ORIENTATION ROBUST OBJECT DETECTION IN AERIAL IMAGES USING DEEP CONVOLUTIONAL NEURAL NETWORK

Haigang Zhu[1], *Xiaogang Chen*[1], *Weiqun Dai*[1], *Kun Fu*[2], *Qixiang Ye*[1], *Jianbin Jiao*[1*]

[1]School of Electronic, Electrical and Communication Engineering
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Electronics, Chinese Academy of Sciences
[1]jiaojb@ucas.ac.cn

ABSTRACT

Detecting objects in aerial images is challenged by variance of object colors, aspect ratios, cluttered backgrounds, and in particular, undetermined orientations. In this paper, we propose to use Deep Convolutional Neural Network (DCNN) features from combined layers to perform orientation robust aerial object detection. We explore the inherent characteristics of DCNN as well as relate the extracted features to the principle of disentangling feature learning. An image segmentation based approach is used to localize ROIs of various aspect ratios, and ROIs are further classified into positives or negatives using an SVM classifier trained on DCNN features. With experiments on two datasets collected from Google Earth, we demonstrate that the proposed aerial object detection approach is simple but effective.

Index Terms— Aerial Object Detection, Orientation Robust, Deep Convolutional Neural Network.

1. INTRODUCTION

In the past few years, there has been a surge of interest in aerial image object detection. Fast and robust object detection in aerial images is potentially applicable in traffic surveillance, emergency, remote sensing and large scale image content analysis.

A considerable number of approaches have been proposed for aerial object detection [1–5] i.e., vehicle and plane detection, yet the orientation robustness problem remains unsolved. In aerial images, objects in multiple orientations have large appearance variation, which challenges existing feature representation and object detection approaches. In addition, the aspect ratios of objects vary with their orientations, which introduces difficulty to object localization.

In most of the aerial object detection approaches, detectors are trained with orientation-registered samples. Various hand-craft features including Haar-like [6], HOG [7], LBP [8] are exploited to represent aerial objects [5], and classification methods including SVM and Partial Least Squares are used for classification [9]. In the detection procedure, test images are rotated to multiple orientated channels, where objects in the registered orientation are detected. Such orientation-registered approaches are simple and intuitive. However, it often suffers from the high computational cost. In addition, merging results from multiple orientated channels could introduce additional false alarms.

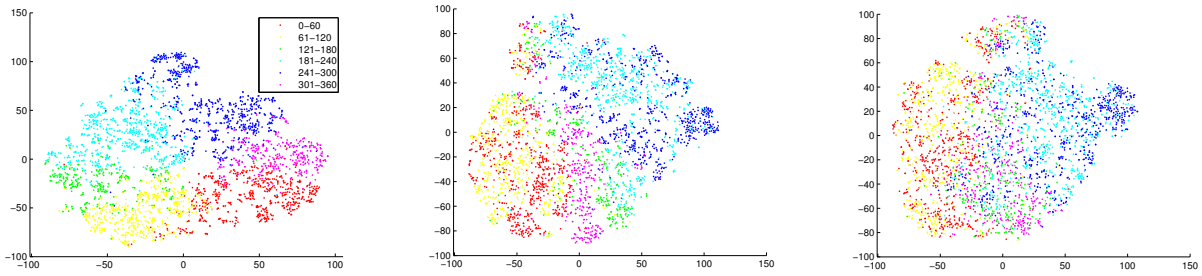
On the other hand, researchers aim to find features that are invariant to specific transformations [10–14]. In [13, 14], deep learning methods including the Convolutional Neural Network and Transformation Invariant Restricted Boltzmann Machine (TIRBM) [12] are studied. However, in these approaches some other challenging factors i.e., the variance of object colors, aspect ratios and cluttered backgrounds are not comprehensively considered.

This work is motivated by a recent leading visual object detection approach rooted in the rich features from deep Convolutional Neural Networks (CNN) and a coarse-localization-fine-classification pipeline [15]. Our main contribution is to explore orientation robust features from combined layers of DCNN. By conducting the t-SNE analysis for feature visualization and analysis, we show how the combined features are related to the recent proposed principle of disentangling feature learning [16]. We argue that rather than extracting rotation invariant characteristics, the combined DCNN features are able to model the rotation factor, which is critical to achieve good performance. To reduce the computational cost as well as process variance of aspect ratios, we employ an image segmentation approach to coarsely localize object candidates, which are then classified with an SVM classifier trained on the orientation invariant features.

The remainder of the paper is organized as follows. In Section 2, the orientation robust feature extraction procedure and the aerial object detection approach are descried. Section 3 presents experimental results. Section 4 concludes the paper with discussion of future works.

2. APPROACH

Our proposed object detection approach comprises two parts: a rotation invariant DCNN feature extraction procedure and



(a) POOL5 (b) FC6 (c) FC7

Fig. 2. t-SNE based visualization of multi-oriented samples. Different colors indicating samples of different orientation in degree (Better view in color).

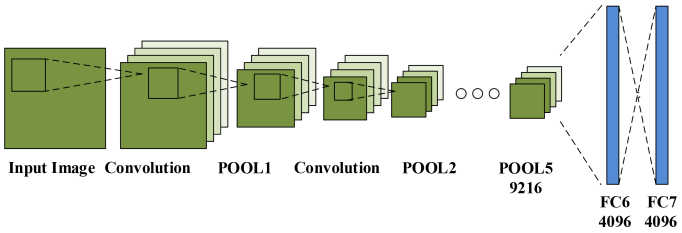


Fig. 1. Structure of the employed Deep Convolutional Neural Network (AlexNet). We use the POOL5, FC6 and FC7 layer features, whose dimensions are 9216, 4096 and 4096 respectively.

an object detection pipeline.

2.1. DCNN Features

For orientation robust DCNN feature extraction, we employ the well-known AlexNet architecture [17]. As shown in Fig.1, the DCNN structure contains five convolutional layers, three of which are followed with pooling layers, and two fully connected layers are stacked at the end, forming a deep CNN architecture. More details about the structure and training protocol please refer to [17].

Recent works have shown that the DCNN trained with a large scale dataset, i.e., ImageNet [18], can generate to other vision tasks [19]. Other works demonstrated that the best performance is achieved by different feature layers in different vision tasks. For example, in the PASCAL VOC multi-class object detection challenge, FC7 performed the best [15]. However, in scene recognition task, FC6 outperformed other layers [20]. In this work, we use DCNN trained with ImageNet to extract multiple layers of features. Three deep layers, i.e., POOL5, FC6 and FC7, are considered for rotation robust feature extraction.

The aim is to find a deep feature representation that is robust to aerial object rotation. To achieve this, we first try to understand how the features distribute with respect to orientations. We propose to use the t-SNE algorithm [21] to visualize the learned features. It can be seen from Fig.2(a) that in the POOL5 feature space, samples form clusters with respect to orientations. In the FC6 and FC7 spaces, however, samples tend to mix together, as shown in Fig.2 (b) and Fig.2(c). Such phenomenon is also observed in the plane samples. With this observation, we can claim that the POOL5 features could well model the rotation factor. FC6 and FC7 layer features seem to model other characteristics rather than orientation variances.

The feature selection procedure is based on the recent advance of disentangling learning [12], which shows that it is proper to use separate groups of features to model distinct factors. With the principle of disentangling learning, features should be able to model the underlying factors of variation, of which rotation is the most obvious one in our case. Since POOL5 shows strong correlation to orientation change, it is expected that employing POOL5 features in the group would achieve better performance.

Beyond the rotation factor, it is expected that the extracted features are robust to other factors, i.e., aspect ratios, color and backgrounds. We further propose to combine features from a different layer for these factors. By concatenating the POOL5 with FC6 or FC7, we experimentally found the best combined feature representation.

2.2. Object Detection

With extracted rotation invariant DCNN features, we conduct a coarse-localization-fine-classification pipeline for objection detection. The flowchart of the pipeline is shown in Fig.3. Details are described in the followings.

2.2.1. Coarse localization

The conventional sliding window detection pipeline, where DCNN features are required to be extracted from millions of image windows, is highly computationally expensive. To reduce the number of windows, a graph-cut based image segmentation approach is firstly used to produce colour consistent regions [19]. On the segmented regions, a similarity measure is used to iteratively group two most similar regions together as a new one. Such an approach, named as Selective

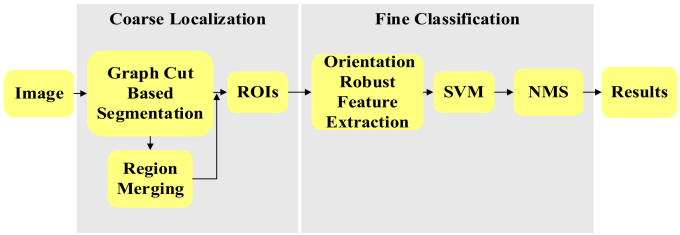


Fig. 3. Flowchart of the proposed object detection pipeline.

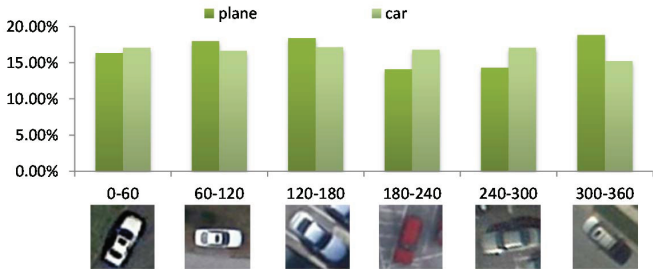


Fig. 4. Distribution of vehicle and plane orientations in degree.

Search, achieved great success, and became the most popular region proposal methods for object detection.

We conduct Selective Search with a simple modification. Considering that the objects are usually in small and uniform scales, we propose stopping the Selective Search merging procedure when the region size exceeds an empirically set threshold: about $100 * 100$ in an $1280 * 659$ images. This simple strategy effectively reduces the region proposals about 60% without performance loss.

2.2.2. Fine Classification

For all the regions of interest (ROIs) generated from the coarse localization stage, we extract features with the method introduced in Section 2.1, and train a linear SVM for classification. The reason we choose linear SVM is that the dimensionality of the DCNN features from combined layers is very high (about 10k), and therefore, could be well learned with a linear SVM classifier. A hard negative mining procedure is conducted to further improve the performance. After classifying all the candidate windows, Non-Max Suppress (NMS) is applied to obtain the final detection results.

3. EXPERIMENTS

We evaluate the proposed approach on a vehicle dataset and a plane dataset collected from Google Earth aerial images. The vehicle dataset contains 310 images with 2819 vehicle samples. The plane dataset contains 600 images, with 3210 plane samples. The samples are carefully selected so that object orientations in the datasets distribute evenly, as shown in Fig.4. Each dataset is split into two subsets: (250 images, 60 images), (500 images, 100 images). One subset is for training, and the other for testing.

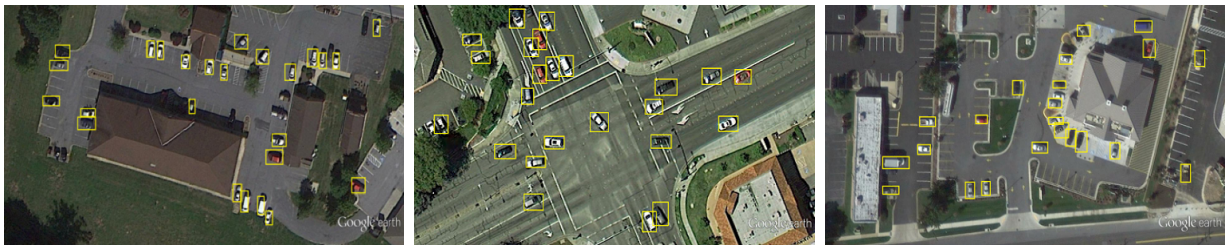
The object detection algorithm returns a list of bounding boxes with SVM classification scores. Evaluation is performed based on this list and the ground truth. According to the PASCAL VOC object detection evaluation protocol [22], a detected bounding box and a ground truth is recognized as matched if their overlap is larger than 50%. The precision-versus-recall curves are presented in Fig.5(a) and Fig.5(b), respectively. Table 1 gives the precision of vehicle detection result when the recall rate is set as 0.8 and the precision of plane detection when the recall rate is set as 0.9.

For comparison, an Aggregate Channel Features (ACF) [23] based sliding window detector is used as baseline. ACF uses HOG, color and gradient as features, and is one of the state of art rigid rigid object detector on many detection tasks. It can be seen that on both the vehicle and plane datasets, the proposed DCNN features significantly outperforms the ACF based detector. For single layer feature, FC7 outperforms POOL5 and FC6 in the vehicle dataset, with precision of 0.861 when recall is set to 0.8. While for the plane dataset POOL5 is the best among the three, with 0.891 precision rate when recall is set to 0.9. On the other hand, combined features significantly improve the performance, especially for features involved with POOL5 layer. In vehicle detection, the POOL5+FC6 combination achieved a 0.945 precision rate with a 0.8 recall rate. In plane detection, the POOL5+FC6, POOL5+FC7 and POOL5+FC6+FC7 combinations show similar results, while POOL5+FC7 is a little better.For convenience, we extract the DCNN features using the RCNN pipeline[15].

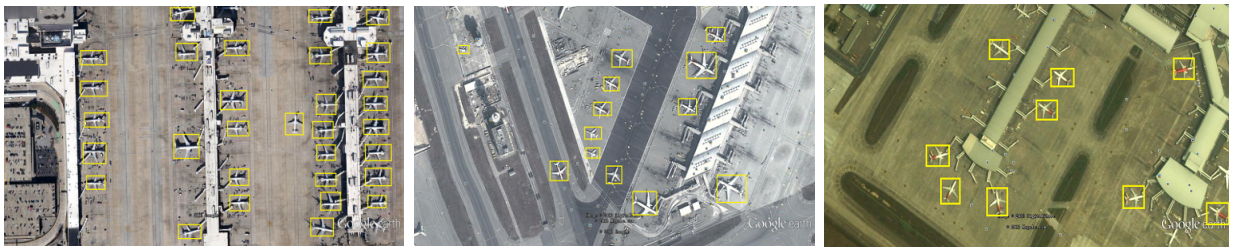
Table 1. Performance of features from different layers.

Feature	Vehicle Recall=0.8	Plane Recall=0.9
ACF(baseline)	0.542	0.511
POOL5	0.548	0.891
FC6	0.635	0.832
FC7	0.861	0.561
FC6+FC7	0.921	0.881
POOL5+FC6	0.945	0.971
POOL5+FC7	0.941	0.972
POOL5+FC6+FC7	0.942	0.971

As indicated in Table 1, combinations involved with POOL5 tends to perform better, which validates the analysis in Section 2, i.e., the POOL5 features can well model the rotation factor. This also explains why the FC6+FC7 combination performs weaker than other combinations in both datasets. This shows that t-SNE provides an effective way to select deep features. It also justifies the hypothesis that disentangling factors of variation helps selecting feature

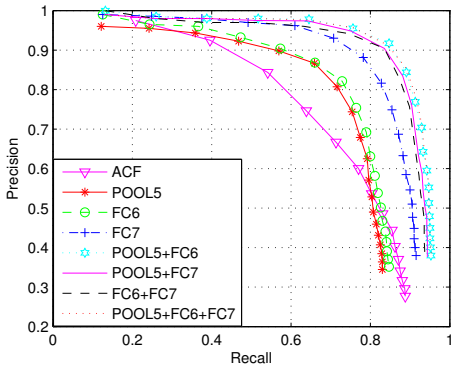


(a)

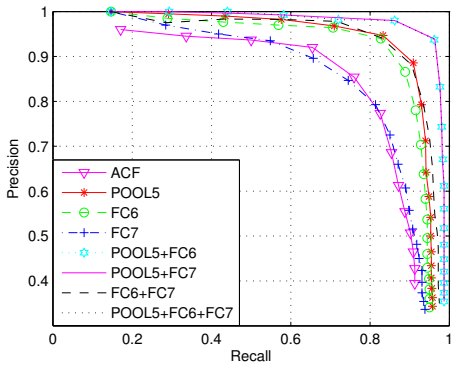


(b)

Fig. 6. Vehicle and plane detection examples. ©2014 Google.



(a) Precision-versus-recall curves of vehicle detection.



(b) Precision-versus-recall curves of plane detection.

Fig. 5. Performance and comparisons of aerial object detection.

representations, although mathematically forming of such a hypothesis remains challenging.

Fig.6 shows some detection examples. It can be seen that our proposed approach detects most of the multi-oriented objects with few false alarms. In the test images, there exists lots of man-made structures, which are correctly classified as negatives. Some vehicles are missed in Fig.6(a), the reason is that segmentation based coarse localization procedure fails on them.

4. CONCLUSION

The experimental results show that the DCNN features from combined layers are competitive when performing orientation robust aerial object detection. With the features, one does not need to perform a rotate-and-detect pipeline, which considerably reduces the computational complexity. We also show that the t-SNE analysis and visualization can be used to find proper DCNN layers. In the future, we plan to detect more kinds of aerial objects with the proposed features and pipeline. To drive the development of aerial image detection research, we will make the aerial object datasets publicly available.

5. ACKNOWLEDGMENT

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB706900, and in part by the National Science Foundation of China under Grant 61271433 and Grant 61202323.

References

[1] Jae-Young Choi and Young-Kyu Yang, "Vehicle detection from aerial images using local shape information,"

in *Advances in Image and Video Technology*, pp. 227–236. Springer, 2009.

[2] Tao Zhao and Ram Nevatia, "Car detection in low resolution aerial images," *Image and Vision Computing*, vol. 21, no. 8, pp. 693–703, 2003.

[3] Line Eikvil, Lars Aurdal, and Hans Koren, "Classification-based vehicle detection in high-resolution satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 1, pp. 65–72, 2009.

[4] Hong Zheng, Li Pan, and Li Li, "A morphological neural network approach for vehicle detection from high resolution satellite imagery," in *Neural Information Processing*. Springer, 2006, pp. 99–106.

[5] Helmut Grabner, Thuy Thi Nguyen, Barbara Gruber, and Horst Bischof, "On-line boosting-based car detection from aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 3, pp. 382–396, 2008.

[6] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*. IEEE Computer Society Conference on, 2001, vol. 1, pp. I–511.

[7] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*. IEEE Computer Society Conference on, 2005, vol. 1, pp. 886–893.

[8] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[9] Aniruddha Kembhavi, David Harwood, and Larry S Davis, "Vehicle detection using partial least squares," *Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1250–1265, 2011.

[10] Henry A Rowley, Shumeet Baluja, and Takeo Kanade, "Rotation invariant neural network-based face detection," in *Computer Vision and Pattern Recognition*. IEEE Computer Society Conference on, 1998, pp. 38–44.

[11] Alireza Khotanzad and Yaw Hua Hong, "Rotation invariant image recognition using features selected via a systematic method," *Pattern Recognition*, vol. 23, no. 10, pp. 1089–1101, 1990.

[12] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee, "Learning to disentangle factors of variation with manifold interaction," *International Conference on Machine Learning*, 2014.

[13] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, and Yann L Cun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in neural information processing systems*, 2010, pp. 1090–1098.

[14] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer vision and pattern recognition*. IEEE computer society conference on, 2006, vol. 2, pp. 1735–1742.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv preprint arXiv:1311.2524*, 2013.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*. IEEE Conference on, 2009, pp. 248–255.

[19] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[21] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.

[22] Mark Everingham, Luc Van Gool, Christopher K-I Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[23] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.