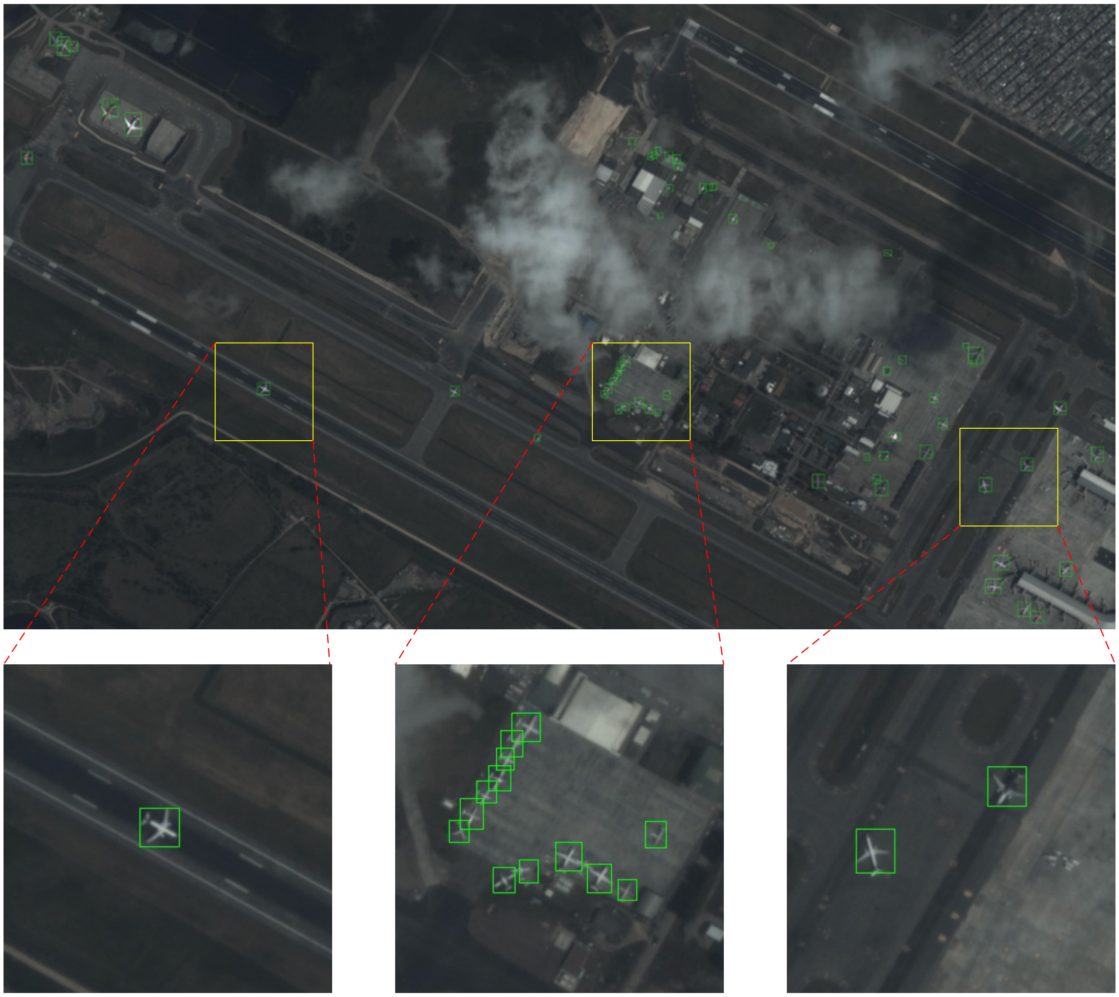

Fig.2. The samples of detection results, the green rectangle denote the detection windows generated by our method. Each of yellow rectangles represents a 300*300 region.

use the VGG-16 pre-trained on ImageNet to initialize our network, while the initialization of the rest layers uses Xavier initialization. The size of sliding window is 300×300 . And S that is the distance between two consecutive positions of the sliding window is set to 250. The NMS thresholds in the detector and the fusing module are all set to 0.3. We train on one Nvidia P5200 GPU for 15 iterations. The learning rate is initialized to 0.001 and is divided by 10 every 8 iterations. We use a weight decay of 0.0005 and a momentum of 0.9.

3.2. Detection results

We use Mean Average Precision (MAP) to evaluate our method on the test video. Tablet.1 shows the mAP of the detection methods. Compared with Faster R-CNN, our method is better at object detection due to the fusion of RoIs. Fig.2. shows the detection results of the 1-th frame. The green rectangles denote the detection windows generated by our method. Each yellow rectangle represents a 300×300 region. Fig.3. shows the detection and tracking results of 30-th, 50-th and 70-th frames for objects covered by cloud. The size of each image cropped from the frame is 100×100 . Fig.4. shows the of 30-th, 50-th and 70-th frames for static objects. The size of each image cropped from the frame is 300×300 .

Table 1. MAP for optical remote sensing video.

Mehtod	Without RSOD data	With RSOD data
Our	**78.80**	**89.23**
Faster R-CNN	78.13	88.09

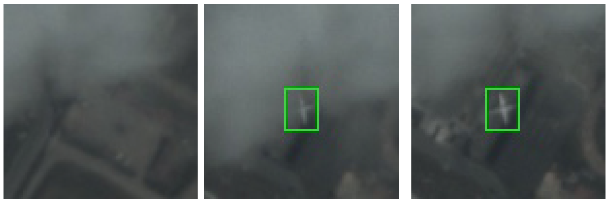

Fig.3. The detection and tracking results of 30-th, 50-th and 70-th frames for object covered by cloud. The size of each image cropped from the frame is 100×100 .

4. CONCLUSION

This paper proposes an efficient object detection and tracking method for wide swath high-resolution optical remote sensing video. The proposed method aims to detect and track objects from high-resolution optical remote sensing videos. Our method firstly segments each frame of a video into sub-samples, so we do not need to worry the challenges about the size of the data. The method tracks the objects by fusing the RoIs of current frame and previous frame. The experimental results show its ability of solving the object detection problem effectively and accurately.

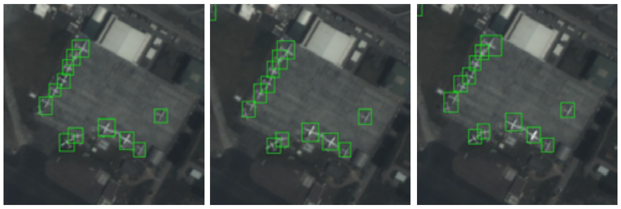

Fig.4. The detection and tracking results of 30-th, 50-th and 70-th frames for static objects. The size of each image cropped from the frame is 300×300 .

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61671350.

REFERENCES

[1] J. B. Poisson, H. Oriot, and F. Tupin, “Performances analysis of moving target tracking in circular SAR,” in *Proc. IRS*, Dresden, Germany, 2013, pp. 531–536.

[2] Z. Xu, Y. Zhang, H. Li, H. Mu, and Y. Zhuang, “A New Shadow Tracking Method to Locate the Moving Target in SAR Imagery Based on KCF,” *Int. Conf. CSPS*, Singapore, 2017, pp. 2661-2669.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” In *ECCV*, Zurich, Switzerland, 2014, 346-361.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *NIPS*, Montréal Canada, 2015, pp. 91–99.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *ECCV*, Amsterdam, Netherlands, 2016, pp. 21–37.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *CVPR*, Las Vegas, NV, USA, 2016, pp. 779-788.

[7] R. Rothe, M. Guillaumin, and L. Van Gool, “Non-maximum suppression for object detection by passing messages between windows,” In *ACCV*, Singapore, 2014, pp. 290-306.

[8] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, “Accurate object localization in remote sensing images based on convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[9] Z. Xiao, Q. Liu, G. Tang, and X. Zhai, “Elliptic Fourier transformation based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images,” *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 618–644, 2014.

[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” In *ICLR*, San Diego, CA, 2015.