RESEARCH ARTICLE

Hangar Detection from Satellite Images with Mask-RCNN and YOLOv2 Algorithms

Nida Kumbasar^a, Asli Nur Ömeroğlu^a, Emin Argun Oral^a, İbrahim Yücel Özbek^a
a Department of Electrical and Electronics Engineering, Ataturk University, Erzurum, Turkey

Abstract

The contribution of this paper is twofold. It first proposes a new dataset of high resolution satellite images of hangars located at civil and military airports. It also presents a hangar detection problem results from satellite images using this new dataset obtained by Mask R-CNN and YOLOv2 algorithms. The satellite dataset contains one thousand pictures obtained from Google Earth at 8, 11, 17, 29 degrees angles from height of 500,800 and 1000 meters. Among 1000 hangar images 650 and 200 images are used for training and validation, respectively, while the remaining 150 images are used for detection test purposes. The total number of hangar object instants in the dataset images is about 3000. The detection of hangars is a challenging problem as the dataset contains camouflaged and non-camouflaged targets in different sizes. Among the two approaches used in the detection problem Mask R-CNN utilizes a regional based algorithm and enables instance segmentation with a bounding box. YOLOv2, on the other hand, is a regression based algorithm, used in real-time applications, and provides a bounding box only. The object detection accuracy in terms of Average Precisions by using Mask R-CNN and YOLOv2 algorithms to detect different sized camouflaged and non-camouflaged hangar objects was obtained as 72% and 74%, respectively.

Keywords: YOLOv2, Mask R-CNN, Satellite Images, Camouflaged and Non-Camouflaged Hangar, Object Detection

1.Introduction

Object detection and classification from remotely sensed images are among the most remarkable subjects in the study of satellite photographs and are applied in many areas such as natural disaster detection, vegetation mapping and urban planning. In recent years, with the developments in satellite technology, very high resolution images can be obtained from any part of the world at any time. The remote sensing technology, which allows the world to be observed with optical or microwave sensors at different heights and angles, enables the user to obtain detailed shape property and structure information. The studies on the detection of objects have become more important especially in the remote sensing images used in military and civilian applications.

Object detection in satellite images is often performed as a remote sensing application. It accomplishes two tasks as defining the class that the object belongs to and surrounding the object with a bounding box. Nowadays, with the increasing interest in deep machine learning and the rapid development of technologies, the Convolutional Neural Network (CNN) has brought the object classification accuracy to a level that can compete with human performance. The CNN-based object detection models consist of two main approaches. Namely, a regional based model and a regression model. In the regional based approaches, candidate region proposals are first created using

Selective Search method. Then the classification takes places based on these region proposals using a pre-trained CNN. In the regression model, the object category and location information are both estimated by using CNN.

Figure 1. Sample images from dataset

Figure 2. Mask R-CNN flowchart

The first object detection model based on the regional proposal is R-CNN [1], proposed by Ross Girshick. In its first phase, approximately 2000 region proposals are created by a search method, namely Selective Search [2], the properties are then extracted with the CNN model in the respective regions. Finally, algorithm classifies the regions using a Support Vector Machine (SVM) classifier.

R-CNN's repeated use of the CNN structure for extracting features increases the complexity of the calculations as well as the processing time. In order to speed up the R-CNN, Ross Girshick proposed the Fast R-CNN model [3] by which the processing time is reduced since a single CNN structure is used for all regions instead of using separate CNNs. Ren Shaoqing suggested the Regional Proposal Network (RPN) in Faster R-CNN [4] to compensate the workload of Selective Search and its blind search effort. In addition to the techniques mentioned above, Mask R-CNN [5] has been recently proposed by Kaiming He et al. It is basically the development of Faster R-CNN. In Mask R-CNN, an example of the semantic segmentation technique, the category and bounding box information of the object as well as its silhouette information are obtained as a part of the output.

Although regional based algorithms, mentioned above, give satisfactory accuracy results, they fail in real time object detection problems due to their computational complexity. As a solution, Joseph Redmon proposed You Look Only Once (YOLO) [6] object detection model. In this method, the object detection problem was investigated as a single regression problem. It is aimed to estimate the coordinates, category and probability of the bounding box with less processing load and computational complexity than its ancestors. YOLO, on the other hand, does not exhibit the expected performance in small object detection problems compared to region proposal based object detection models. Wei Liu has proposed a high speed and accuracy performance Single Shot multibox Detector (SSD) [7] by combining the anchor mechanism of faster R-CNN with the regression model used in YOLO. Then Joseph Redmon developed the YOLO model and offered YOLOv2 [8]. In this proposed model, the displacement of the fully connected layer with the pass-through layer increases the input image size flexibility and combine the feature maps.

In this study, a new dataset, contains camouflaged and non-camouflaged hangars, were created first, and detection of hangars is performed using Mask-RCNN and YOLOv2. This new dataset includes high resolution satellite images of hangars in different sizes located at airports around world obtained from Google Earth. Fig. 1 shows sample images from the dataset.

Section details of this study as follows. Section II describes the training algorithm and methods in detail. Database preparation and the obtained results are given in Section III. Finally, Section IV concludes the study with future works.

2. Training Algorithm and Methods

2.1. Mask R-CNN

Mask R-CNN [5] is an example of object segmentation algorithm developed by the Facebook Artificial Intelligence Survey (FAIR) team. The goal here is to create a separate mask on each of images while detecting specific objects in them. Mask R-CNN not only defines the position of the class and bounding box but also identifies the pixels corresponding to a particular class in that box. This pixel information is utilized, for example, for path planning in autonomous vehicles and to detect objects in robotic grasping problems.

2.1.1 Region Proposal Network

Region based object detection algorithms are composed of two main parts. These are creating of candidate areas that can contain objects in the image to be processed and classifying the target with a bounding box. Region Proposal Network [4] (RPN) creates regional proposals using a basic CNN structures. The input image is passed through convolution layers to obtain a feature map, and this feature map is used as the input of RPN. RPN identifies, detects, different size identification frames in floating positions by using a sliding window, defined by anchors for different size and aspect ratios. If these anchors overlap too much, the one with the highest Intersection over Union (IoU) value is selected. This process, called non-max suppression, provides the intersection regions, namely Region of Interest (RoI).

2.1.2 Region of Interest (RoI)

Two different outputs are generated for regions proposed by RPN. These are class and bounding box info. Unlike the binary classification process in RPN, class info of each region in RoI is determined among multiple classes such as hangar, tree, building, etc. The position and size of the bounding box are further enhanced to border the objects.

As input images with different sizes are difficult to be processed by classifiers, it is generally preferred for fixed input size images. However, thanks to the bounding box extraction step in the RPN, the RoI boxes may be of different sizes. This is where the RoI pooling process, used in region-based algorithms such as Faster R-CNN, takes place. The RoI pooling trims a portion of the feature map and resizes it to a fixed size. Also a method, called RoI Align, samples a property map at different points and applies a bilinear interpolation, proposed by Mask R-CNN. The output of the RoI classifier and the bounding box regressor are shown in Fig. 2.

2.1.3 Segmentation Mask and Fully Connected Layer

This step consists of two parallel processing approaches. One of them is the Full Connection Network (FCN) to create a mask output by applying basic semantic classification, and the other is a process step (FC Layers) to produce bounding box information by classification. The masking phase is a convolutional network that takes the positive regions selected by the RoI classifier and creates a mask for them. The created masks have low resolution with a size of 28x28 pixels. The small mask size ensures that the masking phase has a low computational complexity. To calculate the lost function value during training, the basic accuracy masks are scaled to the size of the RoI bounding box so that the final masks are obtained as shown in the final part of Fig. 2.

In the study, open source libraries, based on Keras and Tensorflow, were used to train the Mask R-CNN model [9]. ResNet50 or ResNet101 is a standard convolutional neural network that both act as feature extractor. First layers of these networks detect low-level properties while the subsequent layers detect higher-level properties. Feature Pyramid Network (FPN) [10] is a method that can better represent objects at multiple scales. It provides access to both lower and upper level properties of each step by improving the standard feature extraction pyramid. Hence, our implementation of Mask R-CNN uses a ResNet101 + FPN backbone. For training, a pre-trained model was used with the COCO [11] dataset.

2.2. YOLOv2

In 2016, Redmon et al. [6] proposed an object detection method, YOLO, which divides the image into $S \times S$ grids and estimates the bounding box and class probability for each grid cell. Each bounding box consists of five estimates, namely w , h , x , y and confidence score. The w and h values here represent the width and height of the box relative to the whole image while (x,y) value pair represent the center coordinate of the box relative to the boundaries of the grid cell. The confidence score, on the other hand, indicates how accurate the boundary box is, and how likely the box contains an object, namely objectness score. All the boxes that are below a given threshold value are removed. The non-maximum suppression (NMS) [12] method is applied to eliminate extra bounding boxes that partially bounds the object. YOLO flowchart is shown in Fig. 3.

YOLOv2 was proposed by Redmon et al. in 2017 to improve the estimation accuracy of YOLO. In YOLOv2 network the target classification as well as identification of target boxes and feature extraction are all integrated into a single network, and an end-to-end training is provided. YOLOv2 utilizes a new network structure, Darknet-19, in which batch normalization was applied to each layer by removing the full connection layers of the network. K-means clustering was used to obtain the anchor boxes. Compared with YOLO, YOLOv2 greatly improves the accuracy and speed of object detection.

Figure 3. YOLO flowchart

In the current study YOLOv2 method, divided into 13x13 grids based on DarkNet-19 backbone, is used. For training, a pre-trained model was used with the COCO [11] dataset.

3. Experimental Work

3.1. Dataset

The dataset used in this study was obtained with the Google Earth Pro application. Total of 1000 high resolution images were collected at 8, 11, 17, 29 degrees angles from height of 500,800 and 1000 meters. The

total number of object instants is about 3000. The dataset is randomly split into three as training, validation and test with 650, 200 and 150 images, respectively.

Two different image labeling tools were used to label the data as required by two different algorithms. The first one, as shown in Fig. 4, is the LabelImg [13] tool that enables labels in xml format required for YOLOv2. The other tool, as shown in Fig. 5, is VGG Image Annotator [14] to create labels in json format required for Mask R-CNN.

Figure 4. LabelImg tool

Figure 5. VGG image annotator GUI

3.2. Performance Criteria

In the literature Average Precision (AP) value is widely used as an evaluation criterion to measure the accuracy of object detection algorithms. Hence, it is also used in this work to measure the performance of the detection method. AP is the average of the precision values corresponding to the different recall values. The precision, calculated using Eq.1, is the ratio of the correct estimates to all results marked as correct. The recall on the other hand, ratio of correct estimates to the correct values, is given in Eq. 2. In both equations TP, TN, FP and FN stands for True Positive, True Negative, False Positive and False Negative, respectively. AP values in the Mask R-CNN algorithm are found assuming the IoU value was greater than 0.7. Moreover, in the YOLOv2 algorithm the confidence score was calculated for objects larger than 0.27.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \quad (2)$$

3.3. Results

Fig. 6 shows some test results of the Mask R-CNN and YOLOv2 algorithms. While YOLOv2 detected all hangars for the image shown in Fig. 6 (a), no targets were found in Mask R-CNN. In Fig. 6 (b) the Mask R-CNN and YOLOv2 both correctly detected the existing target. In Fig. 6 (c) both Mask R-CNN and YOLOv2 detected all target objects correctly while YOLOv2 resulted in incorrect target detection, that is a False Positive result. In Fig. 6 (d) on the other hand, both algorithms made

the same error detecting only one target among all. Based on Fig. 6 results, it is observed that YOLOv2 algorithm detects camouflaged hangars more successfully than Mask R-CNN.

The object detection performance of both algorithms was measured in terms of precision-recall curve based on the number of false and truly placed bounding boxes. Fig. 7 shows that precision-recall curve for both algorithms. The Average Precisions (AP) measure obtained as 72% for Mask R-CNN algorithm and 74% for YOLOv2 algorithm.

Figure 6. Experimental results

Figure 7. Hangar detection Precision-Recall curve

4. Conclusion and Future Work

In this study, a new satellite image dataset is introduced for object detection applications. The dataset consists of 1000 images with camouflaged and non-camouflaged hangars located at civil and military airports. Two detection algorithms, namely YOLOv2 and Mask R-CNN, were applied for the dataset. Among these, the YOLOv2 algorithm is frequently used in real-time applications while the Mask R-CNN is the newest of the regional based algorithms in object detection applications. Based on current study results, the Average Precision bounding box results of both algorithms are measured as 72% and 74% for the Mask R-CNN and YOLOv2 algorithms, respectively. It is also observed that the YOLOv2 algorithm detects camouflaged hangars better than Mask R-CNN algorithm.

As our future studies, we will expand the dataset and make it publicly available for researchers.

References

- [1] Ross Girshick, Donahue Jeff, Darrell Trevor, Malik Jitendra, "Rich feature hierarchies for accurate object detection and semantic segmentation", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014.
- [2] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, Arnold WM Smeulders, "Selective search for object recognition", *International journal of computer vision* 104, vol. 2, pp. 154-171, 2013.
- [3] Ross B. Girshick, Fast R-CNN, pp. 1440-1448, 2015.
- [4] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, pp. 91-99, 2015
- [5] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [6] Joseph Redmon, Divvala Santosh, Girshick Ross, Farhadi Ali, "You only look once: Unified real-time object detection", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788, 2016.
- [7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "Ssd: Single shot multibox detector", *European conference on computer vision*, pp. 21-37, 2016.
- [8] J. Redmon, A. Farhadi, "YOLO9000: Better Faster Stronger", 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517-6525, 2017.
- [9] W. Abdulla, Mask r-cnn for object detection and instance segmentation on keras and tensorflow, 2017, [online], Available: https://github.com/matterport/Mask_RCNN.
- [10] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," in *CVPR* 2017.
- [11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, "Microsoft coco: Common objects in context", *European conference on computer vision*, pp. 740-755, 2014.
- [12] Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Hong Kong, China, 20-24 August 2006; pp. 850-855
- [13] <http://tzutalin.github.io/labelImg/>
- [14] A. Dutta, A. Gupta, A. Zisserman, Vgg image annotator (via), 2016, [online], Available: <http://www.robots.ox.ac.uk/~vgg/software/via>.