Bottle Detection in the Wild Using Low-Altitude Unmanned Aerial Vehicles

Jinwang Wang, Wei Guo, Ting Pan, Huai Yu, Lin Duan, Wen Yang
School of Electronic Information, Wuhan University Wuhan, China
{jwwangchn, weige, ting.pan, yuhuai, duanlin, yangwen}@whu.edu.cn

***Abstract*—In this paper, we propose a new dataset and benchmark for low altitude UAV object detection, aiming to find and localize waste plastic bottles in the wild, as well as to inspire the development of object detection models to be capable of detecting small and transparent objects. To this end, we collect $25,407$ UAV images of bottles with various kinds of backgrounds. Unlike traditional horizontal bounding box based annotation methods, we use the oriented bounding box to accurately and compactly annotate the bottles, which provides more detailed information for subsequent robotic grasping. The fully annotated images contain $34,791$ bottles, each of which is annotated by an arbitrary (5 d.o.f.) quadrilateral. To build a baseline for bottle detection, we evaluate several state-of-the-art object detection algorithms on our UAV-Bottle Dataset (UAV-BD), such as Faster R-CNN, SSD, YOLOv2 and RRPN. We also present an analysis of the dataset along with baseline approaches. Both the dataset and benchmark are made publicly available to the vision community on our website to advance research in the area of object detection from UAVs.**

***Index Terms*—Object Detection, Oriented Bounding Box, Deep Learning, Unmanned Aerial Vehicles**

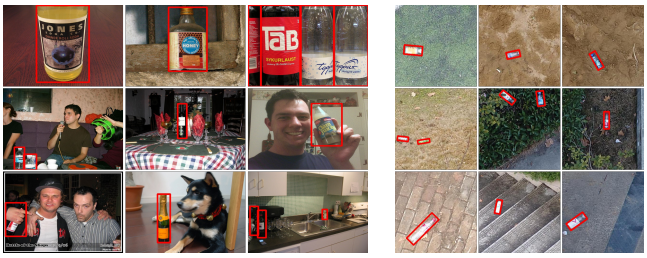
I. Introduction

Nowadays, with the popularity of tourist attractions, there is a lot of rubbish, especially plastic bottles, need to be recycled. However, these bottles are mainly collected by sanitation workers, which is time-consuming, laborious and dangerous, as shown in Fig.1. To solve this problem, we propose to use unmanned aerial vehicles (UAVs) to find and recycle bottles. We also build a UAV bottle dataset (UAV-BD) to detect and locate bottles more effectively. In this paper, we focus on how to detect bottles in UAV images.

Detecting objects in UAV images plays an important role in many applications and has received significant attention in recent years [1]. However, it is still a challenging problem due to the high resolution with the extremely high level of details, various shooting platform, limited annotated data, and limited processing time for real-time applications [2]. In UAV images, the bottles look completely different from the bottles in datasets such as PASCAL VOC [3], Microsoft COCO [4], etc. The difference between PASCAL VOC and our dataset is shown in Fig.2.

As to UAV images, detecting bottles exists several unique challenges. First, the size of bottles is very small, which is generally less than 50×50 pixels. Meanwhile, due to the different altitudes of the UAV system, the size of bottles differs in scale. Second, in UAV images, the backgrounds of the bottles are very complex which usually results in poor detection performance. Third, in contrast to conventional object detection datasets, where objects are generally oriented upward [5], the bottles in UAV images often appear with arbitrary orientations depending on the shooting angle of the UAV camera, as illustrated in Fig.2(b). Fourth, plastic bottles are often transparent, thus the background will can be seen through the bottle, increasing the difficulty of detection. To build a baseline for bottle detection in UAV images, we establish a large scale bottle detection dataset (which we call

Fig. 1. The sanitation workers who are picking up rubbish on the Huashan Mountain[2].



(a) Bottles in PASCAL VOC. (b) Bottles in UAV images.

Fig. 2. Comparison of bottles in PASCAL VOC dataset and UAV images.

[2]http://news.hsw.cn/system/2012/05/16/051321760.shtml

(a) Normal annotation method using horizontal bounding box.

(b) Our annotation method using oriented bounding box.

Fig. 3. The differences between horizontal bounding box and oriented bounding box.

UAV-Bottle Dataset (UAV-BD)) and benchmark.

As we know, convolutional neural networks(CNN) has been applied to solve the object detection problem and the methods based on CNN have achieved state-of-the-art performance [6]. Most of the existing CNN-based detection methods use the horizontal bounding boxes to locate objects in images. The horizontal bounding box is a rotation variant data structure, as shown in Fig.3(a), but it doesn't works well when the detector deals with orientation variations of objects. To overcome this problem, some efforts are made either adjusting the orientation or trying to extract rotation insensitive features. Unlike these methods which try to eliminate the effect of rotation on the feature level, we prefer to make full use of the rotation information for feature extraction so that the detection results involve the angle information. Therefore, the detection results are rotatable, whereas the performance of the detector is rotation invariant [6]. As shown in Fig.3, the same bottle has different horizontal bounding boxes when rotating it, but it has the same oriented bounding box. Moreover, angle information of the bottle is very useful when grasping the bottle with the robotic manipulator.

II. UAV-Bottle Dataset

A. Dataset collection

For dataset collection, we follow four key suggestions: (1) collecting images including bottles a wide range of scale and aspect ratios; (2) collecting images including bottles different background scenes; (3) collecting images including bottles different orientations; (4) collecting as many types of bottles as possible.

The UAV platform used in this work is DJI Phantom 4 Pro quadcopter integrated with a 3-axis stabilized gimbal. Images are collected by a camera mounted on the quadcopter. The resolution of the captured images are 5472×3078 pixels. In order to collect images covering bottles of a wide range of scales and aspect ratios, images at different flight altitudes ranging from $10m$ to $30m$ are collected.

In UAV images, the backgrounds of the bottles are very complex. To increase the diversity of dataset, we divide the collected images into eight scenes. Which are illustrated in Fig.4 and Fig.5. In Fig.4, we show the original images of eight scenes, each scene contains one original image whose sizes are 5472×3078 pixels. In Fig.5, we show the segmented images of eight scenes, each scene contains three subimages whose sizes are 342×342 pixels. Eight background scenes are chosen and annotated in our UAV-BD, including *Bush forest land*, *Waste land*, *Step*, *Mixture*, *Flat ground*, *Plastic stadium*, *Sand land* and *Grassland*.

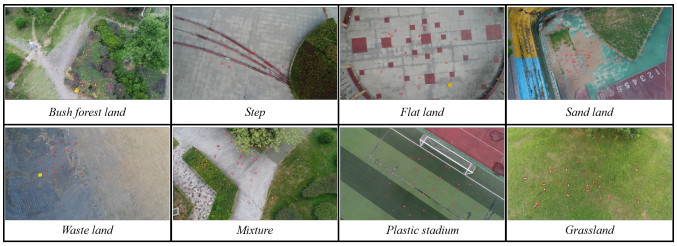


Fig. 4. Samples of annotated images in UAV-BD. We show one full image which size is 5472×3078 per each scene.

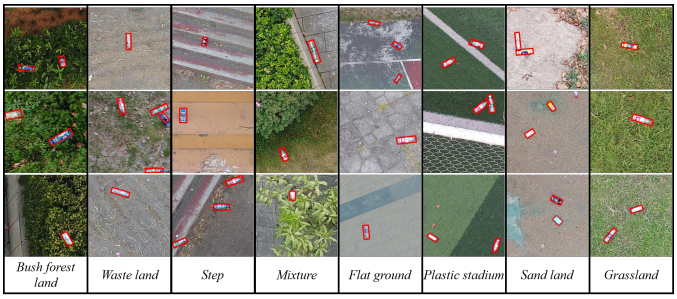


Fig. 5. Sample of annotated images in UAV-BD. We show three images which sizes are 342×342 per each scene.

B. Annotation method

We build the UAV-BD for the bottle detection problem by collecting bottle images using UAV. In the field of computer vision, many visual concepts such as region descriptions, objects, attributes, and relationships, are annotated with horizontal bounding boxes, as show in [5], [7]. A common description of horizontal bounding boxes is (c_x, c_y, h, w) or $(x_{min}, y_{min}, x_{max}, y_{max})$, where (c_x, c_y) is the center location of horizontal bounding box, h, w are the height and width and (x_{min}, y_{min}) is the top left location, (x_{max}, y_{max}) is the bottom right location [5].

Objects with less orientations can be adequately annotated with this method. However, horizontal bounding box cannot accurately or compactly outline oriented instances such as the bottles in UAV images. In UAV images, the overlap between two bounding boxes is sometimes very large that some state-of-the-art object detection methods cannot diffetentiate them [5]. At the same time, horizontal bounding box may contain lots of background pixels while annotating the object, especially when objects with large aspect ratios.

An alternative for annotating oriented objects is using the method of arbitrary quadrilatral bounding boxes. This

TABLE I
IMAGES AND INSTANCES NUMBER IN UAV-BD.

Scenes	n_1	n_2	n_3	n_4
Bush forest land	230	4134	1812	3047
Waste land	379	7598	4355	5800
Step	135	2691	1325	2106
Forest land	285	5724	3702	4891
Flat land	134	2803	1538	2142
Plastic stadium	336	6807	4180	4998
Sand land	249	5570	2704	4008
Grassland	456	9029	5778	7787
Total	2204	44356	25394	34779

annotation method can be expressed as $(x_i, y_i), i = 1, 2, 3, 4$, where (x_i, y_i) denotes the position of the oriented bounding boxes' vertices in the image [5]. The vertices are arranged in a clockwise order. But as bottles are rigid with almost no deformation, therefore we choose a method called θ-based oriented bounding box, as shown in Fig.3(b). This method is often used in text detection benchmarks, expressed as (c_x, c_y, h, w, θ) where θ is the angle from the horizontal direction of the horizontal bounding box [5]. The tool for annotating is roLabelImg[3].

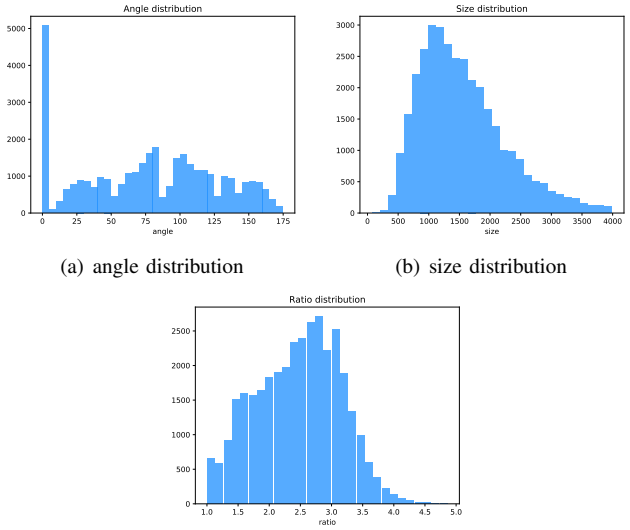
C. Dataset Statistics

UAV images are usually very large in size compared to conventional image datasets. The size of original image in UAV-BD is 5472×3078 pixels, while most images in conventional datasets (e.g. PASCAL VOC and Microsoft COCO) are no more than 1000×1000 pixels [8]. To avoid segmenting the single instances (bottles) into different subimages, we firstly annotate on the original images without segmentation. But we find the original image is too large to be trained for CNN-based algorithms. So we segment each original image into 144 small subimages, and the size of subimages is 342×342 pixels. Note that we abandon the instances at the border. Then we use these subimages to train CNN-based detection model.

The detailed statistics of the UAV-BD is shown in the TableI, where n_1, n_2, n_3, n_4 are the number of original images, subimages, instances in original images, instances in subimages for each scene, respectively. The UAV-BD contains about $34,791$ object instances in $25,407$ images. The "Grassland" scene has the largest number of object instances: $7,795$ instances in $5,785$ images. The "Step" scene has the smallest number of instances: 2106 instances in $1,325$ images.

As bottles usually have rigid body, thus we can get some prior information to train the detection model. For example, we can use the distribution of angle, size and ratio as prior information to improve the performance of detection model. For UVA-BD, we plot the distribution of angle, size and ratio respectively, which are illustrated in Fig.6. From Fig.6(a), we can see that bottles' angles in the dataset are almost uniform. And most of bottles' sizes are range from 500 pixels to 3000 pixels while the ratios of bottles are mostly range from 1.0 to 4.0, which are shown in Fig.6(b) and Fig.6(c). Note that we could use these statistics data to design object-specific detection models.

[3]https://github.com/cgvict/roLabelImg.git



(a) angle distribution (b) size distribution

(c) ratio distribution

Fig. 6. The angle, size and ratio distribution of UAV-BD.

III. BASELINES AND METHODS

The data for all experiments comes from UAV-BD. In order to ensure that the distributions of training and testing data approximately match, we randomly select 64% of the UAV-BD as the training data, 16% as validation data, and 20% as the testing data. The whole UAV-BD contains 16258 images with 22211 instances for training, 5081 images with 6944 instances for testing and 4055 images with 5624 instances for validation. All the original images and segmented images with ground truth for UAV-BD will be publicly available.

Here, we compare four kinds of approaches which differ in the use of detection framework and data annotating method. For horizontal object detection, we select Faster R-CNN[4] [9], SSD[5] [10] and YOLOv2[6] [11] as our baseline testing algorithms for their excellent performance on general object detection. For oriented object detection, we modify the original Rotation Region Proposal Networks(RRPN[7]) algorithm [12] to predict properly oriented bounding boxes. This annotation can be denoted as $\{c_x, c_y, h, w, \theta\}$, where $(c_x,\ c_y)$ is the central coordinate of the oriented bounding box, h, w and θ is the height, width and rotation angle of the oriented bounding box respectively.

A. Baselines with Horizontal Bounding Boxes

Ground truths for horizontal bounding boxes(HBB) experiments are generated by calculating the axis-aligned bounding

[4]https://github.com/rbgirshick/py-faster-rcnn.git
[5]https://github.com/weiliu89/caffe.git
[6]https://pjreddie.com/darknet/yolo/
[7]https://github.com/mjq11302010044/RRPN.git

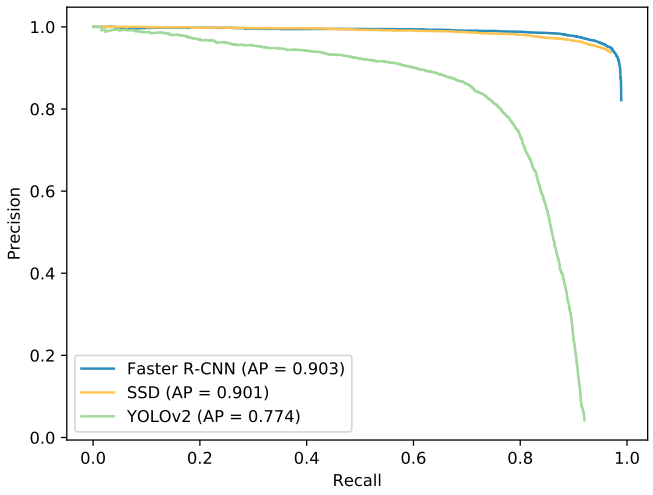


Fig. 7. Numerical results (AP) of baseline models evaluated with HBB ground truths.

boxes over original bounding boxes. To make it fair, we keep all the experiments' setting and hyper-parameters the same as depicted in corresponding papers [9]–[11].

The experimental results of HBB prediction are shown in Fig.7. The blue one illustrates the result for Faster R-CNN, the orange one illustrates the result for SSD and the green one illustrates the result for YOLOv2.

B. Baseline with Oriented Bounding Boxes

Prediction of oriented bounding boxes(OBB) is difficult, because the present state-of-the-art detection methods are not designed for oriented objects. Therefore, we choose Rotation Region Proposal Networks(RRPN) [12] as the framework for its accuracy and efficiency. Then we modify it to adapt UAV-BD with its prior information mentioned in section II-C.

RRPN is based on Faster R-CNN. For Faster R-CNN, the Region of Interests (RoIs) are generated by Region Proposal Network(RPN), and the RoIs are rectangle which can be written as $R = (x_{min}, y_{min}, x_{max}, y_{max}) = (c_x, c_y, h, w)$. These RoIs have regressed from k anchors which are generated by some predefined scales and aspect ratios. But in RRPN, except for predefined scales and aspect ratios, it also uses *angles* to generate RoIs. That is the reason why RRPN can predict oriented bounding boxes which can be written as $R = (c_x, c_y, h, w, \theta)$. In the section II-C, we analyze the size, aspect ratio and angle distributions of UAV-BD, so we can select reasonable scale, aspect ratio and angle values to generate new anchors which are shown in Fig.8. The experimental results of RRPN, SSD, Faster R-CNN and YOLOv2 based on OBB ground truth are shown in Fig.9. The red one shows the result of RRPN.

C. Experimental Analysis

We use precision-recall curves(PRCs) and average precision(AP) values to compare three kinds of baseline models (Faster R-CNN, SSD and YOLOv2), which are evaluated with

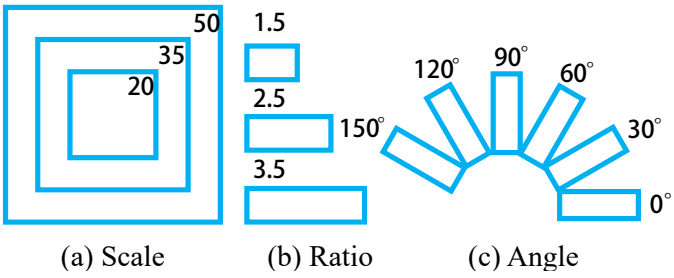


Fig. 8. Anchor strategy in our framework of RRPN.

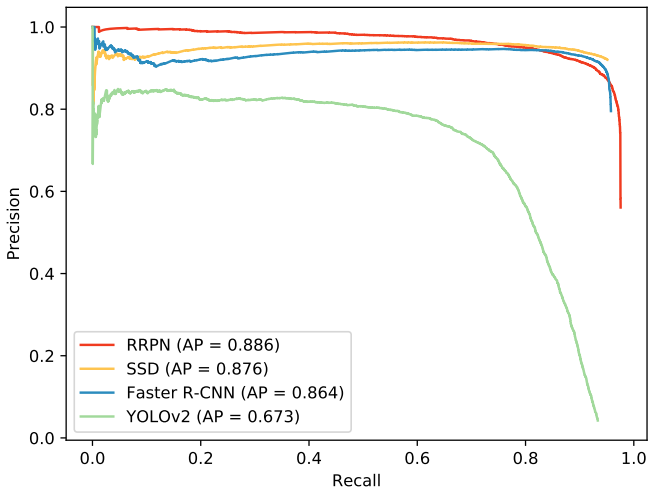


Fig. 9. Numerical results (AP) of baseline models evaluated with OBB ground truths.

HBB ground truth. For evaluation metrics, we use the same AP calculation as for PASCAL VOC. As shown in Fig.7, the AP values of Faster R-CNN, SSD, YOLOv2 are 90.3%, 90.1%, 77.4%, respectively.

Then we still use the same evaluation indicators to compare four kinds of baseline models (SSD, RRPN, Faster R-CNN and YOLOv2). The difference is that we set $\theta = 0$ of Faster R-CNN, SSD, YOLOv2's results and evaluated these models with OBB ground truth. As shown in Fig.9, the AP values of RRPN, SSD, Faster R-CNN and YOLOv2 are 88.6%, 87.6%, 86.4%, 67.3%, respectively. We can observe when using OBB ground truth, the performances of the three baseline methods decrease compared with that using HBB ground truth, thus because of when we set $\theta = 0$, the localization error will increase with OBB ground truth. We can clearly see that the result of RRPN is the best.

In Fig.10, we show the results of different object detection experiments with HBB and OBB ground truth. For oriented bottles shown in Fig.10, the location precision of bottles in HBB are much lower than that of OBB. We can find that OBB regression is the correct way for oriented object detection. Which makes it possible for oriented bottle detection to be efficiently integrated in bottle grasping. In addition, the results of RRPN show the highest localization accuracy and the the

Fig. 10. Visualization results of testing on UAV-BD using well-trained Faster R-CNN, SSD and RRPN. **Top** to **Bottom** respectively illustrate the results for Faster R-CNN, SSD, YOLOv2 and RRPN.

lowest false-alarm and false positive.

IV. Conclusion and Future Work

In this paper, we have built a large-scale dataset for bottle detection in UAV images, i.e., UAV-BD. In contrast to general object detection benchmarks, we annotate a huge number of well-distributed bottles with oriented bounding boxes. We believe this dataset is challenging and very useful for real vision based bottle recycling. Based on this dataset, we also establish a benchmark for bottle detection and show the feasibility to produce oriented bounding boxes which provides more useful information for bottle grasping.

In future work, we will focus on locating and recycling bottles in the real-world using UAV.

Acknowledgment

We thanks Ruixiang Zhang, Jiaqi Xiong, Siyao Zhou, Hao Li, and all the others who involoved in the annotations of UAV-BD.

References

[1] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 3, pp. 1635–1647, 2014.

[2] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "Car detection from low-altitude uav imagery with the faster r-cnn," *Journal of Advanced Transportation*, vol. 2017, 2017.

[3] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[5] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *IEEE CVPR*, 2018.

[6] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," *arXiv preprint arXiv:1711.09405*, 2017.

[7] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[8] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-cnn for small object detection," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 214–230.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[11] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[12] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *arXiv preprint arXiv:1703.01086*, 2017.