International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET

CNN Based Object Detection and Localization of Aerial Images

Prof. Neha Khare, Alok Rajpoot
Takshshila Institute of Engineering and Technology, Jabalpur, Madhya Pradesh, India

ABSTRACT

Aerial vehicles without human for instance drones, are by and large increasingly more embraced in observation and checking undertakings because of their adaptability and extraordinary versatility. They have a wide assortment of uses like following, observation, mapping, land studying and so forth. With the enhancements in CNN and an ever increasing number of models dependent on deep learning are being utilized to find the objects of enthusiasm for the pictures produced by unmanned airborne vehicles.A methodology for aerial pictures is implemented for object localization and detection which can improve execution of the model with less calculation cost. YOLOv3 utilizes Feature Pyramid Network (FPN) as a spine of the system yet in YOLOv4 PANet is utilized for boundary amassing from various layers of the element extraction model. Another import factor is the size of the item in the satellite picture is little, if there should be an occurrence of YOLOv4 PANet deal with this moreover. Along these lines, we executed YOLOv4 on satellite pictures for object discovery assignments. YOLOv4 design is applied on flying pictures for small object identification. We broke down the model on DOTA dataset. Results from YOLOv4 shows that it performs better than YOLOv3 and YOLOv2. It lessens the calculation cost with keeping up the precision of the expectation. With more than 89% precision, YOLOv4 is faster than YOLOv3 and YOLOv2.

Keywords : Object detection and localization, machine learning, neural network, CNN, YOLO, YOLOv3, YOLOv4.

I. INTRODUCTION

Many image and video degradation processes can be modeled as translation-invariant convolution. To restore these visual data, the inverse process, i.e., deconvolution, becomes a vital tool in

motion deblurring [1, 2, 3, 4], super-resolution [5, 6], and extended depth of field [7].

In applications involving images captured by cameras, outliers such as saturation, limited image boundary, noise, or compression artifacts are unavoidable. Previous research has shown that improperly handling these problems could raise a broad set of artifacts related to image content, which are very difficult to remove. So there was work dedicated to modeling and addressing each particular type of artifacts in non-blind deconvolution for suppressing ringing artifacts [8], removing noise [9], and dealing with saturated regions [9, 10]. These methods can be further refined by incorporating patch-level statistics [11] or other schemes [4]. Because each method has its own specialty as well as limitation, there is no solution yet to uniformly address all these issues.

One possibility to remove these artifacts is via employing generative models. However, these models are usually made upon strong assumptions, such as identical and independently distributed noise, which

may not hold for real images. This accounts for the fact that even advanced algorithms can be affected when the image blur properties are slightly changed.
In this paper, we initiate the procedure for natural image deconvolution not based on their physically or mathematically based characteristics. Instead, we show a new direction to build a data-driven system using image samples that can be easily produced from cameras or collected online.

We use the convolutional neural network (CNN) to learn the deconvolution operation without the need to know the cause of visual artifacts. We also do not rely on any pre-process to deblur the image, unlike previous learning based approaches [3]. In fact, it is non-trivial to find a proper network architecture for deconvolution. Previous de-noise neural network [7] cannot be directly adopted since deconvolution may involve many neighboring pixels and result in a very complex energy function with nonlinear degradation. This makes parameter learning quite challenging.

In our work, we bridge the gap between an empirically-determined convolutional neural network and existing approaches with generative models in the context of pseudo-inverse of deconvolution. It enables a practical system and, more importantly, provides an empirically effective strategy to initialize the weights in the network, which otherwise cannot be easily obtained in the conventional random-initialization training procedure. Experiments show that our system outperforms previous ones especially when the blurred input images are partially saturated.

II. RELATED WORK

Deconvolution was studied in different fields due to its fundamentality in image restoration. Most previous methods tackle the problem from a generative perspective assuming known image noise model and natural image gradients following certain distributions.

In the Richardson-Lucy method [7], image noise is assumed to follow a Poisson distribution. Wiener Deconvolution [8] imposes equivalent Gaussian assumption for both noise and image gra- dients. These early approaches suffer from overly smoothed edges and ringing artifacts.

Recent development on deconvolution shows that regularization terms with sparse image priors are important to preserve sharp edges and suppress artifacts. The sparse image priors follow heavy-tailed distributions, such as a Gaussian Mixture Model [1, 11] or a hyper-Laplacian [7, 3], which could be efficiently optimized using half-quadratic (HQ) splitting [3]. To capture image statistics with larger spatial support, the energy is further modeled within a Conditional Random Field (CRF) framework [1] and on image patches [11]. While the last step of HQ method is quadratic optimization, Schmidt et al. [4] showed that it is possible to directly train a Gaussian CRF from synthetic blur data.

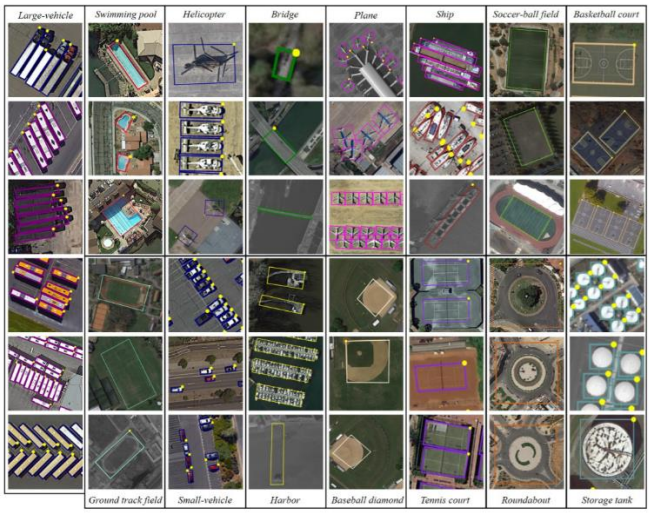
To handle outliers such as saturation, Cho et al. [9] used variational EM to exclude outlier regions from a Gaussian likelihood. Whyte et al. [1] introduced an auxiliary variable in the Richardson- Lucy method. An explicit denoise pass is added to deconvolution, where the denoise approach is carefully engineered [2] or trained from noisy data [1]. The generative approaches typically have difficulties to handle complex outliers that are not independent and identically distributed.

Another trend for image restoration is to leverage the deep neural network structure and big data to train the restoration function. The degradation is therefore no longer limited to one model regarding image noise. Burger et al. [8] showed that the plain multi-layer perceptrons can produce decent results and handle different types of noise. Xie et al. [1] showed that a stacked denoise autoen- coder (SDAE) structure [1] is a good choice for denoise and inpainting. Agostinelli et al. [5] generalized it by combining multiple SDAE

for handling different types of noise. In [2] and [6], the convolutional neural network (CNN) architecture [2] was used to handle strong noise such as raindrop and lens dirt. Schuler et al. [1] added MLPs to a direct deconvolution to remove artifacts. Though the network structure works well for denoise, it does not work similarly for deconvolution. How to adapt the architecture is the main problem to address in this paper.

III. Proposed Work and Result

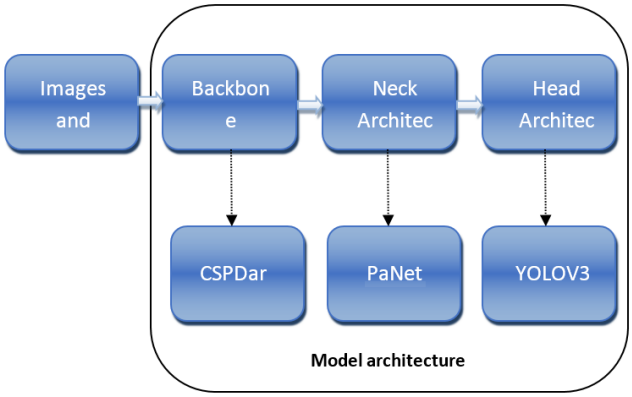
In this work, we are going to use a single-stage object detection scheme for object location in DOTA dataset. Dataset –DOTA[3] dataset is developed and introduced by Xia et al[3] from Cornell university. It may be a large-scale lackey object location dataset. It contains 2806 pictures each of measure 4000x4000 pixels in 15 categories: plane, baseball-diamond, bridge, ground-track-field, small-vehicle, large-vehicle, dispatch, tennis-court, basketball-court, storage-tank, soccer-ball-field, circuitous, harbor, swimming-pool, and helicopter. Fig appears illustration of pictures in each category.



YOLOv4 – YOLOv4 is presented by Alexey Bochkovskiy in April 2020 as an improvement in YOLOv3. It likewise chips away at the idea of passing pictures just a single time through the system for expectation that is the reason named as YOLOv4.

The vast majority of the accessible plan requires high calculation assets, for example, GPUs with enormous bunch size. Consequently the preparation turns out to be moderate. YOLOv4 conquer this issue with the assistance of Weighted-Residual-Connections (WRC), Cross-Stage-Partial-associations (CSP), Cross small scale Batch Normalization (CmBN), Self-antagonistic preparing (SAT) and Mish-actuation, Mosaic data augmentation, DropBlock regularization, and CIoU loss.

For object detection there exist single-stage and two-stage feature detection approaches. Two-stage approaches are accurate but slow when compared with single-stage method. YOLOv4 is a one-stage detector. It consists of Backbone, neck and dense predictor as shown if fig.



Block Diagram of Proposed Work

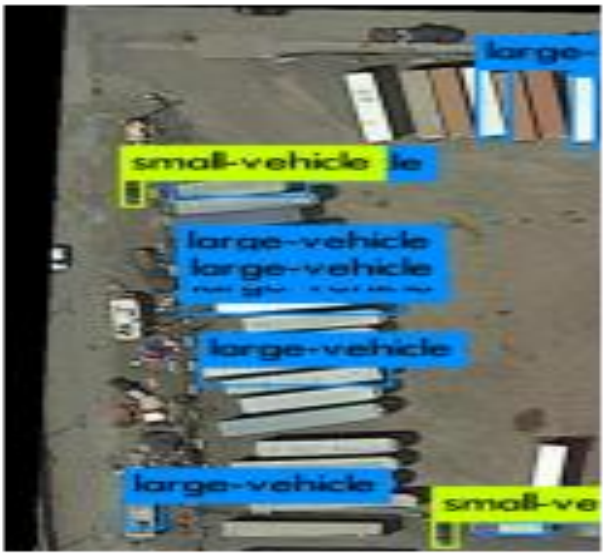
Backbone – various state-of-art image classification methods for imagenet dataset such as VGG16, ResNet50, Inception and DenseNet are used as backbone of the network for image feature extraction. In our case CSPDarkNet architecture with 53 convolutional layers is used in fig.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

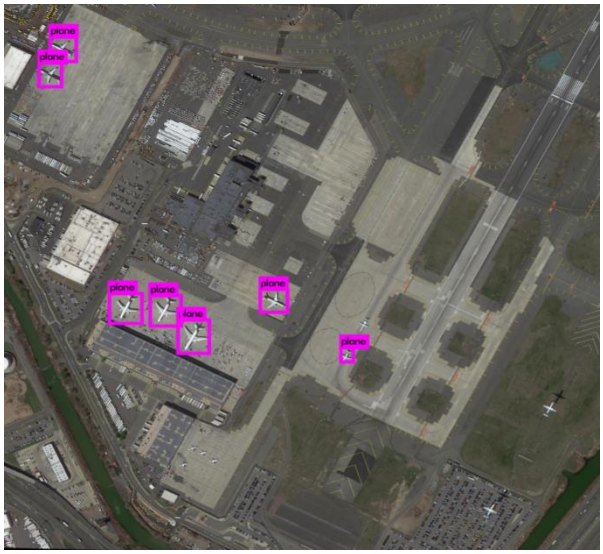
CSPDarkNet

Neck – These layers are present for feature extraction from different stages of backbone architecture. They can be Feature Pyramid Network(FPN), Bi-FPN, PANet. For Neck structure PANet is used which combines multiple layer output so that input information will not be loss.

Head – Task of head is classification and bounding box detection(regression). It usually predicts five values for single object; four for bounding boxes and one for class name prediction. For class and bounding box prediction YOLO V3 architecture is used.



Multiple instances of multiple objects.



Multiple instances of plane from aerial image

Table Precision comparison between YOLOv2, YOLOv3 and YOLOv4

Algorithm	Input Size	Precision	Recall
YOLOv2	512*512	77%	77%
Tiny YOLO V3	512*512	74.4%	81.5%
YOLO V3	512*512	88.47%	91.6%
YOLO V4	**512*512**	**89%**	**89%**

With YOLO V4 architecture we achieved mean average precision (mAP) as 89% and precision, recall and F1-score is 89%.

IV.CONCLUSION

In this work we applied YOLOv4 architecture on aerial images for small object detection. We analyzed the model on DOTA dataset. Results from YOLOv4 shows that it performs better than YOLOv3 and YOLOv2. It reduces the computation cost with maintaining the accuracy of the prediction. The system requirements for running YOLO model are quite high and it consumes a lot of GPU functionalities

to execute. The proposed version YOLOv4 plays a vital role in faster recognition and localization of the object. The schematic systems are able to rapidly create candidate bounding boxes, whereas the more complicated ones can anticipate objects more precisely. Test comes about illustrate that the proposed strategies accomplish the state-of the-art and are vigorous to distinctive question scales.

Further, we can enhance the work with dilated and anisotropic convolutional neural network which can enhance the performance by focusing on the variable shape or size of the object in the image.

V. REFERENCES

[1]. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. ACM Trans. Graph. 25(3) (2020)

[2]. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind decon- volution algorithms. In: CVPR. (2020)

[3]. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-laplacian priors. In: NIPS. (2019)

[4]. Schmidt, U., Rother, C., Nowozin, S., Jancsary, J., Roth, S.: Discriminative non-blind deblurring. In: CVPR. (2019)

[5]. Agrawal, A.K., Raskar, R.: Resolving objects at higher resolution from a single motion-blurred image. In: CVPR. (2020)

[6]. Michaeli, T., Irani, M.: Nonparametric blind super-resolution. In: ICCV. (2020)

[7]. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Trans. Graph. 26(3) (2020)

[8]. Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Progressive inter-scale and intra-scale non-blind image deconvolution. ACM Trans. Graph. 27(3) (2018)

[9]. Cho, S., Wang, J., Lee, S.: Handling outliers in non-blind image deconvolution. In: ICCV.(2019)

[10]. Whyte, O., Sivic, J., Zisserman, A.: Deblurring shaken and partially saturated images. In: ICCV Workshops. (2019).

[11]. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: ICCV. (2019).

Cite this article as :

Prof Neha Khare, Alok Rajpoot, "CNN Based Object Detection and Localization of Aerial Images", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 3, pp. 537-541, May-June 2020.

Journal URL : http://ijsrset.com/IJSRSET2073117