INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Computer Vision Application in Object Detection and Tracking for Aerial Surveillance

Archana Nandibewoor[1]*, L K Prateek[2], Manish Sakaray[2], Abul Hassan[2], Akshay Ravankar[2], Abhilash Hegde[3]

1 Assistant Professor, Research and Development Center, SDMCET, Affiliated to VTU, Dharwad, Karnataka, India
2 Student, Computer Science and Engineering, SDMCET, Dharwad, India
3 Junior Research Fellow, Research Scholar (Ph.D.), SDMCET, Dharwad, India

OPEN ACCESS

Received: 20-03-2023
Accepted: 09-07-2023
Published: 14-08-2023

Citation: Nandibewoor A, Prateek LK, Sakaray M, Hassan A, Ravankar A, Hegde A (2023) Computer Vision Application in Object Detection and Tracking for Aerial Surveillance. Indian Journal of Science and Technology 16(31): 2374-2379. https://doi.org/10.17485/IJST/v16i31.645

* **Corresponding author**.

narchana2006@gmail.com

Funding: ARDB/01/1081990/M/I

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Computer vision duties like object detection, tracking, and counting are significant for surveillance. Factors like altitude, camera angle, occlusion, and motion blur make it a more challenging task. To present a method to overcome all these factors and implement surveillance quickly and accurately for smaller and larger object aspect ratios. **Methods:** Horizontal Bounding Boxes and Oriented Bounding Boxes (HBB and OBB) are evaluated on two ground truths respectively. PASCAL VOC 07 metric is adopted to calculate the mean average precision. Constructed on the score, the original implementation of Mask R-CNN includes the application of a mask head to the highest-scoring 100 HBBs. Subsequently, the mask head was extended to all HBBs remaining after the process of Non-Maximum Suppression. This modification allowed the evaluation of Mask R-CNN, Cascade Mask R-CNN, and Hybrid Task Cascade methods on a wider range of bounding boxes. **Findings:** In summary, this research explores and compares different approaches and techniques in the field of object detection, particularly focusing on oriented object detection and the challenges posed by geometric variations. Furthermore, it addresses the impact of different models, such as Mask R-CNN, Faster R-CNN OBB + RoI Transformer, and Faster R-CNN OBB + Dpool, on performance. Additionally, it highlights the importance of handling numerical instability caused by extremely small instances. The research findings are visually presented in Figure 2, providing a clear representation of the performance of various networks. **Novelty:** The study summarizes the findings of existing research papers and identifies research gaps. The performance parameters of the various algorithms and analysis for various networks show the evolution of various methods over the years. With changes in the network, like mask transferring and dataset, the accuracy for smaller, bigger objects and speed of execution are affected, are explained in results and discussions as well as the conclusions.

Keywords: RCNN; Deep learning; Object detection; Computer Vision; Drones

1 Introduction

Aerial surveillance plays a critical role in various domains such as security, urban planning, environmental monitoring, and disaster management. It involves the use of unmanned aerial vehicles (UAVs) or satellites to capture high-resolution images or videos of large areas from above. These aerial data provide valuable insights and enable the analysis of complex environments with a wide range of applications. With the advancements in machine learning, particularly in computer vision and image processing, aerial surveillance has witnessed significant improvements in its effectiveness and efficiency. Machine learning algorithms have the capability to automatically analyze and interpret aerial images, enabling the extraction of valuable information from vast amounts of visual data. By leveraging machine learning techniques, aerial surveillance can benefit from automated object detection, classification, tracking, and anomaly detection. These capabilities enhance situational awareness, enable real-time monitoring, and support decision-making processes. Machine learning algorithms can learn patterns, detect changes, and identify objects of interest, such as vehicles, buildings, or natural features, in aerial images or videos.

One of the key advantages of using machine learning in aerial surveillance is its ability to handle the vast amount of data generated by UAVs or satellite sensors. These algorithms can efficiently process and analyze large-scale aerial imagery datasets, enabling the detection of subtle changes or abnormalities that might go unnoticed by human observers. Moreover, machine learning algorithms can adapt and learn from new data, improving their performance over time. This adaptability is crucial in aerial surveillance, where environmental conditions, lighting, and objects of interest can vary significantly. By continuously learning and updating their models, machine learning algorithms can maintain high accuracy and reliability in aerial surveillance tasks.

In this paper, we explore the application of machine learning techniques in aerial surveillance and their potential to revolutionize the way we monitor and analyze large-scale environments. We discuss various approaches, algorithms, and methodologies used in object detection, tracking, and anomaly detection tasks. Additionally, we highlight the challenges and future directions in the field, aiming to inspire further research and innovation in the intersection of aerial surveillance and machine learning. By harnessing the power of machine learning, aerial surveillance holds tremendous potential to improve situational awareness, enhance security measures, and support decision-making processes in various domains. The following sections delve into the specific applications, methodologies, and advancements in this exciting field.

Remote sensing is the method of acquiring and recording information about an object without direct contact[(1)]. Drones and UAVs (Unmanned Aerial Vehicles) have become widely used in academics and real-world applications. As a result, one must comprehend and analyze the image data that they collect. Object detectors based on DNNs (deep neural networks) considerably improve object detection performance in the deep learning age. Nevertheless, there are several prominent discrepancies between conventional nature photographs and images obtained by drones, making item detection a difficult process. For starters, the items in such photos are scaled differently. The far items are extremely small in comparison to the close objects. Furthermore, cities have a plethora of dense settings. Because of the density, there are a lot of occlusions, making object detection even more challenging. Object detectors based on deep learning (DL) are now mapped into dual groups.

Two-stage detectors are the first to assess whether the preceding anchors are an object or background, they employ a region proposal network. Several manually defined potential bounding boxes serve as prior anchors. The prospective anchors are then classified into a set of categories using two head networks, and the offset between

the anchors and ground truth boxes is estimated. One-stage detectors are the other kind. Unlike the two-stage detectors, the one-stage detectors do not take the region proposal network. To anticipate the categories and the offset of the prior anchors, they use two detectors directly. The low-resolution image grid is used to produce the previous anchors for these two types of detectors. According to the intersection over union (IoU), each previous anchor can only have a single object bounding box. However, using the image recorded by the drone, the fixed-shape anchor is unable to manage objects of varied sizes. Another sort of detector, the anchor-free detector, has just been proposed. They condense bounding box prediction to a single key point and size calculation. It provides a more accurate method of detecting objects of varying sizes. Despite this, regression is difficult because of the high size difference.

In this study, the RRNet, a hybrid detector is presented. Regardless of the object's scale, the object's center point is always present. As a result, instead of using the anchor box, two detectors were employed to anticipate the center point, as well as the width (w) and the height (h) of each object. The center points and diameters are then converted to coarse bounding boxes. Finally, used a Re-Regression module to feed the deep feature maps and coarse bounding boxes. The Re-Regression module allows you to fine-tune the coarse bounding boxes before generating the final correct bounding boxes. Moreover, the review advises that appropriate data augmentation can help deep models attain state-of-the-art performance without modifying the network design. This strategy allows you to logically extend the objects in the image. Experiments illustrate that the proposed model far exceeds the existing state-of-the-art detectors in the Via Drone 2018 dataset. In principle, the RRNet (Re-Regression Network) is a hybrid model of an anchorless detector and a two-stage detector. It is believed that a regression engine is very important for good results. In addition, it achieves the best outcomes for an average precision and recalls AP50, AR10, and AR100 respectively.

There are various applications of aerial imaging, some of which are as follows. Kolbe et.al. in their paper, special focus on the utilization of model concepts concerning different tasks in disaster management[(2)]. After extensive state-of-the-art, methods belonging to each type of automatic building construction approach. Based on a concrete experiment, the essential points characterizing each approach, including their concept and the obtained model characteristics are analyzed. It is of great interest to study these issues since not many investigations are being carried out before and have delivered useful formulas[(3)]. One of the emerging technologies used to study the rate of vegetation is hyperspectral remote sensing[(4)].

2 Methodology

A detailed literature review was conducted focusing on object detection and tracking using drones concerning different applications. The study summarizes the findings of existing research papers and identifies research gaps. The performance of the various algorithms and analysis for various networks shows the evolution of various methods over the years. With little changes in the network, such as mask transferring and dataset, the accuracy for smaller, bigger objects and speed of execution are affected, which are explained in results and discussions, and also the conclusions.

The task of object detection involves the identification and classification of objects within images[(5)]. Using two location representations horizontal bounding boxes and oriented bounding boxes (HBB and OBB) in this paper. The HBB is a rectangle (x, y, w, h), and the OBB is a quadrilateral {(xi, yi)|i = 1, 2, 3, 4}. Subsequently, two distinct tasks emerge, namely the detection using Horizontal Bounding Boxes (HBB) and the detection using Oriented Bounding Boxes (OBB).

That said, these methods are evaluated on two kinds of ground truths: HBB and OBB ground truths[(6)]. Each detected bounding box has a corresponding confidence score for the two tasks. Then adopted the PASCAL VOC 07 metric for the calculation of the mean average precision (mAP). Average Precision (AP) computes the average precision value for recall value over 0 to 1 (i.e., the area under the precision/recall curve). mAP is the average of AP over all classes[(7)]. The detailed computation of the precision and recall can refer to. The intersection over union (IoU) is crucial in determining true positives and false positives, which are required to compute precision and recall[(8)]. It is important to highlight that, in the case of the Oriented Bounding Box (OBB) task, the intersection over union (IoU) is computed between OBBs, as illustrated in Figure 1. The two OBBs box precision and box groundtruth (Bp and Bgt) respectively, and the intersection between OBBs (Bp T Bgt) are all convex polygons, whose area can be easily computed. The union area of two OBBs can be calculated as |Bp S Bgt| = |Bp| + |Bgt| − |Bp T Bgt|. The code for the mAP and IoU computation between OBBs are found in the development kit.

To ensure fair comparisons among different algorithms in Dataset for Object Detection in Aerial Images (DOTA), a unified code library based on MMDetection was implemented for the implementation and evaluation of all the algorithms. Due to memory limitations, large images were cropped into 1024×1024 patches with a stride of 824 during both training and inference. Temporary results were obtained from the patches, which were then mapped back to the original image coordinates and subjected to Non-Maximum Suppression (NMS) with different threshold settings.

The training process was carried out using 4 GPUs, with a batch size of 81 (2 images per GPU). A learning rate of 0.01 was utilized, and the number of proposals and maximum predictions per image patch were set to 2,000. Baseline models

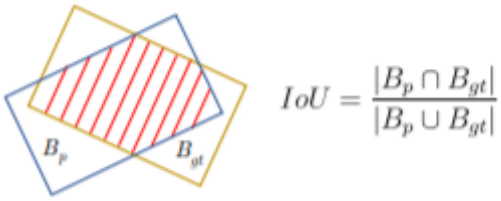

Fig 1. Intersection of Union (IoU) calculation

were developed for both the Horizontal Bounding Box (HBB) and Oriented Bounding Box (OBB) tasks, employing different techniques including RetinaNet, Mask R-CNN, and Faster R-CNN.

To enable the prediction of OBBs using existing object detection methods, two approaches were employed. The first approach involved modifying the HBB head to predict OBBs by regressing the offsets relative to HBBs. The second approach used the Mask Head to treat OBBs as coarse masks and performed pixel-level classification.

For OBB prediction, the Faster RCNN OBB, modified RoI Head of Faster R-CNN, and Anchor Head of the single-shot detector (SSD) were used to regress quadrangles. A special representation (x, y, w, h, θ) was used for OBB regression, where (x, y) represents the center, w and h represent the width and height, and θ represents the orientation angle.

To calculate the learning target, IoUs were computed between horizontal RoIs (anchors) and HBBs of the ground truths. The best-matched ground-truth form was selected based on a distance function, and the corresponding learning target was determined in equation (1).

$$\begin{gathered}\{tx = [[(xb - x)]/w]\}, \{ty = [[(yb - y)]/h]\}, \\ \{tw = [\log(wb/w)]\}, \{th = [\log(hb/h)]\}, \\ \{t\theta = [\theta b - \theta]\}\end{gathered} \tag{1}$$

To address the detection of oriented objects in object detection methods and establish baselines for both the Horizontal Bounding Box (HBB) and Oriented Bounding Box (OBB) tasks in DOTA, several modifications and approaches were implemented.

To incorporate OBB predictions into the Faster RCNN and RetinaNet models, the HBB RoI Head and anchor head were replaced with the OBB Head, resulting in two models: Faster R-CNN OBB and RetinaNet OBB. Furthermore, an adaptation was made to Faster R-CNN to enable parallel prediction of both HBB and OBB, following a similar approach to Mask R-CNN. This variant was referred to as Faster R-CNN H-OBB.

The performance of deformable RoI pooling (Dpool) and RoI Transformer was evaluated by replacing the RoI Align module in Faster R-CNN OBB. This led to the creation of two additional models: Faster R-CNN OBB + Dpool and Faster R-CNN OBB + RoI Transformer. It is important to note that the RoI Transformer used in this study differs slightly from the original version, as it was based on the Light Head RCNN instead of Faster R-CNN.

For OBB predictions in DOTA, the Mask R-CNN model, initially designed for instance segmentation, was adopted. Although DOTA lacks pixel-level annotations, the OBB annotations were treated as coarse pixel-level annotations. During inference, the minimum OBBs encompassing the predicted masks were calculated. While the original Mask R-CNN applies the mask head to the top 100 HBBs based on their scores, in DOTA with its high number of instances per image, the mask head was applied to all HBBs after Non-Maximum Suppression (NMS). These modifications and approaches provided valuable insights into handling oriented objects in object detection and establishing effective baselines for the HBB and OBB tasks in DOTA.

3 Results and Discussion

Object detection still faces challenges in handling geometric variations. In this study, the challenges of handling geometric variations in object detection were addressed. To evaluate different approaches, RoI Transformer and Dpool were utilized as replacements for RoI Align in Faster R-CNN OBB. The resulting models were named Faster R-CNN OBB + RoI Transformer and Faster R-CNN OBB + Dpool. Comparing their performance, it was observed that Dpool generally improves the performance of Faster R-CNN OBB, but RoI Transformer surpasses Dpool, particularly in aerial images. This finding highlights the effectiveness of carefully designed geometry transformation modules like RoI Transformer for aerial images compared to general geometry transformation modules like Dpool.

To tackle the numerical instability caused by extremely small instances during training on DOTA-v1.5 and DOTA-v2.0, a threshold was set to exclude such instances. The impact of different thresholds on DOTA-v2.0 was explored, revealing that small instances have minimal influence on the results.

The number of proposals is an important hyperparameter in modern detectors and exhibits significant variations between aerial and natural images. Optimal settings for aerial images were investigated, and it was observed that for Faster R-CNN OBB, the improvement in mAP slows down at around 8,000 proposals. Furthermore, within the range of 1,000 to 10,000 proposals, Faster R-CNN + RoI Transformer and Faster R-CNN OBB demonstrated mAP improvements of 2.2 and 1.39 points, respectively. However, it should be noted that increasing the number of proposals also increases computational demands. Therefore, 2,000 proposals were chosen for the other experiments conducted in this paper.

The performance of various models, including Faster R-CNN, Faster R-CNN OBB, RetinaNet OBB, Mask R-CNN, and Faster R-CNN OBB + RoI Transformer, was evaluated on challenging scenes, as depicted in Figure 2. The results demonstrate that Faster R-CNN OBB, Mask R-CNN, and Faster R-CNN OBB + RoI Transformer perform well in densely packed large vehicles. However, RetinaNet OBB exhibits lower location precision due to feature misalignment. Detecting long-shaped instances with large aspect ratios poses challenges as all methods tend to generate multiple predictions on single bridges or ships. Additionally, there is a similarity issue where different categories are prone to misclassification due to similar features. Lastly, the detection of extremely small instances (less than or approximately 10 pixels) presents difficulties, resulting in low recall rates.

In Figure 2 the results of the DOTA-v2.0 test-dev are visually represented, showcasing the performance of the five models specifically designed for DOTA-v2.0. Only predictions with scores above 0.1 are displayed. The visualization demonstrates the models' capabilities in handling orientation variations, density variations, instances with large aspect ratios (Ars), and instances with small Ars.

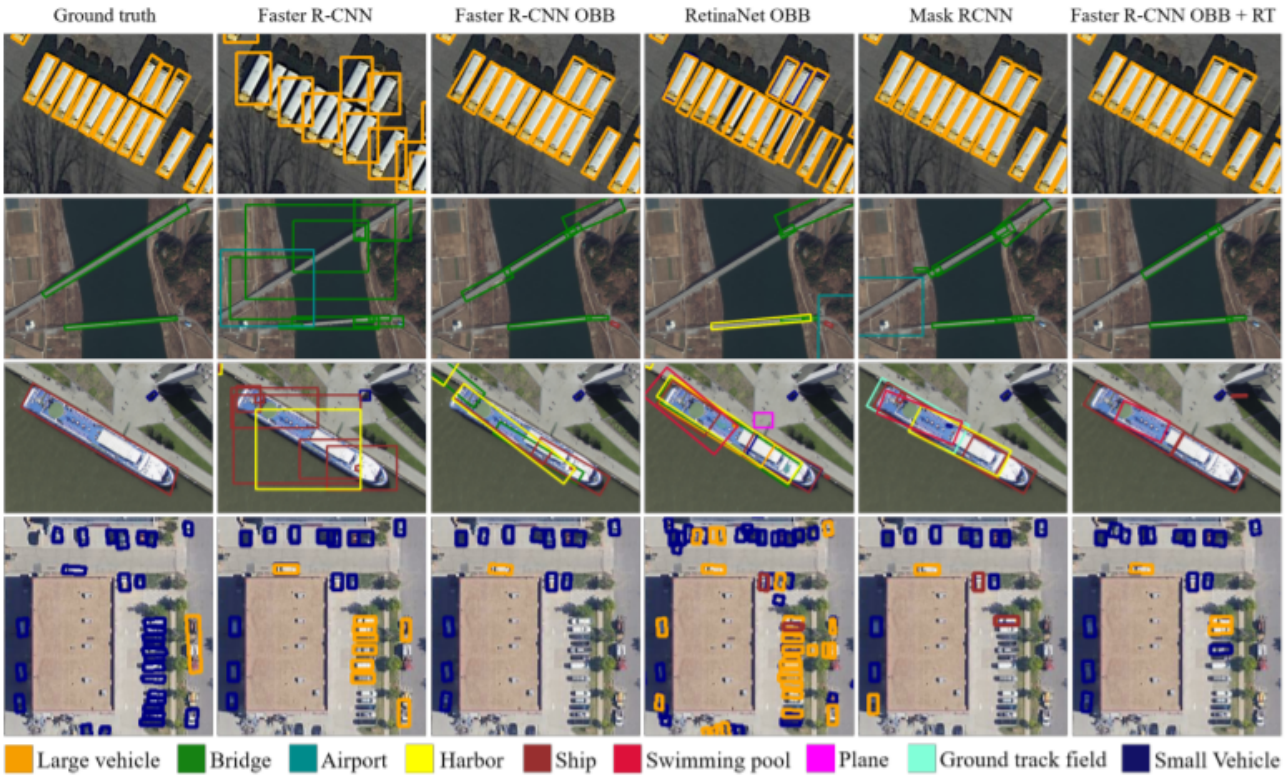

Fig 2. The results of the DOTA-v2.0 test-dev are visually represented, showcasing the performance of the five models specifically designed for DOTA-v2.0

In summary, this research explores and compares different approaches and techniques in the field of object detection, particularly focusing on oriented object detection and the challenges posed by geometric variations. Furthermore, it addresses the impact of different models, such as Mask R-CNN, Faster R-CNN OBB + RoI Transformer, and Faster R-CNN OBB + Dpool, on performance. Additionally, it highlights the importance of handling numerical instability caused by extremely small

instances. The research findings are visually presented in Figure 2, providing a clear representation of the performance of various networks.

4 Conclusion

The detection of oriented objects is tackled by the OBB (Oriented Bounding Box) head, which utilizes a regression approach, while the mask head employs pixel-level classification for the same purpose. Although the mask head shows better results and faster convergence, it comes at a higher computational cost. When evaluating the DOTA-v2.0 test-dev set, Mask R-CNN outperforms Faster R-CNN H-OBB with a higher OBB mAP by 0.57 points. However, it is important to note that Mask R-CNN operates slower than Faster R-CNN H-OBB, with a decrease in speed by 4 fps. It is worth mentioning that the process of transferring the mask to the OBB is not taken into account in this particular comparison, which could potentially further impact the performance and speed of Mask R-CNN. To visually showcase the performance of the five models specifically designed for DOTA-v2.0, the results of the DOTA-v2.0 test-dev are graphically presented. These visualizations demonstrate the models' capabilities in handling orientation and density variations, as well as instances with larger and smaller aspect ratios (Ars).

Acknowledgment

This work is supported and sponsored under the Grant-In-Aid Scheme, ARDB-DRDO, Ministry of Defense, GOI [ARDB/01/1081990/M/I]. The authors would like to thank Dr. P.V Satyanarayana Murthy, Sr. Principal Scientist, CSIR-NAL, Bangalore for his guidance. The authors also like to thank the management, the Principal, and the Staff of SDM College of Engineering for their constant support.

References

1) Cazzato D, Cimarelli C, Sanchez-Lopez JL, Voos H, Leo M. A Survey of Computer Vision Methods for 2D Object Detection from Unmanned Aerial Vehicles. *Journal of Imaging*. 2020;6(8):78. Available from: https://doi.org/10.3390/jimaging6080078.
2) Al-Kaff A, Martín D, García FT, Escalera ADL, Armingol JM. Survey of computer vision algorithms and applications for unmanned aerial vehicles. *Expert Systems with Applications*. 2018;92:447–463. Available from: https://doi.org/10.1016/j.eswa.2017.09.033.
3) Pradhan PK, Baruah U. Object Detection Under Occlusion in Aerial Images: A Review. In: Dutta, N, editors. Lecture Notes in Networks and Systems;vol. 281. Springer Singapore. 2022;p. 215–227. Available from: https://doi.org/10.1007/978-981-16-4244-9_17.
4) Jadhav R, Patil R, Diwan A, Rathod SM, Inamdar M. Aerial Object Detection and Tracking using YOLOv4 and DeepSORT. In: 2022 International Conference on Industry 4.0 Technology (I4Tech). IEEE. 2022;p. 1–6. Available from: https://doi.org/10.1109/I4Tech55392.2022.9952705.
5) Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*. 2021. Available from: https://doi.org/10.48550/arXiv.2010.11929.
6) Rajjak SS, Kureshi AK. Recent Advances in Object Detection and Tracking for High Resolution Video: Overview and State-of-the-Art. In: 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA);vol. 2. 2019;p. 215–228. Available from: https://doi.org/10.1109/ICCUBEA47591.2019.9128812.
7) Brian KS, Isaac-Medina D, Organisciak TP, Breckon M, Poyser CG, Willcocks, et al. Unmanned Aerial Vehicle Visual Detection and Tracking using Deep Neural Networks: A Performance Benchmark. 2021. Available from: https://openaccess.thecvf.com/content/ICCV2021W/AntiUAV/papers/Isaac-Medina_Unmanned_Aerial_Vehicle_Visual_Detection_and_Tracking_Using_Deep_Neural_ICCVW_2021_paper.pdf.
8) Fan GPDP, Ji X, Qin MM, Cheng. Cognitive vision inspired object segmentation metric and loss function. . Available from: https://doi.org/10.1360/SSI-2020-0370.