AIRCRAFT DETECTION FROM REMOTE SENSING IMAGE BASED ON A WEAKLY SUPERVISED ATTENTION MODEL

$JinshengJi^1$, $TaoZhang^1$, $ZhenYang^2$, $LinfengJiang^1$, $WeilinZhong^1$, $HuilinXiong^1$

1. Shanghai Key Lab. of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, 200240 Shanghai, China
2. Jiangxi Science and Technology Normal University, 330013 Nanchang, China

ABSTRACT

Aircraft detection from high resolution remote sensing image is a challenging task due to the lack of annotation information, large-scale image size, and sparse distribution of aircraft. Recently, some convolutional neural network(CNN) based methods explore the attention based weakly supervised way to localize the aircraft without manual annotation information. However, the detection results are not satisfied with high false detection ratio. In this paper, a method of utilizing weakly supervised attention model to localize the multi-scale aircrafts is presented, in which the attention model is carried out in a weakly supervised way. Compared with other CNN based method, the proposed attention model can obtain more accurate attention map and localize the aircrafts more precisely. The experimental results on two challenging datasets demonstrate that the proposed method achieves higher detection accuracy and lower false detection ratio than other methods.

Index Terms— remote sensing, aircraft detection, visual attention, weakly supervised learning, convolutional neural network

1. INTRODUCTION

With the rapid development of remote sensing technologies, the optical images with high spatial resolutions have been widely used in many practical applications, such as natural hazards, land planning, urban monitoring, and so on [1]. Among these applications, aircraft detection from very high resolution(VHR) remote sensing images attracts great attentions recently. Especially, both the civil and military fields have urgent needs to localize the aircrafts precisely and efficiently. However, due to the varieties of scales, appearance, and complex background, aircraft detection is still a challenging task. There has been developing a variety of aircraft detection methods to localize these multi-scale aircrafts of the remote sensing images. Liu and Shi [2] proposed the model of sparse coding and target-oriented saliency for aircraft detecting. Yao et al. [3] explored an algorithm of utilizing rotation invariant features and sparse coding to recognize the aircraft.

Recently, many convolutional neural network(CNN) based methods achieved impressive performance for the task of recognition problems [4, 5], such as object detection [2], image recognition [4] and so on. Additionally, due to the lack of manual annotations, many researchers seek to develop the weakly supervised methods which just utilize the image labels instead of relying on these annotations. Zhang et al. [6] utilized a coupled neural network to localize the aircraft by iterating mining and augmenting the training data set from the original image in a weakly supervised way. Overall, weakly supervised aircraft detection from VHR remote sensing images requires low false detection ratio in addition to high accuracy. Without the accurate manual annotation information, many methods achieves high accuracy at the expense of high false detection ratio. Therefore, the key is how to localize the aircraft precisely and efficiently.

To handle these problems above, we develop a new method of detecting aircrafts from the VHR remote sensing images based on weakly supervised attention(WSA) model. Specifically, the proposed detection model can obtain more accurate attention map and achieve lower false detection ratio.

2. PROPOSED METHOD

2.1. Attention Network

To learn more class specific information about the images, global average pooling(GAP) is utilized for training the attention model. Compared with the fully connected layer, the GAP is more native to the convolution structure by enforcing correspondences between feature maps and categories. Therefore, the three fully connected layers of VGG-16 [5] are replaced with GAP to obtain the attention network, as shown in Figure 1.

Inspired by the great generalization of the model pretrained by ImageNet [4], model trained by aircraft images of scene classification datasets has captured specific information about the aircrafts and can also be utilized for weakly supervised aircraft detection. Based on this, two datasets of

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

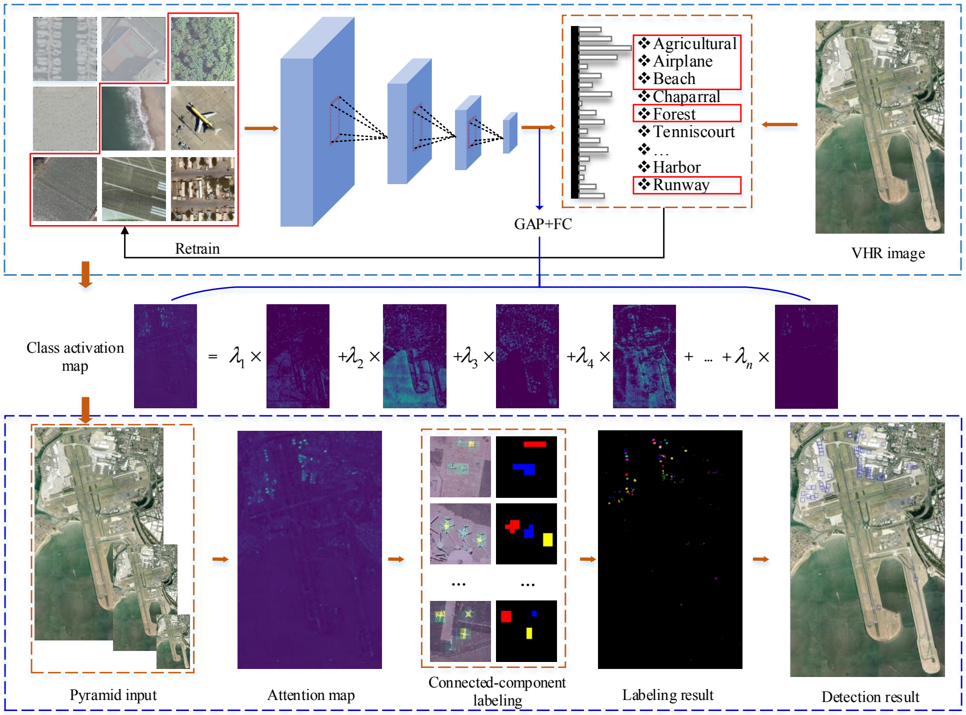

Fig. 1. The framework of proposed weakly supervised attention model for aircraft detection. The convolutional layer of $conv_{5\_3}$ is utilized to calculate the class activation map.

UC-Merced [7] and PatternNet [8] are utilized for training the attention model.

$$GAP(X) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} X_{i,j}^{l}, \qquad (1)$$

where $X \in \mathbb{R}^{N \times W \times H}$, N is channel number of current convolutional layer, the W and H are the width and height of feature map X^l respectively.

As demonstrated in Figure 1, classes that achieving high responses on the test image are reserved for retraining the attention model by(2):

$$L_{cl} = -\sum_{i=1}^{C} y_i log(\widehat{y}_i) \qquad (2)$$

where the y_i is the ground truth label, $\widehat{y}_i$ is the predicted result.

As the scale and annotation information of aircrafts are unknown, relying on the single scaled image can not achieve satisfied results.Additionally, the complex background also decreases the performance of detection model. Therefore, pyramid images are utilized for detecting these multi-scale aircrafts, as shown in Figure 1. Generally, neurons of the deeper convolutional layers, such as $conv_{5\_3}$, can capture high level semantic information and ignore the irrelevant background. Each convolutional layer in CNN is composed of d feature maps with dimension of $w \times h$. For precise aircraft localization, the class activation map(CAM) [9] is calculated using(3). In detail, weights on the aircraft category of fully connected layer after the GAP are selected and the weighted summation is then utilized to localize aircraft. The CAM calculated on the category of airplane has high responses for aircraft compared with other irrelevant background.

$$M_{cam} = \sum_{i=1}^{d} \lambda_i^c M_i^c \qquad (3)$$

where d is the number of feature maps, λ_i^c is the weight of M_i^c on category c.

2.2. Aircraft Localization

As mentioned above, the attention map for localizing aircrafts has been calculated by WSA model. In addition, pyramid inputs provide more details for attention model to localize the multi-scale aircrafts precisely. Different from regarding each responsive neuron of the attention map as a region proposal, connected component labeling is applied to segment these neural responses as described in Figure 1. This can decrease the number of redundant proposals significantly.

Then, the binary result of whole attention map is utilized to localize the detected aircrafts by mapping it to origin im-

Table 1. The detailed information of two VHR remote sensing image datasets used in this paper.

Dataset	Image size(pixel)	Spatial resolution(m)	Aircraft size(pixel)	Aircraft number	Cover area(km^2)
Tokyo Airport	6528*7488	1.0	220∼3800	65	48.75
Sydney Airport	4992*8256	1.0	300∼3500	46	41.2

Table 2. Comparison of different methods.

Dataset	Method	WSA-Pyramid	WSA	LOCNet-C	SPMK	UFL
Tokyo Airport	FPR	**7.35%(5/68)**	7.81%(5/64)	20.80%(47/227)	12.38%(13/106)	16.92%(11/65)
	MR	3.08%(2/65)	9.23%(6/65)	**1.54%(1/65)**	20.00%(13/65)	29.23%(19/65)
	AC	96.92%(63/65)	90.77%(59/65)	**98.46%(64/65)**	80.00%(52/65)	70.77%(46/65)
	ER	**10.43%**	17.04%	22.33%	32.38%	46.15%
Sydney Airport	FPR	20.00%(11/55)	**13.95%(6/43)**	48.54%(100/207)	34.33%(23/67)	45.24%(19/43)
	MR	**4.35%(2/46)**	19.57%(9/46)	10.87%(5/46)	43.48%(20/46)	65.22%(30/46)
	AC	**95.65%(44/46)**	80.43%(37/46)	89.13%(41/46)	56.52%(26/46)	34.78%(16/46)
	ER	**24.35%**	33.52%	59.41%	77.81%	110.46%

age. Each block of the binary map is one region proposal and can be cropped from origin image. Compared with other methods, the proposed WSA model obtains more accurate responses for the test image and less incorrect region proposals. Based on this, the detected proposals are fed into the well trained attention network and the proposals that achieve lower confidence than 0.5 on the category of aircraft are abandoned.

3. EXPERIMENT

3.1. Dataset and Network

We evaluate our method on two benchmark datasets: Tokyo Airport and Sydney Airport [6]. The detailed description about the two datasets are demonstrated in Table 1. It can be find that aircrafts of the two datasets vary greatly in size, ranging from 220 to 3800 pixels. This increases the difficulty of localizing them from VHR remote sensing image. To obtain the WSA model, images from the correlated task as scene classification [7,8] are collected for training. Furthermore, all aircrafts samples of the two datasets are utilized for testing. The backbone network of proposed weakly supervised attention is based on VGG-16 [5].

The experiments were run on a computer with Intel Xeon E5 CPU, 64GB main memory, and four Nvidia Titan GPUs with 48GB memory. The implementation environment is under the Pytorch, which is a popular framework for studying the deep neural network.

3.2. Experimental Result

To evaluate the effect of proposed WSA model, we follow the criteria used in [6] to measure the detection result, i.e., false positive ratio(FPR), missing ratio(MR), accuracy(AC), and error ratio(ER). The four criteria are defined as follows:

$$\begin{aligned} FPR = \frac{N_{fda}}{N_{da}},\ MR = \frac{N_{ma}}{N_a}, \\ AC = \frac{N_{tda}}{N_a},\ ER = MR + ER \end{aligned} \tag{4}$$

where N_{fda} is the number of falsely detected aircraft, N_{da} is the number of detected aircraft, N_{ma} is the number of missing aircraft, N_{tda} is the number of truly detected aircraft, N_a is the number of aircraft.

Comparisons with some state-of-the-art methods are conducted on the two datasets, as shown in Table 2. Experiments were also conducted to evaluate the effectiveness of pyramid inputs for the WSA model. The aircraft proposal achieves intersection overlap with the ground truth bounding box greater than 0.5 is regarded as a truly detected aircraft. In addition, the AC/FPR curves under different probability threshold p are demonstrated in Figure 4. It is obvious to find that WSA-Pyramid with pyramid inputs outperforms the WSA with single scaled input and achieves the largest AUC(area under curve) compared with other methods. As demonstrated in Table 2, pyramid inputs can provide more detailed information for obtaining higher detection accuracy, and this also brings slightly increase of FPR. The WSA-Pyramid achieves lower FPR and MR compared with LOCNet-C [6], SPMK [10] and UFL [11].

4. CONCLUSION

In this paper, the WSA model for weakly supervised aircraft detection from VHR remote sensing image is proposed. The concise structure of the detection model simplifies the procedure of aircraft localization. Meanwhile, the accurate attention map also avoids the time-consuming process of generating and selecting multiple proposals. To evaluate its per-

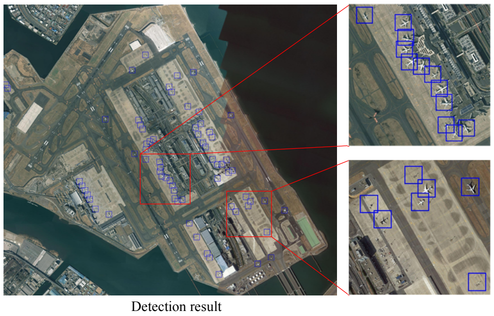

Fig. 2. The detection result of Tokyo airport dataset.

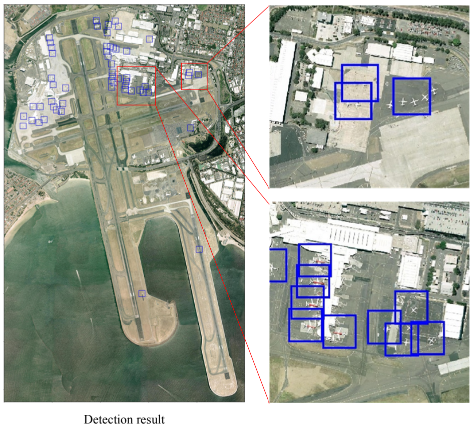

Fig. 3. The detection result of Sydney airport dataset.

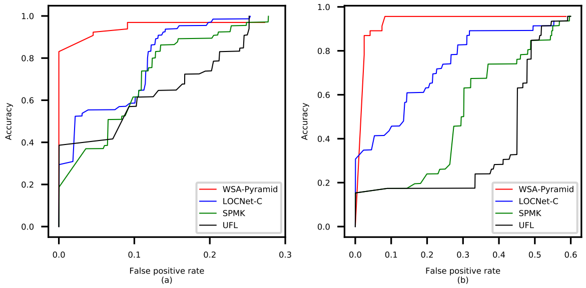

Fig. 4. The AC/FPR curves for the comparison of different methods. (a)Tokyo Airport. (b)Sydney Airport.

formance, two datasets are further adopted. The preliminary results show that the proposed method achieves significantly lower falsely detection ratio and higher accuracy than other methods.

5. ACKNOWLEDGMENT

The work was supported by the National Science Foundation of China (NSFC), under contract No.61866016 and No.61673274, the Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing under No.2016WICSIP031, and the Shanghai Science and Technology Commission Scientific Research Project with project No.17DZ1100803.

6. REFERENCES

[1] Li S et al. He N, Fang L, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. & Remote Sens.*, vol. 56, no. 99, pp. 1–12, 2018.

[2] Shi Z. Liu L, "Airplane detection based on rotation invariant and sparse coding in remote sensing images," *Optik*, vol. 125, no. 18, pp. 5327–5333, 2014.

[3] Guo L et al. Yao X, Han J, "A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and crf," *Neurocomputing*, vol. 164, pp. 162–172, 2015.

[4] Hinton G E et al. Krizhevsky A, Sutskever I, "Imagenet classification with deep convolutional neural networks," *neural information processing systems*, pp. 1097–1105, 2012.

[5] Zisserman A. Simonyan K, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, p. 1409C1556, 2014.

[6] Zhang L et al. Zhang F, Du B, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. & Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, 2016.

[7] Newsam S D. Yang Y, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.* ACM, 2010, pp. 270–279.

[8] Li C et al. Zhou W, Newsam S D, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 197–209, 2018.

[9] Lapedriza A et al. Zhou B, Khosla A, "Learning deep features for discriminative localization," *CVPR*, pp. 2921–2929, 2016.

[10] Ponce J et al. Lazebnik S, Schmid C, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *CVPR*, pp. 2169–2178, 2006.

[11] Cheriyadat A M., "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. & Remote Sens.*, vol. 52, no. 1, pp. 439–451, 2014.