LW-ODF: A LIGHT-WEIGHT OBJECT DETECTION FRAMEWORK FOR OPTICAL REMOTE SENSING IMAGERY

Xin Wu[1,2], *Danfeng Hong*[3,4], *Pedram Ghamisi*[5], *Wei Li*[1,2], *Ran Tao*[1,2]

[1]School of Information and Electronics, Beijing Institute of Technology (BIT), Beijing, China
[2]Beijing Key Laboratory of Fractional Signals and Systems, Beijing Institute of Technology (BIT), Beijing, China
[3]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[4]Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany.
[5]Machine Learning Group, Exploration Division, Helmholtz Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, Freiberg, Germany.

ABSTRACT

In this paper, we propose to extract the multi-scaled and rotation-insensitive deep features to address the issues of object multi-solutions and rotations in geospatial object detection. To this end, we develop a novel object detection framework where a rotation-insensitive convolution neural network is applied for extracting multi-scaled and direction-insensitive feature representation and then the learned features can be fed into the ensemble classifier learning with fast feature pyramid. Such a non-end-to-end learning strategy intuitively reduces the computational cost without the additional performance loss, yielding an effective and efficient light-weight object detection framework. Experimental results conducted on the NWPU VHR-10 dataset demonstrate that the proposed framework outperforms several state-of-the-art baselines.

***Index Terms*—** Deep learning, direction-insensitive, geospatial object detection, light-weight, multi-scaled, optical remote sensing imagery.

1. INTRODUCTION

Recently, the processing and analysis of optical remote sensing imagery (RSI) [1, 2, 3, 4] have achieved a growing interest, particularly for geospatial object detection. However, optical RSI inevitably suffers from all kinds of deformations, e.g., variabilities in viewpoint, scaling, and direction, which have always been challenging. The feature representation of the traditional manual feature extraction methods is incomplete [5]. They not only fail to simultaneously extract the local and global features [6] of the object but also generate deep semantic information with high discrimination. In recent years, the explosive growth of deep learning algorithms [7, 8, 9] with automatic learning ability dramatically improve object detection performance in optical RSIs. However, the single receptive field of a convolutional layer and the direction sensitivity of convolutional kernel limit its generalization ability, while the strong distinguishability of features requires extremely high computational cost. Therefore, it is important to develop a light-weight feature learning method.

More specifically, a light-weight object detection framework (LW-ODF) is proposed to extract multi-scaled and rotation-insensitive features by the rotation-insensitive convolution neural network and to more effectively and efficiently detect objects with the fast feature pyraimd. The workflow of the LW-ODF is illustrated in Fig. 1. Two main technical contributions proposed in this paper are: 1) The low-levels feature maps processed by the scale and direction insensitive modules are fed to the AdaBoost classifier to generate a light-weight detector that is insensitive to scale and direction for optical RSIs. 2) In the testing phase, fast feature pyramid generated by a power law instead of finely sampled are adopted to implement object detection in optical RSIs without sacrificing performance.

2. METHODOLOGY

The purpose of this work is to develop a light-weight object detection framework for optical RSIs. It is insensitive to large-range of scale and direction variation. The flow chart of the proposed LW-ODF is illustrated in Fig. 1, which consists of three phases: network selection, multi-scaled and direction-insensitive feature extraction, and detection with fast feature pyramid. The details of our framework are discussed in the following sections.

2.1. Base Network : VGG-16

The LW-ODF is an improved convolution channel features (CCF) [10]. Similarly, we use a VGG-16 network as the pre-trained network and perform the fine-tuning on the NWPU VHR-10 dataset [11] for further feature extraction .

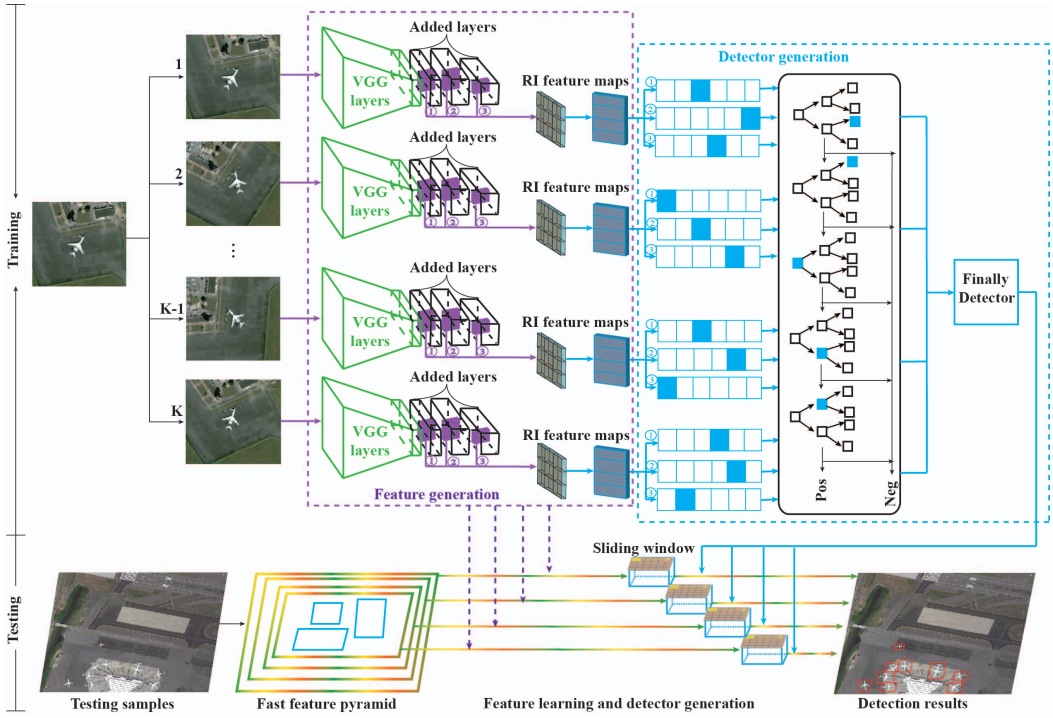

Fig. 1. The workflow of the proposed LF-ODF.

To our knowledge, the VGG-16 is computationally expensive, as massive parameters need to be updated. Consequently, we selected the low-level $conv\_3$ of the VGG-16 as input, and additionally add a 7-layer network inspired by the inception module in order to discriminate the feature representation together with a relatively low computational cost. Table 1 details our network architecture in each layer.

2.2. Scale and Direction insensitive Module

1. Direction insensitive Module: Geospatial objects inevitably suffer from various deformations, such as shift, rotation. This leads to the performance degradation when using VGG-16-like base networks. Inspired by [12, 13], we rotate the training samples with different angles and feed them into the network training. The main idea to construct the rotation-invariant features is to enforce these rotated samples to share a same feature representation. For that, the new object function can be defined as follows

$$J(\theta,\varphi) = \min_{\theta,\varphi} \sum_{i=1}^{n} \widetilde{\omega}_i exp(-\theta_\varphi(x_i)y_i) + \frac{\lambda}{2N} \sum_{x_i \in X} \parallel F_a(x_i) - \overline{F_a(g_\phi x_i)} \parallel_2^2, \quad (1)$$

where φ and θ are sub-classifier and network weights, respectively. $\widetilde{f}(x_i)$ denotes the output of the classifier. x_i is ith sample and y_i is the corresponding label. $\overline{F_a(g_\phi x_i)}$ represents the averaged feature map defined as $\frac{1}{K}\sum_{j=1}^{K} F_a(g_\phi x_i)$, where $F_a(x_i)$ is the feature map of rotated sample.

2. Scale insensitive Module: The single receptive field of a convolutional layer usually fails to detect the objects with varying resolutions. This motivates us to extract the multi-scaled feature maps [14] with the use of the different-size convolutional kernels. We experimentally and empirically select three intermediate layers as the final feature representation. More specifically, The features in the shallow layer are used for the small-scale objects while the deeper features might be used for the large-scale ones.

2.3. Detection with fast feature pyramid

The image pyramid is a simple but effective method used to solve the scale problem in object detection. This can be achieved by sliding a fixed-sized window over a finely sampled image pyramid, leading to an expensive computational cost in general. To speed up the feature pyramid generation, we mathematically estimate a scaling factor by following a power law proposed by [5] to fast and automatically perform the feature pyramid. The resulting expression is

$$\mathbf{P}(\mathbf{F}, s) \approx \Omega(\mathbf{R}(\mathbf{F}, s)) = \mathbf{R}(\mathbf{F}, s) \cdot s^{-\lambda_\Omega}, \quad (2)$$

where $\mathbf{F}$ denotes the feature maps, and $\mathbf{R}(\mathbf{F}, s)$ is a re-sampled feature of $\mathbf{F}$ by s. λ is a scaling factor to be estimated. The size of the sliding window is set to 3×3, 6×3, 3×6 according to three aspect ratios, i.e., 2:1, 1:1, and 1:2.

Table 1. The network architecture in feature extraction of our LW-ODF.

No.	Layer Setting	Patch size /Stride	$\#1 \times 1$	$\#1 \times 1$ /$\#3 \times 3$	$\#1 \times 1$ /$\#5 \times 5$	3×3 /1×1	Active	Padding	Output
0	VGG-16/ Conv_3								$56 \times 56 \times 256$
1	max	$2 \times 2/2$						valid	$28 \times 28 \times 256$
2	inception1	Stride=1	128	128/192	32/96	32/64		same	$28 \times 28 \times 480$
3	relu◁						RELU	valid	$28 \times 28 \times 480$
4	max	$3 \times 3/2$						valid	$14 \times 14 \times 480$
5	inception2	Stride=1	192	96/208	16/48	32/64		same	$14 \times 14 \times 512$
6	relu◁						RELU	valid	$14 \times 14 \times 512$
7	max◁	$3 \times 3/2$						valid	$7 \times 7 \times 512$

Table 2. Quantitative performance comparisons on NWPU VHR-10 dataset. The best is shown in bold.

Method	COPD [11]	BOW-SVM [15]	Exemplar [16]	ACF [5]	YOLO1 [17]	YOLO2 [18]	CCF [10]	Ours
AP	0.5490	0.1394	0.4644	0.5399	0.6584	0.7846	0.6282	**0.8125**
Mean Times/s	2.00	3.5	2.4	0.67	0.15	**0.12**	1.9	0.92
Baseball diamond	0.8259	0.3215	0.7023	0.7592	0.8428	0.9221	0.8215	**0.9507**
Ground track field	0.8525	0.0210	0.2535	0.7320	0.8729	0.9657	0.8005	**0.9700**
Basketball court	0.3528	0.0033	0.4528	0.3901	0.8195	**0.8432**	0.6000	0.7900
Airplane	0.6230	0.0902	0.8389	0.6470	0.5992	0.8667	0.7200	**0.8957**
Ship	0.6910	0.3712	0.3700	0.5207	0.6175	0.8329	0.5891	**0.8571**
Storage tank	0.6459	0.3587	0.7102	0.7990	0.2786	0.4198	**0.8620**	0.6476
Tennis court	0.3235	0.0121	0.3028	0.2980	0.5734	**0.6400**	0.3610	0.6250
Harbor	0.5580	0.1364	0.3295	0.5434	0.7421	0.7887	0.6300	**0.8002**
Bridge	0.1496	0.0004	0.2328	0.3700	0.7195	**0.8790**	0.4551	0.8259
Vehicle	0.4408	0.0795	0.4515	0.3400	0.5187	0.6879	0.4429	**0.7623**

3. EXPERIMENTS

3.1. Dataset Description

The NWPU VHR-10 dataset is a publicly available benchmarking 10-class geospatial object detection dataset. These ten classes of objects are obtained from two different resolution data. One is 715 color images acquired from 0.5 to 2m Google Earth and another is 85 pan-sharpened color infrared images from the 0.08m Vaihingen dataset. In this dataset, all objects labels were manually annotated with axis-aligned bounding boxed. In the experiment, the NWPU VHR-10 dataset was expanded by performing a rotation operation (i,e., 0° to 180° at a 45° interval), and a color space conversion of HSV to avoid over-fitting. All experiments were implemented by a PC with an Intel single-core i7 CPU, NVIDIA GTX-1070 GPU (4 GB RAM) and 32 GB RAM.

3.2. Results and Analysis

Table 2 lists the quantitative results of the eight state-of-the-art baselines. Accordingly, we can have the following conclusions: 1) BOW-SVM is only robust to the spatial variations of the objects, but ignores the spatial background modeling between local features. The generalization capabilities of Exemplar-SVM, COPD, and ACF are limited due to the sensitivity of the HOG to the object rotation. 2) YOLO reconstructs the object detection into a single regression problem, which can speed up the object detection task, but small and dense objects are not well detected and located. It can reduce the probability of detecting the background as an object, but it causes a low recall rate. Therefore, it is necessary to use multiple feature maps to complete the detection in parallel. 3) The AP value of CCF after adding the scale and direction insensitive modules increased by 5 percent to the original CCF, which directly verifies the effectiveness of the algorithm for optical remote sensing object detection. In addition, the local connection and weight-sharing of convolution layer replace the full-connection layer, breaking the network's limitation to the size of the input image.

Fig. 2 visualizes some detection results, where each class is marked by a rectangular box of different colors, that is, the yellow and orange rectangle stand for missed objects and false objects, respectively. It is clear to show in this figure that a small number of unknown objects in the sea is wrongly recognized as a bridge, while those with low detection scores can be removed by the pull-up threshold. Some vehicles, e.g., trucks, fail to be detected. This can be explained by insufficient representation capability.

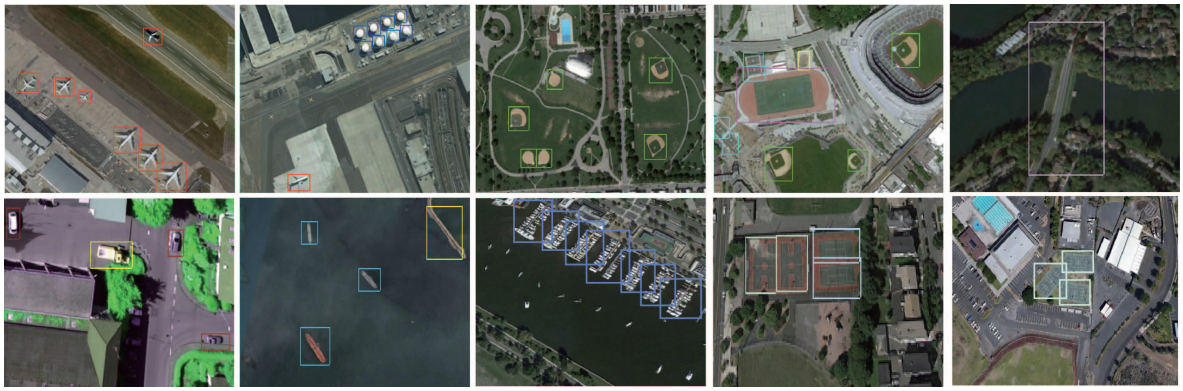

Fig. 2. Some detection results (false positive in yellow, true positive in other colors) by using the proposed method on the NWPU VHR-10 dataset.

4. CONCLUSION AND OUTLOOK

This paper proposes a light-weight object detection framework, called LW-ODF. LW-ODF is capable of automatically extracting the features instead of those hand-crafted methods [19, 20], which is robust to the scale and direction of the objects owing to the use of rotation-invariant CNN and the design of the multi-scaled features. The experimental results on the NWPU VHR-10 dataset show the superiority and effectiveness of the proposed LW-ODF. The AP value is higher than YOLO2 only about 3%, the possible reason is that the feature representation ability is insufficient. Therefore, we will improve the feature representation capabilities by utilizing a more powerful network, e.g., ResNet, or enhancing the quality of the input images [21], or introducing the multimodal data [22] in the future.

5. REFERENCES

[1] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & progressive learning from high-dimensional data for multi-label classification," in *Proc. ECCV*, 2018, pp. 478–493.

[2] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, 2019.

[3] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2019.

[4] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. Zhu, "Learnable manifold alignment (lema): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, 2019.

[5] D. Piotr, A. Ron, B. Serge, and P. Pietro, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, 2014.

[6] D. Hong, N. Yokoya, and X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, 2017.

[7] P. Ghamisi and N. Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial nets," *IEEE Geosci. Remote Sens. Lett.*, vol. PP, no. 99, pp. 1–5, 2018.

[8] X. Wu, D. Hong, P. Ghamisi, W. Li, and R. Tao, "Msri-ccf: Multi-scale and rotation-insensitive convolutional channel features for geospatial object detection," *Remote Sens.*, vol. 10, no. 12, pp. 1990, 2018.

[9] B. Zhang, M. Zhang, J. Kang, D. Hong, J. Xu, and X. Zhu, "Estimation of pmx concentrations from landsat 8 oli images based on a multilayer perceptron neural network," *Remote Sens.*, vol. 11, no. 6, pp. 646, 2019.

[10] B. Yang, J. Yan, L. Zhen, and S. Li., "Convolutional channel features," in *Proc. ICCV*, 2015.

[11] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, no. 1, pp. 119–132, 2014.

[12] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," vol. 54, no. 12, pp. 7405–7415, 2016.

[13] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm Detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, 2019.

[14] D. Hong, Z. Pan, and X. Wu, "Improved differential box counting with multi-scale and multi-direction: A new palmprint recognition method," *Optik*, vol. 125, no. 15, pp. 4154–4160, 2014.

[15] J. Tu, H. Sui, W. Feng, K. Sun, and L. Hua, "Detection of damaged rooftop areas from high-resolution aerial images based on visual bag-of-words model," *IEEE Geosci. Remote Sens. Lett.*, vol. PP, no. 99, pp. 1–5, 2016.

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

[17] J. Redmon, S. K. Divvala, , R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016, pp. 779–788.

[18] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. CVPR*, 2017, pp. 6517–6525.

[19] D. Hong, W. Liu, J. Su, Z. Pan, and G. Wang, "A novel hierarchical approach for multispectral palmprint recognition," *Neurocomputing*, vol. 151, pp. 511–521, 2015.

[20] D. Hong, W. Liu, X. Wu, Z. Pan, and J. Su, "Robust palmprint recognition based on the fast variation vese–osher model," *Neurocomputing*, vol. 174, pp. 999–1012, 2016.

[21] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "Stfnet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, 2019.

[22] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "Cospace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, 2019.