THE EFFICIENCY OF GSD NORMALIZATION FOR OBJECT DETECTION IN AERIAL IMAGES

Taegoo Kim (1), Ildoo Kim (1), Jonghyuck Park (1), Jihoon Lee (1)

[1] Kakao Brain, Seongnam-si, Gyeonggi-do, 13494, Korea
Email: taegoo.kim@kakaobrain.com; ildoo.kim@kakaobrain.com;
jonghyuck.park@kakaobrain.com; jihoon.lee@kakaobrain.com

KEY WORDS: Object Detection, GSD Normalization, Faster R-CNN, Deep Convolutional Neural Network

ABSTRACT: Object detection in aerial images is an active yet challenging task in computer vision due to crowded small objects, its arbitrary orientations and object scale variations. Dealing with various sizes of objects is a common problem in conventional image task. In recent research, a data augmentation technique that transforms various sizes of images has been considered to regularize the size variations. The aerial image, itself contains GSD (Ground Sample Distance) information indicating how much of the actual distance a pixel covers, if the camera altitude and angle are known. This allows the various sizes of real objects to be uniformly normalized. In this research, GSD normalization, an image augmentation technique that modifies sizes of images to satisfy the predefined target GSD value will be introduced to improve the performance measure of object detection in aerial images. To validate the proposed method, DOTA (A Large-scale Dataset for Object DeTection in Aerial Images) has been utilized with Faster R-CNN as an object detector using deep convolutional neural networks. Finally, it is expected that it will be helpful to apply the proposed method when there is insufficient number of images at the altitude suitable for the target task in remote sensing field.

1. INTRODUCTION

Due to the recent rapid development of Convolutional Neural Network (CNN), object detection, one of the most challenging tasks in computer vision, has been much matured to be applied in various applications such as autonomous vehicles (Zhu et al., 2016), surveillance (Lin et al., 2015), monitoring (Guirado et al, 2017), etc. Nonetheless, object detection in aerial image is still remained as unsatisfactory (Yang et al., 2019) due to the various sizes and aspect ratios of objects in aerial images. Xia et al. (2017) indicated several distinct points that explain why an aerial object detector performs relatively low accuracy, compared to an object detector based on ground images, which are huge object size variance, small object crowdedness, unbalanced instance occurrences and arbitrary orientation of objects.

As mentioned earlier, size variance of objects in images deteriorates the performance of aerial object detector due to wide range of a corresponding distance and field of view as images are captured. Theoretically, size invariance problem can be easily resolved by simply resizing images to obtain objects in same scale in respect to photography properties. In the traditional ground image dataset, it is challenging to obtain or estimate the camera information out of the image, required to compute the optimal object size distribution within the image for normalization.

However, in aerial images, it is possible to compute a ground sample distance (GSD) that indicates how much a pixel represents the actual distance (Orych, 2015). If the pixel distance of two points in an image and physical distance of the same points in the real world are given, GSD can be easily calculated by the equation (1). GSD can be computed with the information of shooting altitude and camera's angle of view rather than measuring the actual distance of two points. Using this feature, the size of objects in all images can be uniformly normalized by GSD property. As will be discussed in Chapter 3, the performance of an object detection is distinctly improved by the proposed GSD normalization technique.

$$\text{Ground Sample Distance (GSD; m/px)} = \frac{Actual/Physical\ Distance\ Between\ Two\ Points\ (m)}{Distance\ in\ a\ Image\ Between\ the\ Same\ Two\ Points\ (px)} \qquad (1)$$

1.1 Previous Works

Object detection has been significantly improved by the advent of deep learning. As having been sophisticated, there are mainly two classes of methods for object detection in images, one based on sliding windows and the other based on region proposal classification.

To reduce the computation cost of selecting a huge number of regions, Girshick et al. (2013) proposed a method to use selective search to extract a relatively small number of regions from the image, called region proposals. These regions are generated using the selective search algorithm, which generates initial candidate regions and recursively combine similar regions into larger ones.

Liu et al. (2016) proposed Single Shot MultiBox Detector (SSD), one single shot detector to detect multiple objects within the image, while regional proposal network (RPN) based approaches need two stages, one for generating regions proposals, one for detecting the object for each proposal. By eliminating the object proposals, SSD requires less computation time compared to two-shot RPN-based approaches.

To resolve class imbalance issue that one-stage detectors experienced, Lin et al. (2017) proposed a new loss function, called Focal Loss, that dynamically scaled cross entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. Apparently, this scaling factor automatically down-weight the contribution of easy examples during training and rapidly focus the model on hard examples. The proposed loss function was used to substitute heuristics or hard example mining techniques, the previous one-stage detectors had, which significantly outperforms the existing techniques. This Focal Loss function was validated with a simple one-stage object detector called RetinaNet. The main characteristics of this detector was use of feature pyramid network (FPN) and anchor boxes.

The most significant and challenging problem for object detection in remote sensing era is the various scales and aspect ratios of objects within the image. To resolve this issue, Qiu et al. (2019) proposed a new object detection modal, called A2RMNet composed of gate fusion modules, refine blocks and region proposal networks. The multi-scale feature gate fusion network adaptively aggregates semantic features of different scale features and control feature information in different scale by using the learned weight vector. Furthermore, an attention network was proposed to select the RoI features of appropriate aspect ratios of objects.

For similar purpose, researches on handling image data have also been populated. Touvron et al. (2019) proposed a method to reduce the train-test resolution discrepancy, which allows the training and testing data distributions to be matched by rescaling images at both train and test time accordingly.

Table 1. Descriptive Statistics of Selected Images from DOTA

Feature	Mean	Std. Dev.	Min	Max
GSD (m)	0.249	0.171	0.092	0.988
width (px)	2108.92	1553.09	353	13,383
height (px)	1967.11	1328.24	346	8,115
Instances / Image	181.85	501.32	1	10,180

Table 2. Descriptive Statistics of the Size of Annotated Instances

Category	Count	Proportion	Mean	Std. Dev.	Min	Max
Helicopter	691	0.3%	61.2	49.8	22.4	482.8
Large Vehicle	26,999	10.9%	47.5	26.7	4.6	280.1
Plane	9,719	3.9%	109.4	91.9	8.5	849.0
Ship	41,468	16.7%	38.3	30.4	5.2	1281.9
Small Vehicle	168,983	68.2%	17.7	9.6	2.8	106.3
Total	247,860	100.0%	28.1	32.0	2.8	1281.9

* The size of each bounding box was measured by the geometric mean of height and width (in pixel)

While many studies on image characterization have focused on improving performance using the latent information contained in the image, our approach is clearer than other approaches in that it uses more explicit image information, GSD. The major strength of our method is that it can be quantitatively validated through a more adaptive approach that depends on the ambient condition when image data is captured and capabilities of a camera.

The rest of this paper is organized as follows: materials and methods used for our research are introduced in Chapter 2. The experiments setup and the results are explained and analyzed in Chapter 3. The contribution of this paper and future research are described in Chapter 4.

2. MATERIAL AND METHODS

2.1 Dataset

A fair number of aerial image datasets for object detection have been released to the public and DOTA is one of them. It has been widely used among researchers due to its abundance of size, many different classes of objects, and quadrilateral annotation rule unlike simple rectangle annotation method of other known public datasets (Xia et al., 2017). Thus, DOTA was unquestioningly selected for our research. The recently released DOTA v1.5 dataset is composed of 1,869 images with 280,196 annotated instances, increased by 219% compared to its previous version (excluding the test images of which the annotations were not provided).

In addition, DOTA differs from other datasets due to wide range of GSD distribution over images. In the case of COWC (Mundhenk et al. 2016) published for automobile detection, the data consists only of images with a GSD value of 0.15. Thus, it was undeniable to use DOTA to thoroughly validate our normalization method using GSD property for aerial images.

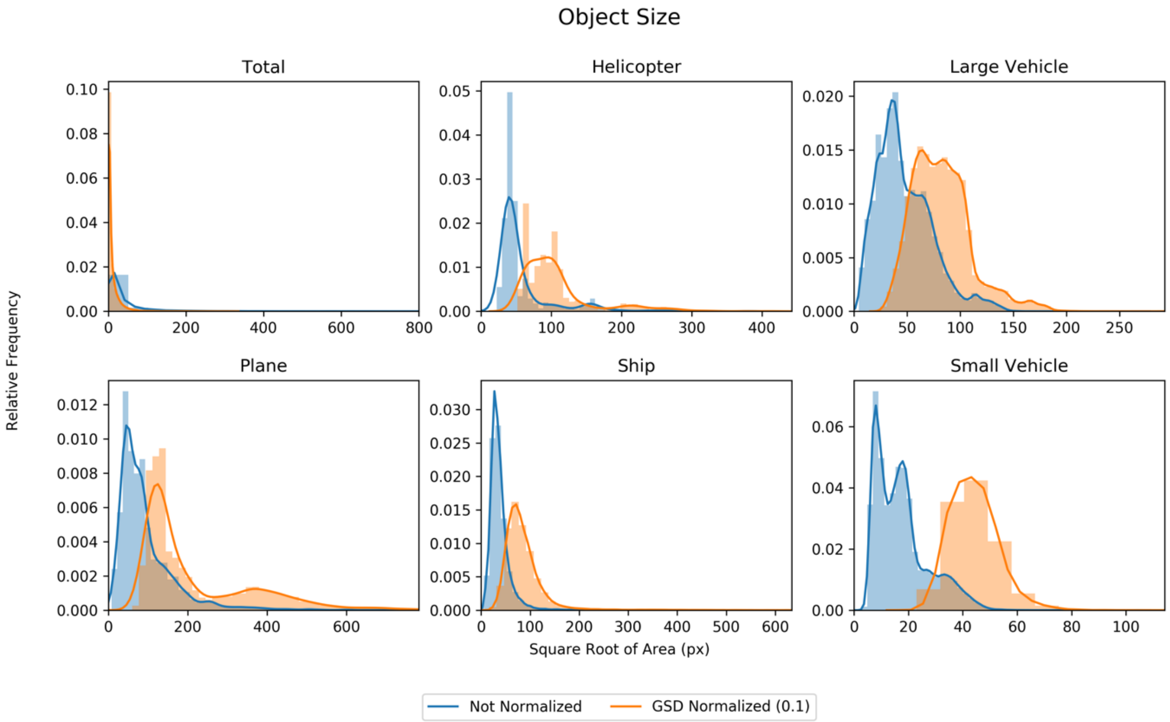

Figure 1. The Size Distribution of Bounding boxes (by Category). The object sizes are measured by the geometric mean of height and width.

DOTA provides annotated instances for a total of 16 categories, but analyzed only by extracting image information about moving objects that are interested in industries such as surveillance, public traffic control, city planning, etc. The moving objects are helicopters, large vehicles, planes, ships, and small vehicles.

Among 1,869 images of the annotated DOTA images, our dataset was composed of a total of 1,363 images which contain the target objects and whose GSD values are lower than 1.0 which cannot distinctly express the target objects. Table 1 and 2 illustrate descriptive statistics of selected images and annotated instances. As shown in Table 1, there were images of various sizes and the average number of instances in one image was 181.85, which confirmed the crowdedness of the object throughout the aerial image dataset. Plus, it was observed that more than two-thirds of instances were a class of small vehicles. In general, an instance with area of less than 32px^2 in an image is categorized as small-sized object according to the public open dataset, COCO (Lin et al., 2014). In that sense, our selected dataset was composed of crowded small-sized instances. Of the total 1,363 images, 272 images (20%) were used as test dataset and the remaining 1091 images (80%) were used for training.

2.2 Methods

GSD Normalization

GSD normalization is a newly proposed image augmentation technique to change the size of images so that the modified image gets the target GSD. The image becomes enlarged when the GSD of the original image is larger than the target GSD and vice versa. Interpolation methods should be selected to modify the size of images. The interpolation method based on resampling

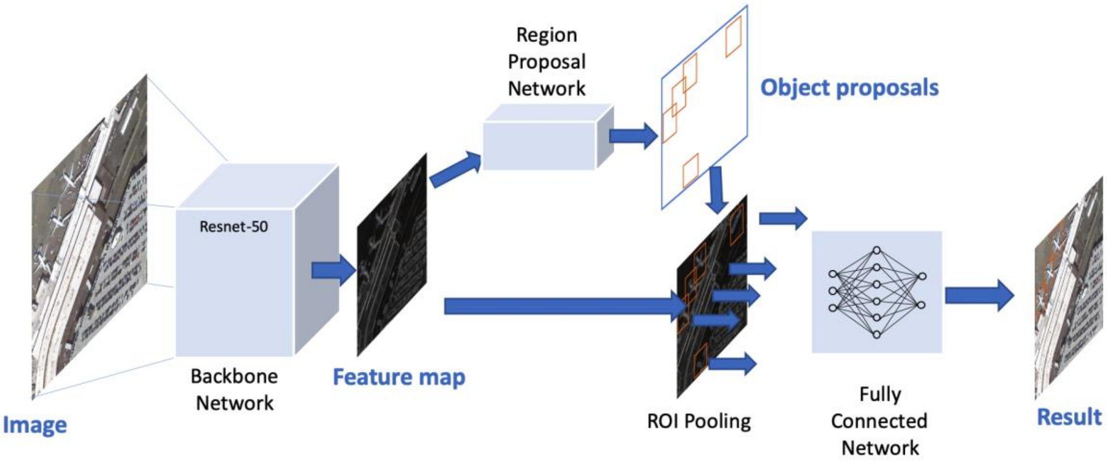

Figure 2. The Structure of Faster R-CNN

using pixel area relation was utilized to decrease the size of images and bicubic interpolation was applied for enlarging images on the basis of OpenCV official documentation (OpenCV team, 2019).

In general, data augmentation in deep learning era feeds the augmented images to the training network while training is being conducted. However, image resizing requires longer computation time. It is also inefficient to crop the image to meet the desired input size of the network throughout the training process. Moreover, random cropping, a common data augmentation method, doesn’t fully guarantee the same outcomes due to its randomness. In order to improve the efficiency of experiments and obtain reproducible results, the dataset was divided into training and test datasets and arranged in a required format in advance.

Figure 1 illustrates the size distribution of bounding boxes grouped by their categories. The sizes of normalized bounding boxes are larger than that of non-scaled dataset; most of the images were enlarged since the average GSD value of unnormalized data is 0.249. Besides this enlarging effect on the size distribution of bounding box become looking more like bell-shape especially in small vehicles.

Object Detector

In this experiment, Faster R-CNN (Ren et al., 2015) was used to build an object detection model for aerial imagery. Figure 2 shows the overall scheme of Faster R-CNN, composed of 3 networks: the Backbone Network that extracts feature maps within an image, the Region Proposal Network that selects areas where objects are more likely and the Fully Connected Network which finds classes of objects and more precise locations of objects based on selected regions.

Even if it is a combination of three independent neural networks, Faster R-CNN is a single and unified network model for object detection. Faster R-CNN is a fully differentiated model that is fast and utilizes advantages of CNN while previous works in R-CNN are based on the selective search to generate region proposals.

Table 3. The Number of Images for Training by Treatment

	Not Normalized	GSD Normalization (target GSD)				
		0.1	**0.15**	**0.2**	**0.3**	**0.4**
Number of Images	10,846	30,012	19,518	13,919	8,389	5,790

Table 4. Average Precision by Category

Category	Not Normalized	GSD Normalization (target GSD)				
		0.1	**0.15**	**0.2**	**0.3**	**0.4**
Helicopter	0.389	**0.719**	0.630	0.659	0.689	0.580
Large Vehicle	0.738	**0.781**	0.763	0.728	0.624	0.513
Plane	0.928	0.935	0.928	**0.938**	0.928	0.915
Ship	0.725	**0.927**	0.899	0.831	0.619	0.490
Small Vehicle	0.473	**0.771**	0.735	0.680	0.332	0.070
mAP	0.650	**0.827**	0.791	0.767	0.638	0.514

3. Experiments

3.1 Experimental Design

We trained datasets with target GSD of 0.1, 0.15, 0.2, 0.3, and 0.4, respectively. For comparison, the dataset without GSD normalization was also trained by identical Faster R-CNN model. The size of the input image in the Faster R-CNN object detector were set to 800 x 800. Object detector training and inferencing were conducted with the use of MMDetection: Open MMLab Detection Toolbox (Chen et al, 2019), and original config file of Faster R-CNN provided by the toolbox was used except the input size of images and total number of epochs (set to 300).

As the target GSD values is decreased, the size of augmented image becomes larger. If the GSD of the original image was 0.4, the image is increased by two times horizontally and vertically when the target GSD is set to 0.2, and the area is increased by four times in total. As described in Section 2.2, the original image is augmented by our GSD normalization method. Then, the modified image is cropped using sliding window, allowing 20% overlap for both width and height and fit to the desired input size. The number of images for each augmented training dataset according to each experimental condition is shown in Table 3.

All experiments were performed on the environments called Brain Cloud, GPU-powered cloud computing service developed by Kakaobrain. Each training was conducted on V4.XLARGE instance (4 NVIDIA Tesla V100 installed machines with 56-cored CPU and 488 GB RAM).

3.2 Result

Table 3 shows average precision score for each class. The evaluation metric for PASCAL VOC (Everingham et al., 2010) was applied. Due to the fact that we use the independently modified

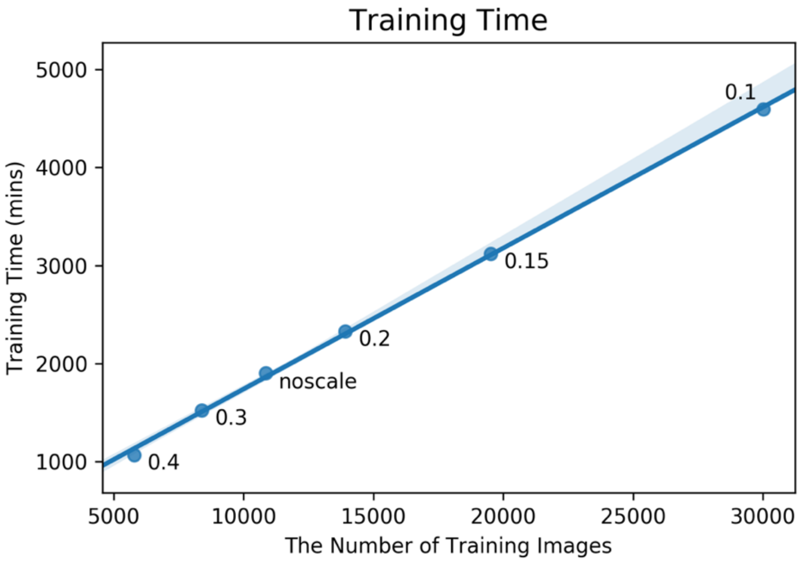

Figure 3 Training Time Dependence of the Number of Training Images

training and test datasets, aggregated by the moving window method explained in the earlier section, there may little discrepancy between our scores with the official DOTA benchmarks. Compared to the result without GSD normalization, our proposed method outperforms in all categories.

The GSD value to achieve the most higher AP score was when it was set to 0.1. Its mean average precision (mAP) score was 27.2% higher than the result without GSD normalization, and the AP score for small vehicles was increased by 63.0%. It was investigated that the overall performance was improved with lower GSD values.

Basically, GSD normalization is an augmentation technique based on resizing for both training and test images. When target GSD is set to small number, excessive number of images need to be generated for training. Figure 3 illustrates that the relationship between training time and the number of training images in our experiments. The training process spends a certain amount of time to learn the entire images generated by GSD normalization, which indicates that the GSD normalization helps the model to learn more data.

However, for certain categories such as plane, GSD value of 0.2 achieved better AP score than the score of when GSD value was set to 0.1. It implies that there are certain sets of GSD values for each category to attain the most outstanding performance. In other words, the performance for an object detector based on deep learning seems to be sensitive to the size distribution of objects not just depends on the size of objects. It is important to reduce these variances by suitable normalization techniques.

4. CONCLUSION

In this paper, we have investigated how our GSD normalization performs on object detection in aerial images. In addition, it was proved that the GSD normalization evidently helps to build a better object detector by adjusting the overall distribution over its object size for both training and testing phases. It is expected that the proposed GSD normalization is highly applicable in remote sensing field. The existing aerial images from satellite and planes could be used to develop an object detector for Unmanned Aerial Vehicles (UAVs) by applying our GSD normalization technique. Therefore, it would save the budge to collect aerial images using UAVs.

However, it is imprecise whether the improvement depends on the excessive number of training images generated by GSD normalization or the normalized distribution of object sizes. In addition, Faster R-CNN was the only object detection model used to examine the efficacy of our GSD normalization. In the future, we hope to experiment our proposed GSD normalization method on the state-of-the-art object detection models that are robust to scale.

REFERENCES

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., … Lin, D., 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. ArXiv:1906.07155 [Cs, Eess].

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303–338.

Girshick, R., Donahue, J., Darrell, T., & Malik, J., 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. ArXiv:1311.2524 [Cs].

Guirado, E., Tabik, S., Alcaraz-Segura, D., Cabello, J., & Herrera, F., 2017. Deep-learning Versus OBIA for Scattered Shrub Detection with Google Earth Imagery: Ziziphus lotus as Case Study. Remote Sensing, 9(12), 1220.

Lin, K., Chen, S.-C., Chen, C.-S., Lin, D.-T., & Hung, Y.-P., 2015. Abandoned Object Detection via Temporal Consistency Modeling and Back-Tracing Verification for Visual Surveillance. IEEE Transactions on Information Forensics and Security, 10(7),

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P., 2017. Focal Loss for Dense Object Detection. ArXiv:1708.02002 [Cs].

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., … Dollár, P., 2014. Microsoft COCO: Common Objects in Context. ArXiv:1405.0312 [Cs].

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C., 2016. SSD: Single Shot MultiBox Detector. ArXiv:1512.02325 [Cs], 9905, 21–37.

Mundhenk, T. N., Konjevod, G., Sakla, W. A., & Boakye, K., 2016. A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision – ECCV 2016 (Vol. 9907, pp. 785–800).

OpenCV team, 2019. Geometric Image Transformations. In OpenCV Document for v4.1.1 Retrieved Jul 26, 2019, from https://docs.opencv.org/4.1.1/da/d54/group__imgproc__transform.html

Orych, A., 2015. Review of Methods for Determining the Spatial Resolution of UAV Sensors. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-1/W4, 391–395.

Qiu, H., Li, H., Wu, Q., Meng, F., Ngan, K. N., & Shi, H., 2019. A2RMNet: Adaptively Aspect Ratio Multi-Scale Network for Object Detection in Remote Sensing Images. Remote Sensing, 11(13), 1594.

Ren, S., He, K., Girshick, R., & Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. ArXiv:1506.01497 [Cs].

Touvron, H., Vedaldi, A., Douze, M., & Jégou, H., 2019. Fixing the train-test resolution discrepancy. ArXiv:1906.06423 [Cs].

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., … Zhang, L., 2017. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. ArXiv:1711.10398 [Cs].

Yang, F., Fan, H., Chu, P., Blasch, E., & Ling, H., 2019. Clustered Object Detection in Aerial Images. ArXiv:1904.08008 [Cs].

Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., & Hu, S., 2016. Traffic-Sign Detection and Classification in the Wild. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2110–2118.