Context Semantics for Small Target Detection in Large-Field Images with Two Cascaded Faster R-CNNs

To cite this article: Lanfang Kong et al 2018 J. Phys.: Conf. Ser. 1069 012138

View the article online for updates and enhancements.

Related content

- An Efficient Method of Small Targets Detection in Low SNR
  Y L Wang, J M Dai, X G Sun et al.
- Real-time vehicle detection and tracking in video based on faster R-CNN
  Yongjie Zhang, Jian Wang and Xin Yang
- Understanding of Object Detection Based on CNN Family and YOLO
  Juan Du
Context Semantics for Small Target Detection in Large-Field Images with Two Cascaded Faster R-CNNs

Lanfang Kong, Xiuhong Zhu and Guoyou Wang
National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan, China.
Email: kong875387859@outlook.com, M201672536@hust.edu.cn, gywang@mail.hust.edu.cn.

Abstract. Computer vision and image processing techniques have been widely applied to power transmission line inspection. However, the successful detection of small targets in large scenes is still challenging due to their low resolution and poor feature representation. Existing methods, such as multi-scale image pyramid, multi-scale feature pyramid and multiple heterogeneous feature fusion, etc. can extract more representative features of small objects, but they usually require high computation cost. In this paper, we propose an effective two cascaded Faster R-CNN strategy, which is based on multi-scale features and semantic information between the objects and the background, to address the small target detection in large scenes. Specially, we detect large object candidate proposals that may contain small objects at first and then map them to the original images to detect the small-sized targets on the high resolution regions. Experiments show that our strategy could lead to higher (83.0%) accuracy of small target detection and recognition than the one-stage Faster R-CNN (78.3%) on the dataset of aerial images.

1. Introduction

With the wide-range application of the unmanned aerial vehicle (UAV) in the inspection of transmission line, computer vision is increasingly applied in the image understanding and electrical devices detection. In transmission line inspection, small-sized electrical devices, such as grading ring and suspension point are crucial to fault diagnosis. It is also important to remove nests which are usually built in the electricity pylons because they can potentially hinder the maintenance of the grid operation. Thus, detecting electrical device by analysing UAV obtained images has become a trend in intelligent grid.

However, the aerial images acquired by UAV are complex. Many of our objects: suspension point, grading ring and nest are extremely small compared to the large-field scenes. For instance, targets are commonly below 300×100 pixels while the original images are sized at 4800×2704 pixels, which makes the unsuccessful detection of small-sized electricity devices.

Traditional electrical devices detection algorithms mainly adopt human-designed features like SIFT [1], HOG [2], LBP [3], and edge detection operators. However, these features need manual designs and lack generalization ability because they usually aim at specific scenes or categories.

On the other hand, deep learning can obtain amazing results in image recognition and object detection. However, few people applied these methods into electrical devices detection in scene images before. So, how is the performance? To delve this question, we consider region-based CNN methods, including R-CNN [4], Fast R-CNN [5] and Faster R-CNN [6]. With the help of UAV obtained images, we have the dataset to train very deep convolutional networks.
In this paper, we propose a cascaded Faster R-CNN to improve the accuracy of small target detection in scene images. Inspired from the semantic segmentation process before object detection [7], we perform Faster R-CNN to localize the large regions that contain contextual information of the small targets at first and then map these candidate regions to the original images to improve the feature map resolution for small objects. Small targets can be finally localized and detected in the candidate regions precisely. Details of our approach can be found in section III and experiments are reported in section IV.

2. Related Work

Since 2012, training deep networks, especially convolutional neural networks, has dominated computer vision field, including object detection. For object location, there are two kinds of methods: sliding window based and region proposal based.

The sliding window based methods use a sliding window to search on the input images and estimate whether or not each window contains the objects. However, this method is high in computation.

Recent great progresses made in object detection are followed by the deep learning pipelines that learn feature representation from region of interest (RoI) and then perform classification to complete detection tasks such as R-CNN, Fast R-CNN and Faster R-CNN. However, for small targets in the aerial images such as grading rings, framing hammer, spacer and suspension gear, region proposal methods usually fail to detect them due to the poor-quality appearance and structure.

People devote some efforts into researching small targets detection. HyperNet [8] raised a Hyper Feature to combine deep, coarse information with shallow information to make features more representative and generate more accurate region proposals. But the accuracy is not improved a lot; FPN [9] proposed a method called feature pyramid which combines top feature and bottom feature, and predicts on each feature layer. It improves the detection accuracy slightly but the detection speed is much slower; Feature-fused SSD [10] proposed a multi-level feature fusion method in SSD[11] baseline to improve small objects detection accuracy. It is also harmless to detection speed. However, multi-scale feature representation works like a black box and we could not guarantee the constructed features are beneficial to our tasks.

Another simple solution[15] is to enhance the resolution of input images. However, naively enlarging the scale of input images can bring problems of heavy computation and high cost on hardware during training and testing periods because of the extremely high feature dimensions.

The famous Generative Adversarial Nets [16] which was proposed in 2014 provides us another path to solve this problem. In [17], GAN is applied in small targets detection. We can also use this method to do data augmentation in case of lacking samples of small targets.

Thus, we propose a cascaded Faster R-CNN as our method to perform semantic comprehension at first which can help us to acquire large targets’ location in UAV images and then search for small targets on the large target areas. This approach is effective because in the second stage, the original small targets are at high resolution and no longer small compared to the background to make detection more accurate.

3. Small Target Detection in UAV Images using Two-stage Faster R-CNNs

Faster R-CNN was proposed to reduce the computational burden of region proposals generation. It breaks through the bottleneck of computing time and guarantees the ideal accuracy at the same time. It contains two modules: one is called the Region Proposal Network (RPN) and the other is Fast R-CNN detector. The first module RPN is a fully convolutional network, which introduces 3 aspect ratios of anchors to generate candidate proposals; the second module takes these proposals as input and then refine them to predict. However, if we directly apply Faster R-CNN to detect small targets in the aerial images, it tends to fail to achieve high accuracy because in the deep layers of networks, the feature maps corresponding to small targets may become coarse and lose their response ability.

In order to locate small targets in UAV images, we split this progress into two stage. Firstly, we should generate object candidates that may contain small objects. One way to generate these candidates is to perform semantic segmentation like [7]. We adopt neural-based method Faster R-
CNN to generate object candidates at first. It extracts several basic large objects precisely which our small targets are contained in. Then we feed the extracted regions to another Faster R-CNN to search for small targets precisely. The pipeline illustration is shown in figure 1.

3.1. Train with RPN
In order to generate region proposals, we use a small window sliding on the final feature map and map the corresponding feature to a low-dimension vector, which is 256-d for ZFnet [12] and 512-d for VGG16. This vector is input to two sibling fully connected layers: bounding box regression layer and bounding box classification layer. Faster R-CNN also introduces anchor to adapt to multi-scale object detection. It adopts 3 different aspect ratio (1:1, 1:2, 2:1) combinations and 3 scales to predict windows.

3.2. Approximately Joint Training
The Faster R-CNN discusses 3 plans to train networks with features shared. In this paper, we adopt the approximately joint training strategy for simplicity. At first, the candidate objects are fed into the second Faster R-CNN training network. When training Fast R-CNN detector, region proposals, which are generated by RPN, are seen as fixed. The RPN loss and the Fast R-CNN are combined for the shared layers during the backward propagation.

4. Experiments
In this section, we evaluate the two cascaded Faster R-CNN on our dataset and compare the detection performance of different methods for small objects carefully.

4.1. Dataset
Our dataset taken by UVA on power transmission lines is provided by Wuhan NARI Co. Ltd, which is owned by China's State Grid Corp. In figure 2, we demonstrate some sampled images from the dataset. All the images are in high resolution and contain multiple categories of small-sized objects. It is clear that there exist large range of variations in scale and pose. Our dataset has 3700 images. We use the dataset into training-validation-test split, which is 8:1:1. Each image is in size of 4800×2704 pixels.
Our experiment focuses on following 4 small targets: nest, ID label (denote the number or identification of electricity pylon), suspension point (the device to suspend insulators) and grading ring. As previously described in the section 3, object proposals, which contain contextual information, need to be acquired in the first stage, we also labelled annotations of 2 electrical equipment: electricity pylon and insulator.

**Figure 2.** Some samples in our dataset, where blue, green, yellow, orange, red and pink are ground-truth annotations of pylon, insulator, nest, ID label, suspension point and grading ring, respectively.

4.2. **Setup**

We use a pre-trained ImageNet [14] model, VGG16 [13] to extract the features and then to train the two-stage Faster R-CNN. The CNN architecture is implemented by PyTorch. We randomly sample 300 images per batch for training. In order to fit it in the GPU memory, it is resized based on the ratio 1200/ w if the width of image is greater than 2000 pixels, otherwise w and h retain unchanged, where w and h are the width and height of the image, respectively. We run the stochastic gradient descent (SGD) solver for 10,000 iterations with a base learning rate of 0.001. In addition, IoU overlap is set at 0.7 for all region proposals with a ground-truth box as positive. Considering the circumstance that positive samples are extremely scarce, we bias sampling toward positive windows.

4.3. **Large Objects Detection Results in the First Stage**

In the first stage, we perform Faster R-CNN on the original scene images to do semantic segmentation, thus to extract the main objects which could possibly contain our small targets. The average precision of the first stage and the overall pipeline are calculated and shown in the table 1. It is no doubt that we get a very high accuracy in pylon and insulator detection because they have large scales and discriminative features, which can be learned easily.

4.4. **Small Targets Detection Results in the Second Stage**

In the second stage, we adopt the same detection process to evaluate the small targets detection performance. Followed by the first stage, the pre-detected contextual regions are fed into the second detection network. Table I shows the final detection result.

**Table 1.** The Results of Cascaded Faster R-CNN

| Table Head | The First Stage | Overall Accuracy of Two Stage |
|------------|----------------|------------------------------|
|            | Electricity Pylon | Insulator | nest | ID label | suspension point | grading ring |
| AP         | 0.951            | 0.923     | 0.752 | 0.732     | 0.923         | 0.914       |
| mAP        | 0.937            |           |       |           |               |             |
| Test Time (ms) | 10               |           |       |           |               | 12          |

4.5. **Comparison with one-stage Faster R-CNN**

In order to prove our method is effective for small target detection, we also experiment the end-to-end Faster R-CNN [15], which directly detects small targets on the original images. The results of these two different strategies are illustrated as figure 3. For objects in close shots, they can be detected easily and the bounding boxes are more accurate. However, for the objects in distant views, our
method can detect them as well, which is much better than direct application of Faster R-CNN. We also list the average accuracy of each class in the Table II.

![Image](image_url)

**Figure 3.** The comparison of the detection results of two strategies. Yellow is the result of direct application of Faster R-CNN and blue is the result of our system.

| ID Label               | Nest          | Suspension Point | Grading Ring   | Test Time (ms) |
|------------------------|---------------|------------------|---------------|---------------|
| mAP                    | 0.682         | 0.876            | 0.883         | 10            |
| mAP                    | 0.732         | 0.923            | 0.914         | 12            |
| mAP                    | 0.783         | 0.830            |               |               |

The accuracy of suspension point and grading rings is much better than the accuracy of the nest and ID label for the reason that the samples of nest and ID label are extremely limited and the size of both two objects are much smaller than the other two classes.

4.6. Comparison with Different Anchors

The aspect ratios for different objects in scene images vary a lot, so we experiment the effects of different anchors on small target detection. We enumerate the following 3 types of anchors: 1:1, 1:2, 2:1; 1:1, 1:3, 3:1; 1:1, 1:5, 5:1. Results for these 3 are shown in Table III. We can conclude that the anchor of aspect ratio: 1:1, 1:2, 2:1 has the best performance compared to the other aspect ratios.
Table 3. The result comparison of different anchors

| Aspect ratio | 1:1 | 1:2 | 2:1 | 1:1 | 1:3 | 3:1 | 1:1 | 1:5 | 5:1 |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Nest         | 0.752 | 0.746 | 0.741 |     |     |     |     |     |     |
| ID Label     | 0.732 | 0.711 | 0.704 |     |     |     |     |     |     |
| Susp. Point  |     | 0.923 | 0.924 | 0.918 |     |     |     |     |     |
| Grading Ring |     | 0.914 | 0.915 | 0.910 |     |     |     |     |     |
| mAP          |     | 0.830 | 0.824 | 0.818 |     |     |     |     |     |

5. Conclusion
In this paper, we describe a two-cascaded Faster R-CNNs to address the problem of the small target detection in large scene images taken by UAV over the transmission lines. By mapping the candidate objects with contextual information to the original images, we increase the resolution of the candidate regions, thus improving the final detection performance with comparable testing time. The experiments demonstrate that our method superior to the end-to-end Faster R-CNN.

6. Acknowledgement
Image Processing & Intelligent Key Laboratory in Huazhong University of Science and Technology supports this work. Assistance from my tutor Guoyou Wang is gratefully acknowledged. These authors contributed to the work equally and should be regarded as co-first authors.

7. References
[1] Lowe, D.G., Object Recognition from Local Scale-Invariant Features (SIFT). 1999.
[2] Dalal, N. and B. Triggs. Histograms of Oriented Gradients for Human Detection. in IEEE Computer Society Conference on Computer Vision & Pattern Recognition. 2005.
[3] Ojala, T., M. Pietikainen, and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002. 24(7): p. 971-987.
[4] Girshick, R., et al., Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2013: p. 580-587.
[5] Girshick, R., Fast R-CNN. Computer Science, 2015.
[6] Ren, S., et al., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017. 39(6): p. 1137.
[7] Garcia, G.M., et al. Semantic segmentation priors for object discovery. in International Conference on Pattern Recognition. 2017.
[8] Kong, T., et al., HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. 2016: p. 845-853.
[9] Lin, T.Y., et al. Feature Pyramid Networks for Object Detection. in IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[10] Cao, G., et al., Feature-Fused SSD: Fast Detection for Small Objects. 2017.
[11] Liu, W., et al., SSD: Single Shot MultiBox Detector. 2015: p. 21-37.
[12] Simonyan, K. and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014.
[13] Zeiler, M.D. and R. Fergus, Visualizing and Understanding Convolutional Networks. 2013. 8689: p. 818-833.
[14] Deng, J., et al. ImageNet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2.
[15] Ku, J., et al., Joint 3D Proposal Generation and Object Detection from View Aggregation. 2017.
[16] Goodfellow, I.J., et al. Generative adversarial nets. in International Conference on Neural Information Processing Systems. 2014.
[17] Li, J., et al., Perceptual Generative Adversarial Networks for Small Object Detection. 2017: p. 1951-1959.