Object Detection Method Based on Aerial Image Instance Segmentation by Unmanned Aerial System in the Framework of Decision Making System

Serhiy Kovbasiuk
Research Department
Zhytomyr Military Institute named after S. P. Korolov
Zhytomyr, Ukraine
klasik552008@gmail.com

Leonid Kanevskyy
Research Department
Zhytomyr Military Institute named after S. P. Korolov
Zhytomyr, Ukraine
leo10k10@ukr.net

Mykola Romanchuk
Research Department
Zhytomyr Military Institute named after S. P. Korolov
Zhytomyr, Ukraine
romannik@ukr.net

***Abstract*—The article analyses the capabilities of unmanned aerial system application in the framework of decision making in the crisis situations that require the object detection on aerial images acquired by the unmanned aerial system. To increase the operational capability and credibility of the automotive vehicles detection at the aerial images acquired by the unmanned aerial systems for more efficient use of acquired information in the framework of decision making support model Mask R-CNN was selected. This model is more appropriate for solving the problem of multiclass classification and object detection of small-size objects on the image. To improve this model, the article recommends using small-size anchors taking into account height-to-width aspect ratio according to greater amount of classes that along with test time augmentation usage enables to augment the mAP.**

Keywords—recognition, object detection, aerial images, instance segmentation, focal loss, unmanned aerial systems.

I. Introduction

The rapid growth of engineering industry in the XX and XXI centuries bears the substantial threats for contemporary and future generations. The problem is not only in the environmental contamination by harmful substances from the cars with internal combustion engines. The large numbers of cars on the urban roads and parking places in some crisis situations increase the risk of untimely threat localization.

Taking into account the existing problems [1, 2] (e.g., low operational capability of terrain evaluation after the destructive flood for optimum planning of relief operations – failure to take into account the approaching ways blockage by the cars when the fire fighting team arrives etc.) the requirements to the decision making support in crisis situation increase. To solve the task of operational capability increase for receiving the reliable data about the situation in places hard to access in urban areas the use of unmanned aerial systems (UAS) become more and more prospective with implementation of automated means for the object detection, recognition and classification above ground.

II. Problem Statement in General Terms

One of the main components affecting the operational capability and quality of decision making in the crisis situations is the visualization system (information display) and such information processing technologies. The relevance of UAS usage for such task solution is predetermined by possibilities of information acquisition about the objects above ground and about the Earth as underlying terrain. Besides, the prospects of UAS usage are monitoring improvement of simple and complex objects condition that happened to be in the thick of events, as well as operational capability increase of crisis situation development forecast and optimization of management decision agenda.

The contemporary visualization systems enable to represent great volumes of information from various sources including the UAS. In most cases such information does not contain the intermediate conclusions of monitoring work that complicates the course of events forecasting and management decision making. To solve such problem additionally in the automated mode the acquired information processing is carried out. One of such examples may be the quantitative and qualitative analysis of the cars on the urban roads and parking places to forecast the passing ability of the special-purpose machines of nonstandard size, saturation of parking places with the cars in the heart of crisis event where the need for optimal evacuation emerges or corresponding passing corridor creation.

The set problem solution using UAS requires the search and development of efficient (operative with substantial credibility) means of automotive vehicles detection at aerial images received from UAS.

The purpose of the article is evaluation of neural networks as a tool for operational capability and credibility of automotive vehicles detection improvement at the aerial images received from UAS for acquired information efficient use in the decision making support.

III. Review of the Latest Research and Published Materials

As of today a lot of systems have been developed that show the capabilities of artificial neural networks. The conducted analysis of such systems technologically arranged at neural networks usages showed a lot of advantages. It is stipulated by the neural network capabilities for adaptive learning, self-organization, generalization, making calculations in the real-time mode and failure resistance. It was determined that the main areas for neural network application is function approximation, associative memory, data compaction, recognition and classification, task optimization, management of complex processes and forecasting.

978-1-7281-2399-8/19/$31.00 ©2019 IEEE

For task solution next architectures used in task Object Detection were analyzed: R-CNN [3], Fast R-CNN [4], SSD [5], YOLO [6], Faster R-CNN [7], Mask R-CNN [8]. As a result, it was determined that SSD and YOLO belong to single-pass detectors and consequently the family of double-pass models R-CNN has the speed advantage. However, SSD using network VGG [9] are optimized such way that minor layers in the neural network may not generate sufficiently high functions to detect the minor objects. YOLOv1 uses the layers for fusion that also decreases the possibility to detect the minor objects. In the next YOLO versions as a result of architecture change the minor object detection accuracy is lower than Faster R-CNN. The main deficiency of R-CNN, Fast R-CNN is low speed of image processing; Faster R-CNN demonstrates good results during detection, however, the same way as the previous models uses the limitation boxes during the object detection image visualization. In case of close location, it may result in multiple coverage of the minor objects. Mask R-CNN uses an additional instance segmentation that does not affect the computational capabilities.

Among all the mentioned models the winner of the latest COCO Detection Challenge 2018 was the model based on Mask R-CNN. This model fits better the task solution of multiclass classification and small-size object detection at the image within the set task, however, to improve the accuracy of object location detection it is necessary to improve it.

IV. Core Material Summary

Regarding the architecture and development of model Mask R-CNN for solving the set task it is suggested performing the stages of dataset increase where it is envisaged to take into account the conditions of object photographing, set the model hyper specifications to detect the transport vehicles and additionally to perform the post-processing to augment the accuracy of object detection and classification.

The main problem for using the models of deep learning is insufficiency of object depiction during their shaping. Besides, it is difficult to achieve the class balancing during the set of images shaping. The object dimensions according to classes may change from 12 to 120 pixels. The object images made under various angles contain noises, minor occlusions, various illumination levels; objects of one class may have various colors and their different saturation and may vary in sizes and height-to-width aspect ratio. Some classes contain insufficient number of object images that leads to the problem of neural network relearning during exercises.

Selection of Mask R-CNN as the basic model for the task solution is based on next approaches which application and optimization enables to solve the task of object detection at aerial images and improve the performance of learnt model indicators. To provide substantial speed of incoming image processing and maintenance of minor object detection accuracy in the incoming Regional Proposal Network (RPN), it is suggested using 7x7 convolution filter. Such approach will enable to realize the opportunity of using the anchor set with dimensions [8, 16, 32, 64, 128, 256] and corresponding variation will enable to improve the adaptation of the anchor height-to-width ratio. Adding the branch for forecasting the segmentation masks at each region of interest (RoI) augmented by Faster R-CNN in parallel with existing branch for classification and regression of bounding box (Fig. 1).

The mask branch is a little FCN used for each RoI stipulating the segmentation mask in pixel-to-pixel manner. Mask R-CNN is simple to implement and it is trained taking into account Faster R-CNN that relieves the wide spectrum of the architecture flexible constructions. Besides, the mask branch adds only little overhead costs on computer system enabling it to perform fast experimenting. And application of RoIAlign in Mask R-CNN with harder localization parameters will enable to augment its accuracy two times (more precisely from 10% to 50%).

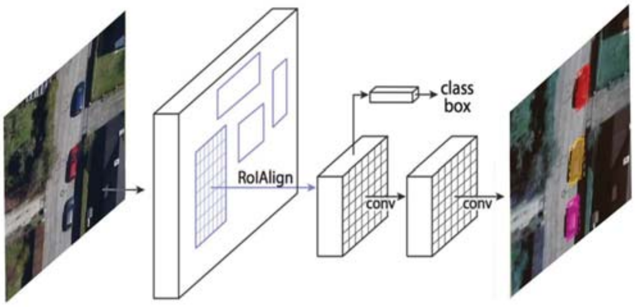

Fig. 1. Mask R-CNN for instance segmentation

During the exercise we determine the multi-task losses at each selected RoI is conducted with equation (1):

$$L = Lcls + Lbox + Lmask \tag{1}$$

where loss classification *Lcls* and limited loss box *Lbox* are identical to those that are determined in Faster R-CNN. The mask branch has Km^2 dimensional output for each RoI that codes *K* bit patterns with resolution $m \times m$ for one for each of *K* classes. For that we apply the sigmoid unit of pixels and determine *Lmask* as average loss of binary cross-entropy.

For RoI related with class ground-truth, *Lmask* is determined only at kth mask (other outputs do not aid the loss). Definition *Lmask* enables the network to create the masks for each class with no competition among the classes; the designers rely on assigned classification branch as opposed to generally accepted practice during FCN [10] application to semantic segmentation in order to envisage the class mark used for selection of the output mask. The mask allocation is performed in class-agnostic style: the masks are envisaged separately for each class without preliminary knowledge what exactly is depicted in the region, and then only the mask of class is selected that won in the independent classifier. As a result, the mask and forecast class during the sigmoid and pixel losses do not compete.

As a trunk architecture it was used the original performance Faster R-CNN with ResNets [11] on 50 or 100 layers that pulls the attributes from the final convolution layer of the 4th stage named C4 and which may be frozen and used for learning transfer.

V. Experimental Results

To research the suggested approaches and to have qualitative modeling process according to the set task it was used Dataset with Vehicle Detection in Aerial Images that contained 10 photos made by camera Canon EOS-1Ds Mark

III (focal distance is 51 mm) at height 1595-1600 m with resolution 5616 x 3744 pixels. As a result, the distribution of objects was formed 10 classes of vehicles distributed, shown in Fig. 2.

The results of research on the distribution of the area of objects on the aerial image are presented in Fig. 3 shows the quantitative majority of small-sized objects in comparison with large-sized vehicles.

In accordance with the ratio of geometric dimensions of vehicles, the main part lies in the range from 0.6 – 0.8 as shown in Fig. 4

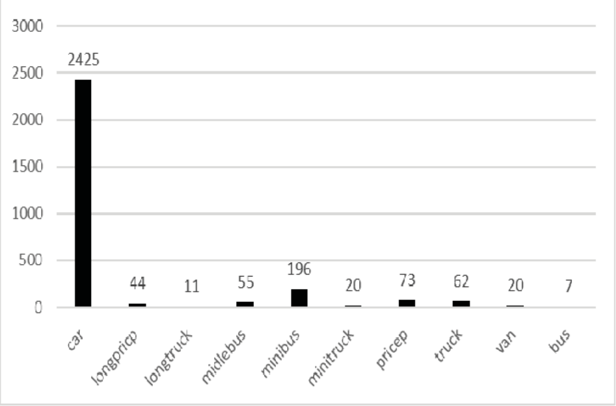

Fig. 2. Distribution of objects by class

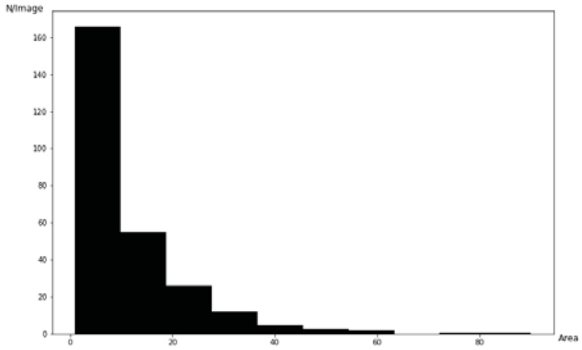

Fig. 3. Distribution by area of objects on an aerial photograph

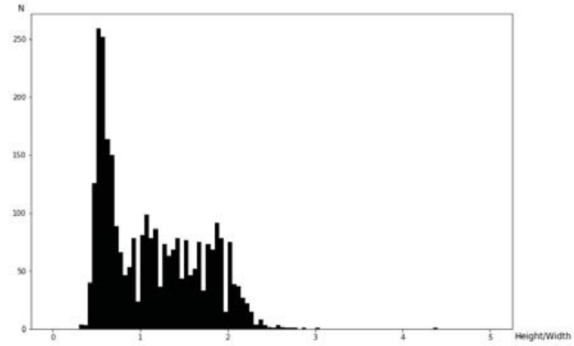

Fig. 4. Distribution of objects according to their geometrical dimensions

The incoming image was augmented two times for object mask augmentation, which improves the minor object detection quality on images.

The online augmentation for augmenting the object image set was used (D4 turns, adding Gaussian noise, contrast, sharpness and color saturation change). Transfer Learning approach was used through the trained models at datasets COCO [12], ImageNet [13], the weights of convolution layers were recorded and the model completed learning.

For evaluation of the work model we apply the metrics mean average precision (mAP) that calculates the average score value mAP for the variables intersection of unit (IoU*)* in order to fine a large number of bounding boxes with incorrect classifications and to avoid the maximum degree of specialization in several classes, due to weak performances in others.

RPN where 7x7 convolution filter is used has improved, for anchor changing to [4, 8, 16, 32, 64] creating anchor crossing as well as for taking into account the transport vehicle height-to-width ratio next Anchor ratio value was used [0.33 ;0.5; 1.0; 2.0; 3.0].

For uniting of all detection objects to exclude object detection duplication the post-processing algorithm non-maximum supression is used with value 0.7.

In training, light negatives can suppress learning and lead to degenerate models due to the imbalance of classes. When training the model, Focal Loss [14] was used to train RPN and Head (Faster R-CNN), which is designed to solve class imbalance by weight reduction (easy examples) in such a way that their contribution to the overall damage was small, even if their number is high. Instead of loss of cross-entropy $CE\ (p_t) = -log\ (p_t)$, focal loss was used, is conducted with equation (2):

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t),\ \gamma = 2 \qquad (2)$$

where FL – focal loss; CE – loss of cross-entropy; p_t – probability of ground truth class; γ – focusing parameter.

To finish the work of model we apply 3 TTA (negligible image compilation with resolution 600x650, 700x750 and turn (0, 90, 180, 270), 800x850 to 1000x1000). To take into account the place of image location at photo (minimizing the object influence with distorted signs at the image edge) it was used the probability reduction for adjustment factor 0.8.

The learning took place for Head – 15 epochs, backborn Resnet-50 with C4 – 10 epochs, for all layers – 15 epochs with *learning rate* = 0.001. The received results are represented in table 1.

TABLE I. DEPENDENCE OF ACCURACY OF MAP FROM CHANGE OF HYPERMETERS

Changes	**Mask R-CNN**						
anchor size change	No change (nc)	x	x	x	x	x	x
value of Anchor ratio		nc	x	x	x	x	x
NMS=0.7		nc	nc	x	x	x	x
3 TTA		nc	nc	nc	x	x	x
Probability reduction at the edge of photo		nc	nc	nc	nc	x	x
Focal Loss		nc	nc	nc	nc	nc	x
mAP (at IoU>=0.7), %	60.2	61.2	63.2	64.7	66	66.4	66.7

Due to model Mask R-CNN adjustment along with increase of image set and post-processing the mAP accuracy was increased 6.5%. It enables to increase the detection credibility of automotive vehicles at aerial images received from UAS. As far as this approach has little computing capability, it enables us to implement it on UAS board increasing the operational capability of received information usage within the framework of decision making support.

CONCLUSIONS

As a result of performed analysis of the advanced neural network models within the set task it was selected the best variant with corresponding architecture. From some of the performed research it was suggested performing the stages for increasing the datasets taking into account the conditions of object shooting, set the model hyper parameters for detecting the transport vehicles and additionally perform post processing which would enable to augment the localization accuracy and object classification. The direction for further research is focused on detection resolution improvement of minor object at the images.

Further research should be directed for the studies of FPN application in trunk architecture Mask R-CNN and Cascade R-CNN, ZN and usage of model assemblage to enlarge the possibilities for their application within the framework of decision making support. Additionally, it is necessary to perform research to enlarge the possibilities of UAS use in the complicated conditions of crisis situation and complicated space orientation.

REFERENCES

[1] New US Geological Survey-led research helps California coastal managers prioritize planning and mitigation efforts due to rising seas and storms. – Available at https://www.preventionweb.net/news/view/64251

[2] Alekseev V.O. Interactive monitoring of highways: monograph / V.O. Alekseev, O.P. Alekseev, A.A. Vidmish, V.O. Khabarov - Vinnitsa: VNTU, 2012. - 144 p.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[4] R. Girshick. Fast R-CNN. In ICCV, 2015.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[8] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask R-CNN. arXiv:1703.06870, 2017.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in ICLR, 2015.

[10] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in ECCV. Springer, 2014, pp. 740–755.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," IJCV, vol. 115, no. 3, pp. 211–252, 2015.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, pp. 2980–2988, 2017.