Real-Time Vehicle Detection in Aerial Images using Skip-connected Convolution Network with Region Proposal Networks

No Author Given

No Institute Given

Abstract. Detection of objects in aerial images has gained significant attention in recent years, due to its extensive needs in civilian and military reconnaissance and surveillance applications. With the advent of Unmanned Aerial Vehicles (UAV), the scope of performing such surveillance task has increased. The small size of the objects in aerial images makes it very difficult to detect them. Two-stage Region based Convolutional Neural Network framework for object detection has been proved quite effective. The main problem with these frameworks is the low speed as compared to the one class object detectors due to the computation complexity in generating the region proposals. Region-based methods suffer from poor localization of the objects that leads to a significant number of false positives. This paper aims to provide a solution to the problem faced in real-time vehicle detection in aerial images and videos. The proposed approach used hyper maps generated by skip connected Convolutional network. The hyper feature maps are then passed through region proposal network to generate object like proposals accurately. The issue of detecting objects similar to background is addressed by modifying the loss function of the proposal network. The performance of the proposed network has been evaluated on the publicly available VEDAI dataset.

Keywords: Vehicle Detection · Hyper Maps · Skip connected · Region Proposal Network(RPN) · Aerial Images · Aerial Videos

1 Introduction

There is a growing need of aerial surveillance for civil and military purposes in today's world. This helps in maintaining the decorum of an area while simultaneously ensuring the safety and security of the place. Aerial Surveillance using drones/ Unmanned Aerial Vehicles(UAVs) have been proved to be very useful in this context. Visual Inspection of areas is the most common practice used in surveillance. Use of an UAV to monitor and stream videos to the base station works as a better alternative as compared to the regular visual surveillance through CCTV system. Videos taken from an higher altitude provides a better visibility of the area. It also adds up to the advantage of tracking high speed objects with ease without increasing its own pace.

Real-time monitoring of unwanted and suspicious activities happening in an area is an important concern. This involves Detection and Recognition of objects followed by subsequent tracking of identified suspicious objects. Detection of objects, such as, vehicles, animals, is a challenging task in aerial images due to the small size of objects with respect to the complete frame. This often creates a misclassification scenario for the detector. The misclassification mainly happens due to the fact that there are several small surrounding structures and ambient background objects that look very similar to the objects of interest from the high altitude. Another major issue that makes the task of a detector more difficult is the use of low resolution cameras with UAVs. This is mainly due the limitation of the payload handling capacity of the drone.

Object detection in aerial images is performed by finding the most salient features of the objects and use them for further processing. Many Saliency based approaches have been developed which are effective to find small objects with salient spectral features but did not perform well in real-time aerial videos.

The requirement of a real-time object detection has pushed the research towards using deep learning approaches. It has paced up the performance of object detection and recognition in real-time. Many Convolutional Neural Network (CNN) based architectures have been proposed in recent works that can detect objects with high accuracy. Along with high accuracy, the detectors has also boosted the speed of detection. The convolutional framework of region proposal networks helped in detecting very small objects even when the object is partially occluded or looks similar to the background. The use of high resolution camera payloads with the UAVs has substantially improved the quality of aerial image datasets and consequently has reduced the chances of misclassification.

2 Related Works

Vehicle detection in aerial images has been an active area of research since last three decades. Most of the object detection algorithms uses sliding window approach to generate candidate regions. The candidate regions which are similar to the object properties are considered as the detected objects. But these approaches are very time-consuming and computationally heavy as these methods use several different sized windows and slides over the entire image. Thus these techniques are not well suited for real-time object detection in videos. Region proposals provide computationally less complex solution for object detection. Several region proposal methods have been developed and the performance has improved a lot.

Many saliency based approaches [13] have been developed that produces good results in images where number of objects is not very high and the foreground objects are significantly different from the background. But these methods fail to provide good results in real-time object detection problem.

To overcome the problem of dedicated feature extraction in images containing multiple objects, several deep learning based approaches have been developed in the last decade. Several CNN based frameworks [14] have been proposed that

showed good results for object detection. Zhu et al. have developed a CNN architecture [3] based on the AlexNet framework with Selective Search with a simple modification using empirically set threshold range. A CNN based GoogleNet architecture has been adopted in [1] to detect objects in UCMerced dataset and classify the objects based on a threshold based decision process. A CNN based salient object detection has been proposed in [15], which uses nonlinear regression for refinement of saliency map generated.

The CNN based object detection architectures can be broadly classified into two categories: one-stage and two-stage detectors. The one-stage detectors, namely, YOLO (you only look once), SSD (single shot multi-box detector) etc., provide very fast detection. But these detectors fail to detect small size objects. The two-stage detectors, on the other hand, are very accurate but detect objects at a subsequently low frame rate due to the high computational complexity.

CNN based object proposals classification have been proposed in [8]. which uses region proposal network (RPN) that makes the object detection very accurate. To incorporate location invariance in the model, position-sensitive score maps for the proposals have been introduced in [12]. Due to the accurate object detection, several researchers have adopted the concept of RPN in their work [16], [2] on aerial images in DLR 3K and VEDAI [17] datasets. The accuracy of two-stage object detection has been improved further by adopting the faster RCNN like framework and modifying it by using hierarchical feature maps [?], [6], [4]. A coupled R-CNN based vehicle proposal network has been proposed in [18] that reported good accuracy in DLR-3K dataset.

The above mentioned two-stage classifiers perform well but are noticeably slower as compared to the one-stage classifiers. YOLO [9] and SSD [10] provides very fast detection by using only one CNN architecture for both classification and localization of the objects. But these method fails to detect objects in aerial images as the size of the objects are very small as compared to the size of the proposals. To provide a solution to this problem, focal loss [11] has been introduced by Lin et al. which involves a scaling factor that puts more weightage to the hard classifiable objects as compared to the easily classifiable objects.

The current approaches for vehicle detection are mainly focused on aerial images and thus the speed factor is an issue in these detectors. Even the methods, that performs fast detection of vehicles, are unable to produce very accurate results. This sums up to the unavailability of a suitable vehicle detection framework for real-time aerial videos. In the following section, a two-stage CNN based framework have been proposed that addresses both the speed and accuracy issue.

3 Proposed Framework

The proposed object detection framework is a two-stage model that uses a skip connected convolutional network followed by a region proposal network. The framework generates features from the image using CNN. The features derived are then used to generate multiple object-like regions with scores using a region proposal network.

The proposed framework as shown in Fig. 1 uses first 5 convolutional layers of the ZF-Net architecture [19]. The features from the shallow layer provides low level information about the images. The deeper layers compute more fine details about the image. The features from the shallow layers and the deep layers are merged into a single feature map to define new hyper feature map. This incorporates low level details and deep level highly semantic representation of data together.

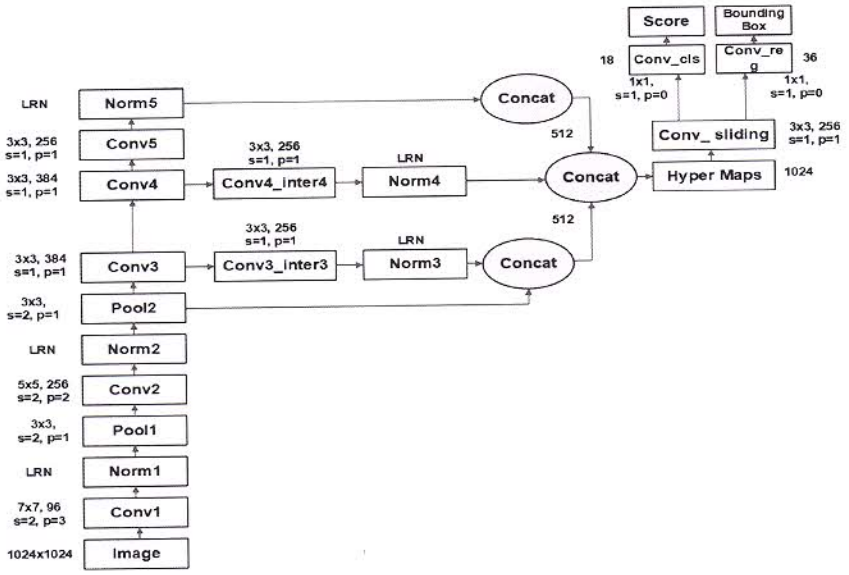

Fig. 1. Proposed Convolutional Hyper Maps based Framework

The output of each convolution layer is passed through Rectified Linear Unit (ReLU) activation function to induce non-linearity in the data. The ReLU output is normalized using Local Response Normalization (LRN).The output of corresponding units of the same convolution layers from different kernels are used to normalize the value of the particular unit in the image.

The shallow and deep layers are first scaled to an intermediate size. The ReLU normalized output of the third convolution layer is passed through an intermediate convolutional layer with 256 kernels. The ReLU output of the *conv3inter*3 layer is normalized using LRN and concatenated with the *Pool*2 output, to define the shallow level features of the images. Similarly, the ReLU output of the fourth convolution layer is passed through an intermediate convolutional layer with 256 kernels to generate *conv4inter*4 output. These output is then passed through the ReLU nonlinear unit and then further normalized. These outputs are merged with the normalized output of *conv*5 to generate deep feature maps.

The deep and the shallow feature maps are merged together to define the Hyper Maps that provide a complete description of the images.

The hyper map contains 1024 feature maps. To define the region proposals, sliding window of size 3×3 is traversed through the entire feature map. The output of the sliding operation is passed through two sibling 1×1 convolution layers to consider the features from all the kernels at each sliding window location. For each of the sliding window location, one network computes the possible locations of the proposals in terms of a vector, $(x_0, y_0, width, height)$, through regression. The other convolution network performs classification of the objects into predefined classes and generates a score for each of the predicted region.

The predicted regions are compared with the ground truth boxes in the images. Intersection over Union (IoU) metric is used to define the similarity of the predicted regions to the ground truth. The total loss at each epoch is computed using a composite loss function that includes the loss of classification as well as regression network. We have employed different loss function for the two components in RPN. We have used Softmax classifier for calculation of loss in score generation of each predicted region of the classification network. Smooth L1 loss has been used to compute loss in the regression layer. The smooth L1 and cross entropy loss functions have been described in Equation 3 and Equation 2 respectively. The overall loss function is defined in Equation 1 where L_{cls} represents classification loss and L_{reg} represents regression loss. The parameter α in Equation 1 represents a trade-off factor between the classification and regression loss.

$$L_{total} = L_{cls} + \alpha L_{reg} \tag{1}$$

$$L_{reg} = \begin{cases} 0.5x^2 & , for|x| < 1 \\ |x| - 0.5 & , otherwise \end{cases}$$

$$L_{cls} = -\sum_i y_i log(p_i) \tag{2}$$

The size of final hyper maps are large enough as compared to the deeper convolutional layers. The size constraint reduces the computational complexity of the algorithm and thus helps in improving the speed of the object detection algorithm. But the increased size of the feature map adds up to the problem of detection of more number of region proposals. Most of the proposals detected are prone to be a part of the background and thus scope of detecting false positives increases.

To handle the number false positive and solve the problem, the concept of Focal Loss has been used as mentioned in [11]. The loss function has been described in Equation 3. The parameter γ decides on the scaling factor to keep or neglect the hard classified examples.

$$FocalLoss(p_i) = -(1 - p_i)^{\gamma} log(p_i) \tag{3}$$

Thus in the Proposed Framework V2, the cross entropy loss function has been replaced by the focal loss function. We have presented the experimental results

for the two frameworks in the following section. The Proposed Framework V1 uses cross entropy loss and Proposed Framework V2 uses focal loss.

4 Experimental Results

The proposed architecture is trained using the publicly available VEDAI dataset. The dataset consists of various backgrounds such as agrarian, rural and urban areas. The dataset is available in two different image sizes 512×512 and 1024×1024 with Annotation. For our experiment, we selected the VEDAI 1024 dataset that contains 1268 images of size 1024×1024 in .png format. We have used 935 images for training the proposed model while the remaining images have been utilized for testing the model. We have trained our model by using pre-trained weights of a ZF-Net model obtained from training the ZF-Net on the ImageNet dataset. The pre-trained ZF-Net model consist of lot of good lower level features which are important in feature extraction for the RPN.

For the proposed framework, the values of the hyperparameters in the LRN are taken as $\alpha = 0.00005$, $\beta = 0.75$ and $\kappa = 0$. The value of the parameter n is taken as 3. The hyperparameters for the focal loss computation is taken as $\gamma = 2$ and $\alpha = 0.25$. This optimal pair of values provide a correct balance between the hard classified and well classified classes. We have further used the trained models to generate inference on aerial videos. The videos were captured at 25 fps and the background in the videos were completely different from the train image scenarios.

The proposed framework has been compared with Faster RCNN architecture. The mean average precision (mAP) and the speed of the Faster RCNN, the Proposed Framework V1 with cross entropy loss and the Proposed Framework V2 with focal loss has been enlisted in Table 1.

Table 1. mAP and Detection Speed of different object detection frameworks

Method	mAP	Speed(fps)
SSD-Inception V2	0.460	7
Faster RCNN	0.454	5
Proposed Framework1	0.644	14
Proposed Framework2	0.659	14

The output of the Proposed Framework V1 on the VEDAI aerial images is shown Fig. 2. The Framework V2 improves both the speed and accuracy of detection. The Proposed Framework V2 can thus be used as a suitable model for vehicle detection in aerial images and aerial videos.

5 Conclusion

In this paper, we have proposed a two-stage hierarchical region-based CNN framework for detection of vehicles in aerial images and videos. The designed

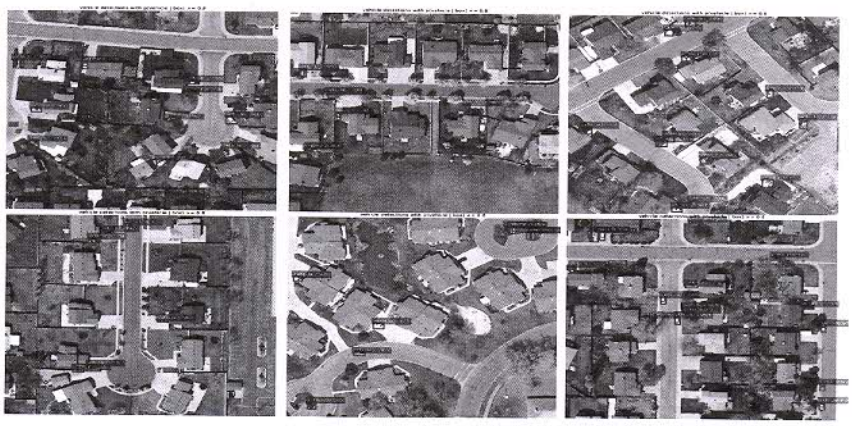

Fig. 2. Object Detection Results on Test Images using Proposed Framework V1

hyper maps based framework produces very accurate and fast vehicle detection result. Scalability of the object size in the videos has also been addressed using the proposed framework due to the use of the shallow and deep layer features. The Proposed Framework V1 produces very accurate results, but the speed of operation is slow as compared to Proposed Framework V2. The inclusion of focal loss in the Proposed Framework V2 helps in reducing the number of proposals per frame by solving the class imbalance problem and thus improves the speed of operation. Thus Proposed Framework V2 generates superior result in terms of accuracy as well as speed and thus can be used as a suitable vehicle detector in aerial videos. However, the proposed framework sometimes do not detect all the vehicles in subsequent frames in aerial videos. The mAP can be further improved to get address the problem and improve the accuracy of the detection algorithm.

References

1. Ševo, I., & Avramović, A. (2016). Convolutional neural network based automatic object detection on aerial images. IEEE geoscience and remote sensing letters, 13(5), 740-744.
2. Sommer, L. W., Schuchert, T., & Beyerer, J. (2017, March). Fast deep vehicle detection in aerial images. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 311-319). IEEE.
3. Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., & Jiao, J. (2015, September). Orientation robust object detection in aerial images using deep convolutional neural network. In 2015 IEEE International Conference on Image Processing (ICIP) (pp. 3735-3739). IEEE.
4. Deng, Z., Sun, H., Zhou, S., Zhao, J., & Zou, H. (2017). Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(8), 3652-3664.

5. Tayara, H., Soo, K. G., & Chong, K. T. (2018). Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network. IEEE Access, 6, 2220-2230.
6. Tang, T., Zhou, S., Deng, Z., Zou, H., & Lei, L. (2017). Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors, 17(2), 336.
7. Chen, Z., Wang, C., Wen, C., Teng, X., Chen, Y., Guan, H., Li, J. (2016). Vehicle detection in high-resolution aerial images via sparse representation and superpixels. IEEE Transactions on Geoscience and Remote Sensing, 54(1), 103-116.
8. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
9. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.
11. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).
12. Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems (pp. 379-387).
13. Borji, A., Cheng, M. M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. IEEE transactions on image processing, 24(12), 5706-5722.
14. De Oliveira, D. C., & Wehrmeister, M. A. (2016, May). Towards real-time people recognition on aerial imagery using convolutional neural networks. In 2016 IEEE 19th International Symposium on Real-Time Distributed Computing (ISORC) (pp. 27-34). IEEE.
15. Li, X., Zhao, L., Wei, L., Yang, M. H., Wu, F., Zhuang, Y., & Wang, J. (2016). Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing, 25(8), 3919-3930.
16. Lee, J., Wang, J., Crandall, D., Šabanović, S., & Fox, G. (2017, April). Real-time, cloud-based object detection for unmanned aerial vehicles. In 2017 First IEEE International Conference on Robotic Computing (IRC) (pp. 36-43). IEEE.
17. Razakarivony, S., & Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation, 34, 187-203.
18. Tang, T., Zhou, S., Deng, Z., Lei, L., & Zou, H. (2017, July). Fast multidirectional vehicle detection on aerial images using region based convolutional neural networks. In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1844-1847). IEEE.
19. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). springer, Cham.