Low Altitude, Slow Speed and Small Size Object Detection Improvement in Noise Conditions Based on Mixed Training

J. F. Que*, H. F. Peng, and J. Y. Xiong

National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu, China.

*E-Mail: 525706660@qq.com

Abstract. The low altitude, slow speed and small size object which we call LSS-object for short, such as small UAV (unmanned aerial vehicles) have become a hot issue of air defense security, which is difficult to detect and identify accurately from the image. In this paper, aiming at the problem of LSS-object detection under noise environment, the detection method based on deep learning is proposed. Firstly, a standard training dataset consisting 5 classes of typical objects is constructed. Then, the standard dataset is augmented with noise of different intensity. Finally, YOLO v3 algorithm is used to form a LSS-object detection system which can adapt to environment noise. The training and detection experiments were carried out on the GPU server. After only using the noise-free dataset for training, the mAP(mean Average Precision) of the noise-free test set detection reached 81.07%, but the mAP decreased to 20.68% when the noise variance was 0.03. After adopting the mixed training strategy of the dataset with noise variance of 0.01 and noise-free data, the mAP for the test set detection with noise variance of 0.03 was increased to 70.61%, and the mAP still reached 79.85% in noise-free test set detection. The experiment results show that the mixed training strategy can greatly improve the accuracy in the noisy images detection while maintaining a higher accuracy in noise-free images.

1. Introduction

In recent years, with the development and commercialization of consumer-grade UAV technology, the LSS-object represented by rotor-wing UAV have brought enormous air defense pressure to airport security, conference security, anti-terrorism and stability, and national border protection. It has become one of the most important focuses in the national air defense security field. Such objects have low contrast in optical images due to their low flying height, weak size, and inconspicuous features. The detection and identification of such objects have become an important technical problem. In order to improve the detection sensitivity of weak objects in the imaging process, a large amount of noise is present in the image, and a more efficient algorithm is needed for image detection. On the other hand, for such LSS-objects, the system detection range is generally only a few kilometres may even within 1 km, resulting in short warning and disposal time. It is required that optical observation equipment not only achieves fast object detection, but also gives object position. It is necessary to quickly and automatically detect the object, and provide basis for subsequent decision. However, in addition to small fixed-wing and rotor-wing UAV, objects such as balloons and birds often appear in the sky, especially birds, which have high probability of occurrence, and often have similar characteristics to
rotor-wing UAV. Such objects brings great difficulties for classification and identification, so the efficient detection algorithm is very important for the LSS-objects.

Classical object detection algorithms include optical flow method, threshold segmentation method and Deformable Part Model (DPM) [1]. These methods are not ideal using in complex backgrounds. Among them, the best DPM algorithm can only reach 40.4% on the VOC2007 test set. In recent years, with the advance and continuous maturity of deep learning technology, it has achieved fruitful results in the field of object detection, and has become one of the mainstream methods. In 2013, Ross Girshick et al. proposed a region-based convolutional neural network (R-CNN) [2, 3], which is a object detection method combining Region Proposal and Convolutional Neural Network (CNN). The algorithm uses selective search [4] and image pyramid in the generation of candidate regions [5], and subsequently forms SPP-Net [5], Fast R-CNN [6], Faster R-CNN [7] and R-FCN [8] detection algorithm, in which the latest R-FCN algorithm has reached 83.6% mAP on the VOC2007 dataset. However, the region-based object detection algorithm needs to be completed in two steps: one is to extract candidate regions, and the other is candidate region feature extraction, which leads to low object detection rate [9].

Based on the idea of regression, Redmon et al. proposed an end-to-end object detection algorithm YOLO v1, which greatly improved the detection speed while maintaining a high object detection accuracy rate. However, but the positioning error was large due to the inaccuracy of the predicted bounding boxes[9], SSD [10], YOLO v2 [11], YOLO v3 [12] and other algorithms have been improved on the basis of YOLO v1, the detection accuracy has been significantly improved. At the exactly same high accuracy of detection, the speed of the current YOLO v3 method can reach 3.5 times that of R-FCN [8, 12]. The YOLO algorithm's one-time detection is characterized by faster speed and maintaining high accuracy. It meets the high requirements for timeliness and accuracy in LSS-object detection and recognition. However, there is no public standard dataset yet, and its detection performance is unknown especially under special conditions such as noise pollution. Therefore, this paper constructs a LSS-object dataset, and studies the detection and recognition performance under the noise background.

2. Proposed method

2.1. Problem definition

Optical image is the main means to realize LSS-object recognition and confirmation. In this paper we intend to use the latest YOLO v3 deep learning algorithm to achieve efficient detection of LSS-object. The typical deep learning detection training system flow is shown in figure 1.

Aiming at the problem of LSS-object detection and recognition in the background of noise, the main ideas of the deep learning method adopted in this paper are as follows:

1) Construct a standard LSS-object standard training sample set. Aiming at the problem that there is no LSS-object training sample set in public currently, the image in training dataset that meets the actual application standard is selected from the real video and the picture.

2) Training the deep learning network with the dataset. Firstly, the obtained original training dataset is input into the deep neural network model for training, and then the noisy image data is
simulated by original dataset and combined original dataset to input the deep neural network model for mixed training.

3) Using the training results to detect and identify LSS-object. The object detection and recognition experiments were carried out using the network obtained from the original training dataset and the mixed training.

4) Establish evaluation rules. In order to test the network performance, a method is introduced to evaluate detection accuracy, then compare and analysis between the two training results.

2.2. dataset construction
In this paper, the LSS-object dataset will be collected according to the actual scene requirements, and marked according to the VOC2007 dataset format. The dataset covers five categories of LSS-objects, including “rotor-wing UAV”, “fixed-wing UAV”, “micro-helicopter”, “bird”, ”hot air balloon”. In a realistic scene, subject to flight control, it is unacceptable to obtain flight images of LSS-object in the city. Samples are further collected from a large amount of real videos and images on the Internet. Dataset acquisition needs to be ensured that these images have different resolutions, different weather conditions, and different complex scenes. That is to say, the video and image have varies resolution, and the overall quality is poor. The scene includes clouds, ground and water surfaces, and the weather involves sunny, cloudy, and rainy. There are a total of 1429 images, of which 942 are used as training set and the remaining 487 are used as test set. The number of LSS-object dataset constructed in this paper is shown in table 1:

| Classes            | Rotate-wing UAV | Fixed-wing UAV | Micro-helicopter | Bird | Balloon |
|--------------------|-----------------|----------------|------------------|------|---------|
| Number             | 355             | 659            | 244              | 91   | 80      |

The dataset constructed in this paper has two main characteristics:

1) At present, the public image quality in the object detection dataset is generally high, the texture in the image is clear and the area has a lot of pixels. The LSS-object in the dataset of this paper, distinct from these high quality images, intentionally selects some images with unclear and blurred texture features, whose object is tiny used as the dataset, so that the LSS-object detection system obtained by the training has good adaptability to objects of varied scales and sizes.

2) The common methods to augment the dataset, such as “rotation” and ”multi-scale”, are not suitable for the detection of complex noise environments. Although there is some atmospheric noise in the constructed data, there may be more unknown noise in the real environment during the camera-component acquisition. Hence, Gaussian noise of different intensity needs to be added for simulation.

2.3. Object detection framework and training method
This paper selects the YOLO v3 to optimize the network structure based on the framework of YOLO v1 and YOLO v2, and adds a series of improved strategy methods. Compared with the previous traditional network structure, the network used to extract features from the residual model proposed in the resnet network [13]. The residual learning unit is used to allow the original input information to be directly input into the subsequent convolutional layers. Hence, the model can still converge and learn better characteristics, the detection and classification recognition rate can be improved. According to the actual needs, the real-time detection network structure based on YOLO v3 is shown in figure 2. To further improve the recognition accuracy of the model, the input picture pixel size is increased from 416×416×3in original YOLO v3 model to 608×608×3. The network called “darknet-53” has a total of 53 convolutional layers and 5 maxpooling layers, which finally output 19×19×3 feature map. The feature map forms a total of 361 grids. Here, three scale predictions are used, and design of the borders at each scale is predicted. The clustering method is used to obtain 9 cluster centers, and the
scales are equally divided into three types of scale. Each grid in the feature map predicts 3 bounding boxes, each of which will include three kinds of information: 4-dimensional position information (center coordinates $t_x, t_y$), height and width $h,b$), 1-dimensional bounding box confidence and $N$-dimensional category information. Therefore, the number of channels finally output is $9 \times (4+1+N)$.

During the training process, the set is trained by the stochastic gradient descent algorithm in deep learning. The number of samples per batch is set to 64 pictures, that is, the weight parameters of the network is updated by training 64 pictures. Training was advanced according to the parameter settings in table 2 below, and the training error of the final network converges at 10000 iterations.

Table 2. Training parameters

| Parameters       | Value                     |
|------------------|---------------------------|
| Momentum         | 0.9                       |
| Weight Decay     | 0.0005                    |
| Iterations       | 10000                     |
| Learning rate    | 0.00001(0-100)            |
|                  | 0.001(100-2500)           |
|                  | 0.0005(2500-4000)         |
|                  | 0.0002(4000-5500)         |
|                  | 0.0001(5500-7000)         |
|                  | 0.00003(7000-8500)        |
|                  | 0.00001(8500-10000)       |
2.4. Object detection and evaluation method

In order to evaluate the overall performance of the algorithm in the detection of LSS-object detection, the precision and recall of the object detection are used to measure the ability of the final system [14]. First you need to calculate the IOU value which indicates the ratio of the Intersection and union between prediction box and the labeled box. The prediction in the type of each bounding box needs to predict whether there is a LSS-object whose center falls into the inner bounding box, so the confidence formula is calculated as:

\[ \text{Confidence} = \Pr(\text{Object}) \times IOU \]  \hspace{1cm} (1)

\[ IOU = \frac{\text{Detection} \cap \text{GroundTruth}}{\text{Detection} \cup \text{GroundTruth}} \] \hspace{1cm} (2)

In the formula (1), \( \Pr(\text{Object}) \) is indicated whether there is a LSS-object falling into the bounding box, and if there is one, then 1 is taken, and if not, 0 is taken. Detection in Equation (2) represents the prediction table box in system, which represents the test bounding box of the data set. For instance, the "rotor-wing UAV" class, we can definite as following. (1) TP: \( IOU \geq 0.5 \), the object is detected as "rotor-wing UAV" while the label is "rotor-wing UAV"; (2) FP: \( IOU \geq 0.5 \), the object is detected as "rotor-wing UAV" while the label is not "rotor-wing UAV"; (3) FN: \( IOU \geq 0.5 \),the object is not detected as "rotor-wing UAV". (\( IOU < 0.5 \) or classify the wrong category)

In summary, for a total of 487 pictures in 5 categories, an overall evaluation of the network structure detection capability can be made. The formula for calculating the accuracy is:

\[ \text{Precision} = \frac{TP}{TP + FP} \] \hspace{1cm} (3)

The calculation formula for the recall rate is:

\[ \text{Recall} = \frac{TP}{TP + FN} \] \hspace{1cm} (4)

If a classifier performs well for a certain class, then the recall can be increased while maintaining the precision at a very high level. Usually, according to the confidence level of a certain type detected by the classifier, the precision-recall curve can be obtained. A specific value can be used to more intuitively express the detection performance of the system. Usually, Average Precision is used as the metric, which is referred to as the AP value. Its formula is:

\[ AP = \int_{0}^{1} P(r)dr \] \hspace{1cm} (5)

In this integration, \( P \) represents the precision and \( r \) represents the recall. In general, the AP value for the five categories of appeals can be averaged and then measured by a mean Average Precision, generally referred to as the mAP value.

2.5. Mixed training under noisy conditions

Studies have shown that in the process of acquiring an image using an optical sensor, the noise in the image is generally additive noise [15], that is, noise superimposed on the image and independent of the image signal, which is generally considered to be Gaussian distribution and white noise with uniform power spectral density, the magnitude of the noise is measured by its standard deviation or variance [16]. In summary, we can add Gaussian noise to the image of test set to simulate real-world noise. In this paper, the variance of Gaussian noise is used to measure the noise intensity. Due to the randomness of ambient noise, the mean value of Gaussian noise added to the sample is set to zero. The variance of the random Gaussian noise added to the data set of the overall test set is 15 steps: 0.00001, 0.0001, 0.001, 0.005, 0.001, 0.003, 0.005, 0.01, 0.015, 0.02, 0.03, 0.05, 0.1, 0.15, 0.2. In the process of mixed training, the data set is augmented and trained by comprehensively selecting the noise with a variance of 0.01.
3. Experiment results and analysis

3.1. Object detection of the original test set
Using the transfer learning method, the network structure parameters of the "Darknet53" pre-trained in imagenet were selected as the initial weights of the training network. After 942 pictures were trained for about 96 hours, total 10000 iterations, the LSS-object detection system were obtained.

After detecting 487 images in the test data, the results for the overall test set are as follows: TP=472, FP=38, FN=88, the precision rate is 92.55%, and the recall rate is 84.29%. The five classes of AP values are as shown in table 3 below. The mAP value of the original test set is 81.07%. It can be seen from the test results that the AP of birds is low. To analysis the reason, we can know that the birds are too complicated due to the various categories and the texture features are not clear to distinguish when the object pixels are small. And the morphological characteristics of birds are relatively high similar with various types of UAV. Moreover, the training data of the bird data we set in this paper is small, the training data is only 61, which leads to the low test accuracy.

3.2. Object detection of test set in the environment of Gaussian Noise
Based on the original test set, the random Gaussian noise with the variance from 0.00001 to 0.2 is added. Choosing the noise intensity 0.01 as an example, the AP value and AP decrease of the 5 classes of LSS-objects are shown in table 4 below. TP=162, FP=44, FN=398, the accuracy rate is 78.64%, and the recall rate is 28.92%. The mAP value of the final model is 41.91%, and the mAP value is decreased by 39.16%. It can be seen that the current LSS-object network is less capable of combating random Gaussian noise in the background, and the accuracy of the network model is reduced by 13.91%. The full rate has dropped by 55.37%. The detection results of the test set under various intensity Gaussian noise are shown in table 3 and figure 3.

| Gaussian noise variance | Rotate-wing UAV | Fixed-wing UAV | Micro-helicopter | Bird | Balloon | mAP value |
|-------------------------|-----------------|----------------|------------------|------|---------|-----------|
| 0                       | 89.30%          | 90.52%         | 79.95%           | 57.59% | 87.99% | 81.07%    |
| 0.00001                 | 89.30%          | 90.52%         | 79.82%           | 57.34% | 88.04% | 81.00%    |
| 0.001                   | 89.25%          | 90.47%         | 79.93%           | 57.05% | 88.09% | 80.96%    |
| 0.005                   | 81.56%          | 87.85%         | 78.56%           | 55.37% | 88.62% | 78.39%    |
| 0.03                    | 80.56%          | 62.27%         | 69.81%           | 45.77% | 80.47% | 67.78%    |
| 0.05                    | 68.85%          | 43.53%         | 58.73%           | 38.34% | 79.66% | 57.82%    |
| 0.1                     | 51.34%          | 17.85%         | 46.05%           | 25.22% | 69.10% | 41.91%    |
| 0.015                   | 35.50%          | 9.09%          | 35.73%           | 22.89% | 58.29% | 32.30%    |
| 0.02                    | 26.92%          | 9.09%          | 29.36%           | 19.69% | 49.96% | 27.01%    |
| 0.03                    | 23.83%          | 9.09%          | 20.97%           | 11.29% | 38.22% | 20.68%    |
| 0.05                    | 9.09%           | 9.09%          | 12.25%           | 9.41%  | 16.16% | 11.20%    |
| 0.1                     | 9.09%           | 0              | 9.09%            | 0      | 0.01%  | 3.64%     |
| 0.15                    | 0              | 0              | 9.09%            | 0      | 0      | 1.82%     |
It can be seen that such a LSS-object detection system is very sensitive to possible environment noise. Therefore, the practical application of the system needs to be improved.

3.3. Adding Gaussian noise to construct new training set for mixed training

The 942 images of the test set under the original standard data were added the Gaussian noise with a variance of 0.01 and the original test set of 942 pictures to form a new training set of 1884 pictures. The training parameters are not changed in the 10000 iterations, and the training error of the final network can still reach convergence.

In order to evaluate the new test results, the test set remains unchanged. The experiment results of the improved LSS-object detection system under the test of various intensity Gaussian noise are shown in table 4 and figure 4.

To show the detection results in test set with Gaussian noise 0.03, we select representative detection images for the five classes of objects. The results of the improved object detection system are shown in figure 5(b), figure 6(b) and figure 7(b), comparing with the original system, we can find that the detection in both position and categories are correct through the improved system. However, the objects in figure 5(a), figure 6(a) and figure 7(a) are failed to detect in the original system.

| Gaussian noise variance | Rotate-wing UAV | Fixed-wing UAV | Micro-helicopter | Bird | Balloon | mAP value |
|------------------------|-----------------|----------------|------------------|------|---------|-----------|
| 0                      | 89.80%          | 90.82%         | 79.27%           | 51.21% | 88.15%  | 79.85%    |
| 0.00001                | 90.01%          | 90.82%         | 79.17%           | 51.35% | 88.15%  | 79.89%    |
| 0.0001                 | 89.94%          | 90.82%         | 79.26%           | 51.68% | 81.53%  | 78.64%    |
| 0.001                  | 90.06%          | 90.44%         | 71.36%           | 51.82% | 81.53%  | 77.04%    |
| 0.003                  | 90.01%          | 90.55%         | 77.90%           | 56.03% | 81.39%  | 79.18%    |
| 0.005                  | 87.97%          | 90.14%         | 70.26%           | 54.21% | 81.53%  | 76.82%    |
| 0.01                   | 81.19%          | 89.78%         | 70.62%           | 52.60% | 81.53%  | 75.15%    |
| 0.015                  | 87.41%          | 89.32%         | 69.36%           | 53.59% | 81.52%  | 76.24%    |
| 0.02                   | 80.54%          | 88.69%         | 60.59%           | 52.27% | 81.82%  | 72.78%    |
| 0.03                   | 79.85%          | 81.42%         | 60.99%           | 49.09% | 81.67%  | 70.61%    |
| 0.05                   | 65.86%          | 78.92%         | 53.65%           | 47.35% | 80.69%  | 65.86%    |
| 0.1                    | 40.82%          | 59.29%         | 35.88%           | 38.83% | 63.45%  | 47.66%    |
| 0.15                   | 30.27%          | 30.98%         | 25.96%           | 31.60% | 62.11%  | 36.18%    |
| 0.2                    | 25.03%          | 14.92%         | 15.86%           | 14.96% | 52.27%  | 24.61%    |
Figure 3. Experiment 1: detection results of the test set under various Gaussian noises in original system.

Figure 4. Experiment 2: detection results of the same test set under new training set with various Gaussian noises after improvement.

Figure 5. Comparison of the detection result of a fixed-wing UAV: original system (a) and improved system (b).

Figure 6. Comparison of the detection result of a balloon: original system (a) and improved system (b).
Combined with table 3 and table 4, the change of the mAP value after the improved LSS-object detection system are shown in figure 8.

![Graph showing mAP value vs Gaussian noise intensity](image)

**Figure 7.** Comparison of the detection result of 5 birds: original system (a) and improved system (b)

Combined with table 3 and table 4, the change of the mAP value after the improved LSS-object detection system are shown in figure 8.

**Figure 8.** Experiment 2: comparing detection results of the test set under various Gaussian noises by original system and improved system.

It can be seen from figure 8 that applying the way to add noise into the original images to augment the training set can assist the detection system has better adaptability to the environment of Gaussian noise. The training data set constructed by original data and noisy data enhances the feature learning ability of the system. From table 3 and table 4, the mAP value of the improved detection system is only 1.22% lower than the original detection system when detect original test data without Gaussian noise, but when the variance of the Gaussian noise in the environment is not more than 0.03, the mAP value is above 70.61% which has been greatly improved to detect test data. The experiment results show that the improved LSS-object detection system can adapt to the interference generated by Gaussian noise in the environment in practical applications.

4. Conclusions

The above experiment results show that, based on the YOLO v3 framework, after adding a Gaussian noise extended dataset for mixed training, the improved object detection system can achieve higher object detection accuracy in noisy conditions. At present, in the research of detection for LSS-object images, the construction of datasets and the specific needs for actual object types is relatively
insufficient. The method of this paper has carried out a certain amount of dataset annotation and construction work in this field. In the future work, the dataset will be augmented in categories and quantities in combination with practical engineering applications, and the impact of complex environment noise will be further studied. It is expected that more classes of LSS-objects can be detected and identified while maintaining a high detection accuracy.

References
[1] Szegedy C, Toshev A and Erhan D 2013 Deep Neural Networks for Object Detection NIPS vol. 26 pp 2553-61.
[2] Girshick R, Donahue J, Darrell T and Malik J 2014 Rich feature hierarchies for accurate object detection and semantic segmentation IEEE CVPR pp 580-87.
[3] Girshick R, Donahue J, Darrell T and Malik J 2016 Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 38 pp 142-58.
[4] Jasper R, Uijlings R, Koen E, Sande A, Theo G, Arld W and Smeulders M 2013 Selective Search for Object Recognition International Journal of Computer Vision 104 pp 154-71.
[5] He K, Zhang X, Ren S and Sun J 2015 Spatial pyramid pooling in deep convolutional networks for visual recognition ECCV vol.37 pp 1904-16.
[6] Girshick R 2015 Fast R-CNN ICCV Preprint arXiv:1504.08083.
[7] Ren S, He K, Girshick R, Sun J 2015 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks NIPS vol.39 pp 91-99.
[8] Dai J, Yi L, He K, and Sun J 2016 R-FCN: Object Detection via Region-based Fully Convolutional Networks NIPS Preprint arXiv:1605.06409.
[9] Redmon J, Divvala S, Girshick R and Farhadi A 2015 You only look once: Unified, real-time object detection CVPR Preprint arXiv:1506.02649.
[10] Liu W, Anguelov D and Erhan D 2015 SSD: Single Shot MultiBox Detector Preprint arXiv:1512.02325.
[11] Redmon J and Farhadi 2016 A YOLO9000: Better, Faster, Stronger CVPR Preprint arXiv:1612.08242.
[12] Redmon J and Farhadi A 2018 YOLOv3: An Incremental Improvement CVPR Preprint arXiv:1804.02767.
[13] He K, Zhang X, Ren S and Sun J 2015 Deep Residual Learning for Image Recognition CVPR Preprint arXiv:1512.03385.
[14] Dollar P, Wojek C and Schiele B 2009 Pedestrian Detection: A Benchmark IEEE CVPR pp 304-11.
[15] Chen QL and Xue YQ 2000 Estimation of signal-noise-ratio from Data Acquired with OMIS Journal of Remote Sensing 4 pp 284—89
[16] Corner B R 2003 Noise estimation in remote sensing imagery using data masking International Journal of Remote Sensing 24 pp 689—702