SCALABLE MULTI-CLASS GEOSPATIAL OBJECT DETECTION IN HIGH-SPATIAL-RESOLUTION REMOTE SENSING IMAGES

Gong Cheng, Junwei Han, Peicheng Zhou, Lei Guo

School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China

ABSTRACT

In this paper we present a conceptually simple but surprisingly effective multi-class geospatial object detection method based on Collection of Part Detectors (COPD), which can be easily scaled to a larger number of object classes. The presented COPD is composed of a set of representative and discriminative part detectors, where each part detector is a linear support vector machine (SVM) classifier trained using a weakly supervised learning method that only requires image labels indicating the presence of objects for the training data. Here, each part detector corresponds to a particular viewpoint of an object class, so the collection of them provides a feasible solution for rotation-invariant and simultaneous detection of multi-class geospatial objects. Comprehensive evaluations on high-spatial-resolution remote sensing images and comparisons with a number of state-of-the-art approaches demonstrate the effectiveness and superiority of the presented method.

Index Terms—Object detection, remote sensing, detectors, image analysis, image recognition

1. INTRODUCTION

Automated object detection and recognition in remote sensing images is a core requirement for high-level scene understanding and semantic information extraction. In recent years, the rapid development of remote sensing technology has increasingly facilitated us the acquisition of remote sensing images with high spatial resolution, but how to automatically detect and locate geospatial objects in remote sensing images is still a fundamental yet challenging problem in computer vision field.

A number of recent works in literatures have proposed various methods for different object detection tasks [1-10]. For example, Aytekin et al. [1] proposed a novel airport runway detection method by using an Adaboost learning algorithm employed on a large set of textural features. Bo and Jing [2] developed a simple region-based airplane detection method by using region segmentation and binary image processing. Grabner et al. [5] developed an online boosting algorithm for car detection from large-scale aerial images. Segl and Kaufmann [6] presented an approach for the detection of small objects by combining supervised shape classification with unsupervised image segmentation. In addition, the detection methods for some other object classes such as ships [4, 7] and landslide [3, 8], etc., have also been explored in the literature.

Although the topic of geospatial object detection has been deeply investigated, most of the current object detection methods are still dominated by the detection of a single object category and fewer concerns have been given to scalable multi-class objects detection. Furthermore, the features extracted in most existing individual detectors are customized for the particular type of objects, which are not able to generate the generally good results across different categories of objects. Therefore, this group of approaches is more difficult to scale towards large numbers of object classes. Generally, a large-scale remote sensing image always contains multiple classes of interesting objects instead of only a single one, so it is an important and promising issue to develop a scalable multi-class objects detection method.

More recently, part-based methods have shown promising potential for object detection [11-13] and image classification [14, 15] on natural scene (non-overhead) images. However, these methods cannot be directly used to detect geospatial objects from remote sensing images because they are difficult to handle the rotation variations problem. In this paper we present a conceptually simple but surprisingly effective multi-class object detection method based on Collection of Part Detectors (COPD), which can be easily scaled to a larger number of object classes. Comprehensive evaluations on high-spatial-resolution remote sensing images and comparisons with a number of state-of-the-art approaches demonstrate the effectiveness and superiority of the presented method.

2. MULTI-CLASS GEOSPATIAL OBJECT DETECTION METHOD

2.1. Method overview

Figure 1 gives an overview of the presented multi-class geospatial object detection method. It is mainly composed of two stages: COPD training and object detection. In the COPD training phase, the approach learns a moderate

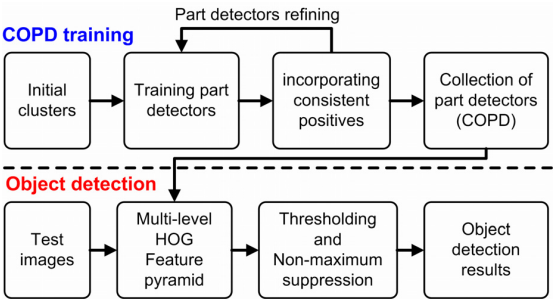

Figure 1. Overview of the presented multi-class geospatial object detection method.

number of seed-based part detectors using a weakly supervised learning method where only image labels indicating the presence of targets need to be provided for the training data. Here, each part detector is a linear support vector machine (SVM) classifier, which corresponds to a particular viewpoint of an object class, so the collection of them provides a feasible solution for rotation-invariant and simultaneous detection of multi-class geospatial objects. To be specific, we firstly pick a set of seeds to serve as initial clusters, where each cluster corresponds to a part detector needed to be trained. Then, we train a set of part detectors using an iterative procedure [11, 13, 15] that alternates between refining part detectors and incorporating consistent positives. In the meantime careful cross-validation is applied at each step to prevent over-fitting. Given K seeds, we can finally obtain a COPD that is composed of K seed-based classifiers. In the object detection stage, given a new test image, we first run all detectors simultaneously on the input image, in HOG [16] feature pyramid space, to obtain the response and potential object class label for each sliding-window. Then, multi-class object detection is implemented by thresholding the responses and eliminating repeated detections via non-maximum suppression [11, 13].

2.2. COPD training

When training a COPD for multi-class object detection, the input is composed of a "positive image dataset" P in which each image contains at least one target of interest and a "negative image dataset" N in which all images do not contain any targets of the given object classes. The COPD training is performed in terms of the following steps [12-15]:

(1) Pick a set of seeds from P to serve as initial clusters, where each seed corresponds to a particular viewpoint of an object class. Given K seeds, we have K part detectors to be trained.

(2) Train a linear SVM classifier $\Gamma_k = (\boldsymbol{w}_k, b_k)(k = 1, \cdots, K)$ for each cluster, in HOG [16] feature space, using image patches within the cluster as positive examples and all hard negative examples of N as negative examples. It is noted that when we train the SVM classifiers first time, this can be seen as a special situation of [12] because each cluster contains one image patch only, i.e. the picked seed. For each cluster, learning the parameters $\boldsymbol{w}_k$ and b_k amounts to optimizing the following objective function:

$$(\boldsymbol{w}_k, b_k)^* = \arg\min_{(\boldsymbol{w}_k, b_k)} \left\{ \begin{array}{l} \frac{1}{2}\|\boldsymbol{w}_k\|^2 + C \sum_{x^+ \in X_k^+} h(\boldsymbol{w}_k^{\mathrm{T}} \Phi(x^+) + b_k) \\ + C \sum_{x^- \in X_k^-} h(-\boldsymbol{w}_k^{\mathrm{T}} \Phi(x^-) - b_k) \end{array} \right\}, \quad (1)$$

where X_k^+ and X_k^- denote the sets of positive examples and negative examples of kth cluster. $\Phi(x^+)$ and $\Phi(x^-)$ denote the feature vectors of positive example x^+ and negative example x^- obtained by concatenating all the HOG feature vectors within the examples. $h(\tau) = \max(0, 1-\tau)$ is the standard hinge loss function that allows us to use hard negative mining technique to cope with millions of negative examples [11-13]. C is a constant and we set $C = 0.1$ in our work.

(3) Run $\Gamma = \{\Gamma_k\}_{k=1}^K$ on P to obtain new clusters by selecting the top-m high-scoring patches for each part detector. In our work, we set $m = 5$ to keep each cluster having a high purity.

(4) Repeat the steps of (2) and (3) $L1$ iterations until convergence is reached. In our experiment we empirically set $L1 = 2$.

2.3. Object detection

Given a test image I, its HOG [16] feature pyramid $H(I)$ is first constructed. Then, for each sliding-window S, we run all detectors $\Gamma = \{\Gamma_k\}_{k=1}^K$ on $H(I)$ to obtain its response $R(S)$ and potential object class $O(S)$. $R(S)$ is defined as the maximum response of all part detectors:

$$R(S) = \max_{(\boldsymbol{w}_k, b_k) \in \Gamma} \left(\boldsymbol{w}_k^{\mathrm{T}} \Phi(S) + b_k \right), \quad (2)$$

where $\Phi(S)$ denotes the feature vectors of sliding-window S by concatenating all the HOG feature vectors within it, $\boldsymbol{w}_k^{\mathrm{T}} \Phi(S) + b_k$ is the response of sliding-window S of detector Γ_k. $O(S)$ is defined by the class of part detector with the maximum response. Finally, multi-class object detection is implemented by thresholding the responses using a threshold ρ (the optimal threshold can be derived from the highest F1-measure), and each hypothesis is defined by a response, a potential object class and a bounding box.

In practice, when we use the above described detection approach solely, a number of sliding-windows near each instance of an object are likely to be detected as the targets, which results in multiple overlapping detections for a single object. We therefore apply non-maximum suppression [11,

13] to eliminate repeated detections. In brief, the bounding boxes are sorted by their responses, and we greedily select the highest scoring ones while removing those that are at least 50% covered by a previously selected bounding box.

3. EXPERIMENTS

3.1. Dataset description

In our experiments, we used the task of detection of three different types of objects to evaluate the performance of the presented method. These three classes of objects are airplanes, ships, and storage tanks. We collected 300 1000×800 high-spatial-resolution remote sensing images from Google Earth and divided them into four independent datasets: a "negative image set" containing 50 images, a "positive image set" containing 60 images, an "optimizing set" containing 40 images, and a testing set containing 150 images. The "negative image set" and "positive image set" were used for the COPD training, the "optimizing set" was used for parameter optimization and the testing set was used for testing the performance of the presented method. From the "positive image set", we picked 13 seeds (eight for airplane, four for ship, and one for storage tank) for initializing clusters. Furthermore, we also labeled 146 airplane targets, 62 ship targets, 363 storage tank targets from the optimizing set, and 561 airplane targets, 214 ship targets, 1326 storage tank targets from the testing set, respectively, which are used for ground truth. The sizes of the targets in these images vary from about 45×45 to 140×140 pixels.

Figure 2 shows the visualization of 13 detectors (1st row) for airplane, ship, and storage tank classes, respectively, where brighter "pixel" represents bigger weight and vice versa. Their corresponding top-five high-scoring positives are also shown from 2nd to 6th rows.

3.2. Experimental results and comparisons

We consider a detection to be correct if its bounding box overlaps more than 50% with the ground truth bounding box, otherwise the detection is considered as a false positive. In addition, if several bounding boxes overlap with a same single ground truth bounding box, only one is considered as true positive and the others are considered as false positives. We adopted the standard Precision–Recall curve (PRC) [19] to quantitatively evaluate the performance of an object detection system.

In the implementation of multi-class object detection, to address the problem that the sizes of targets may be different in images, each image is represented by an eight-level HOG [16] feature pyramid and each octave contains five levels (i.e. for l th level, the sub-sampling factor is $2^{(l-1)/5}$). We follow the construction in [16] to extract the HOG feature for each pyramid level. Specifically, we

Figure 2. The visualization of 13 detectors (1st row) for airplane, ship, and storage tank classes, respectively. Their corresponding top-five high-scoring positives are shown from 2nd to 6th rows.

partition the image at each pyramid level into non-overlapping cells of 6×6 pixels and use nine orientation bins to accumulate a one-dimensional histogram of gradient orientations over pixels in each cell. Then, each 2×2 neighbourhood of cells is grouped into one block (with a stride of one cell) and a robust normalization process based on 2-norm is run on each block to provide greater invariance to local illumination and spatial deformation, which finally forms a 36-dimensional HOG feature vector. Rather than using the 36-dimensional vector directly, in this work we project it onto a lower 32-dimensional space as described in [29] and [33]. In addition, the size of each part detector is 8×8 blocks, i.e. an 8×8 HOG descriptors. Consequently, the sizes of image patches that each part detector can detect are 54×54, 62×62, 71×71, 82×82, 94×94, 108×108, 124×124, and 143×143, respectively, which correspond to eight different image scales.

Using the pre-trained COPD, as illustrated in Figure 2, we performed multi-class objects detection on our testing dataset which contains 561 airplane targets, 214 ship targets, and 1326 storage tank targets. To evaluate the presented method, we compared it with some state-of-the-art object detection methods [12, 17, 18]. The method in [17] is based on bag-of-words (BOW) feature description and SVM classifier, which is called BOW-SVM in this paper. The method in [18] is based on spatial sparse coding bag-of-words and SVM classifier, which is called SSCBOW in this paper. The method in [12] is based on a set of exemplar-based SVMs, which is called Exemplar-SVMs in this paper. For fair comparison, we (1) adopted the same training dataset and test dataset for various approaches, and used the same seeds as our method to train Exemplar-SVMs [12]; (2) implemented these three comparison methods by adopting multi-scale scanning window scheme for each test image which was similar to our multi-level HOG feature pyramid. Following the works of [17, 18], the vocabulary size was set to 400 and 450 for SSCBOW and BOW-SVM, respectively. Figure 3 shows the quantitative comparison results of four different methods measured by PRC. Especially for our method, the false alarm rate (i.e. $1-Presision$) is nearly zero before *Recall* is bigger than 0.8, which demonstrates

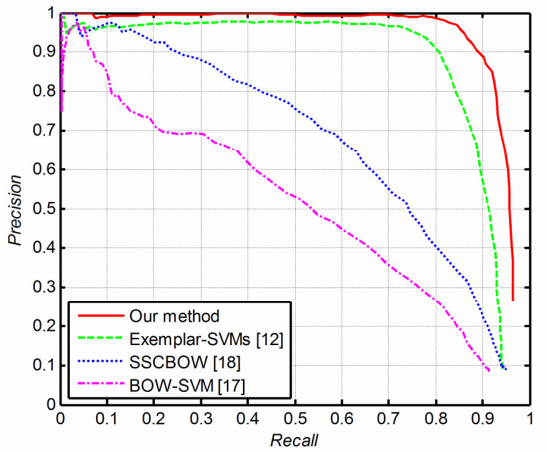

Figure 3. Precision–Recall curves of the presented method and three state-of-the-art approaches.

that the presented multi-class objects detection method is highly competitive.

4. CONCLUSIONS

In this paper, we presented an effective multi-class object detection method, which can be easily scaled to a larger number of object classes. Comparisons with a number of state-of-the-art approaches demonstrated the effectiveness and superiority of the developed method. In the future work, we will (1) share part detectors to reduce the overall number of model parameters and improve the computational efficiency; (2) extend the presented method for more visual recognition tasks, such as geographical image classification.

5. ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China under Grants 61005018 and 91120005, Program for New Century Excellent Talents in University under grant NCET-10-0079, and China Postdoctoral Science Foundation under Grant 2014M552491.

6. REFERENCES

[1] Ö. Aytekin, U. Zöngür, and U. Halici, "Texture-based airport runway detection," *IEEE Geosci. Remote Sens. Lett.,* vol. 10, no. 3, pp. 471-475, May 2013.

[2] S. Bo, and Y. Jing, "Region-based airplane detection in remotely sensed imagery," in *Proc. CISP*, 2010, pp. 1923-1926.

[3] G. Cheng, L. Guo, T. Zhao, J. Han, H. Li, and J. Fang, "Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA," *Int. J. Remote Sens.,* vol. 34, no. 1, pp. 45-59, Jan. 2013.

[4] C. Corbane, L. Najman, E. Pecoul, L. Demagistri, and M. Petit, "A complete processing chain for ship detection using optical satellite imagery," *Int. J. Remote Sens.,* vol. 31, no. 22, pp. 5837-5854, Nov. 2010.

[5] H. Grabner, T. T. Nguyen, B. Gruber, and H. Bischof, "On-line boosting-based car detection from aerial images," *ISPRS J. Photogramm. Remote Sens.,* vol. 63, no. 3, pp. 382-396, May 2008.

[6] K. Segl, and H. Kaufmann, "Detection of small objects from high-resolution panchromatic satellite imagery based on supervised image segmentation," *IEEE Trans. Geosci. Remote Sens.,* vol. 39, no. 9, pp. 2080-2083, Sept. 2001.

[7] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.,* vol. 48, no. 9, pp. 3446-3456, Sept. 2010.

[8] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Trans. Geosci. Remote Sens.,* vol. 49, no. 12, pp. 4928-4943, Dec. 2011.

[9] G. Cheng, J. Han, L. Guo, X. Qian, P. Zhou, X. Yao, and X. Hu, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 85, pp. 32-43, 2013.

[10] J. Han, P. Zhou, D. Zhang, G. Cheng, L. Guo, Z. Liu, S. Bu, and J. Wu, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 89, pp. 37-48, 2014.

[11] L. Bourdev, and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proc. ICCV*, 2009, pp. 1365-1372.

[12] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *Proc. ICCV*, 2011, pp. 89-96.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 32, no. 9, pp. 1627-1645, Sept. 2010.

[14] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that Shout: Distinctive Parts for Scene Classification," in *Proc. CVPR*, 2013.

[15] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. ECCV*, 2012, pp. 73-86.

[16] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886-893.

[17] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.,* vol. 7, no. 2, pp. 366-370, Apr. 2010.

[18] H. Sun, X. Sun, H. Wang, Y. Li, and X. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," *IEEE Geosci. Remote Sens. Lett.,* vol. 9, no. 1, pp. 109-113, Jan. 2012.

[19] M. K. Buckland, and F. C. Gey, "The relationship between recall and precision," *J. Am. Soc. Inf. Sci.,* vol. 45, no. 1, pp. 12-19, Jan. 1994.