# Road Analytics

Yagnik Bhavsar[1] and Mehul S Raval[2]

***Abstract*— In recent years, unmanned aerial vehicle (UAV) has been increasingly applied to traffic monitoring field, driver behaviour analysis, road infrastructure analysis and , traffic flow analysis. Here, primary task is vehicle recognition. However, there are many challenges for vehicle detection in the UAV aerial video, such as camera shake, different camera orientation, different UAV altitude, interferential targets and a wide range of change of scene. Also, vehicle tracking is also useful for road analytic. In this project, we propose an end-to-end solution for vehicle tracking.**

***Index Terms*— UAV, ITS, deep learning, computer vision, object detection, object classification, VeRI 776, VisDrone 2019, Yolov5s, DeepSORT**

## I. INTRODUCTION

In today's world, unmanned aerial vehicles (UAVs) are very widely used in intelligent transport system(ITS), road structure analysis, traffic flow analysis and, enhanced road safety fields. In such fields, on road object detection and tracking are prominent steps. On-road objects differ in size, shape, color and also faces problems like occlusion, scale change. Also, there are several challenges related to UAVs itself such as camera shake, different camera orientation, different UAV altitude. That's why, it is important to have robust system to detect road objects. There are various machine learning and deep learning based algorithms used for it [1]. In order to detect and classify on-road objects light detection and ranging (LIDAR) and vision based method are available. A lidar based system uses pulsed laser to form 3D point cloud for objects. It uses laser which will work in every condition with high accuracy but as it is using laser, problems may arise when laser is reflected from dark objects and it uses very complex hardware, which is quite expensive. Vision based system uses camera, which can detect every object's color, size and, shape. Nowadays, low-price cameras available with high resolution which provides cheaper solution than the lidar. Also it can easily import image processing algorithms in use to solve the problem.

The project proposes vision based solution with help of deep learning algorithm. VisDrone 2019- a UAV based dataset is used for training and testing of detector and tracker models.

## II. RELATED WORK

There are various multi object trackers are available to find out trajectories of vehicles. They can use different object detectors for detection.

[1]Y. Bhavsar is PhD student at SEAS,Ahmedabad University.
[2]M. Raval is faculty at SEAS,Ahmedabad university.

| size | mAP@0.5 | Params | FLOPS |
|---|---|---|---|
| 640 | 56 | 7.2M | 16.5 |

TABLE I
YOLOV5S PRETRAINED MODEL PARAMETERS

- **Yolov5 Object detector** Alexey et el. [2] present yolov4 a CNN based network with : CSPDarknet53 as backbone, SPP, PAN as neck of the network and yolov3 as head. This is an extension of yolov3 model which provides optimal speed and accuracy. Yolov5 is just implementation of yolov4 in pytorch. Also pretrained models are available for COCO dataset which has 80 classes. The results for pretrained model yolov5s are as given in Table. I:
- **VisDrone 2018 dataset** Zhu et al. [3] present a VisDrone2018 dataset, with carefully annotated ground-truth for various important computer vision tasks, to make vision meets drones. Their dataset consists of 263 video clips with 179, 264 frames and additional 10, 209 static images. The videos/images are acquired by various drone platforms, i.e., DJI Mavic, Phantom series (3, 3A, 3SE, 3P, 4, 4A, 4P), including different scenarios across 14 different cites in China, i.e., Tianjin, Hongkong, Daqing, Ganzhou, Guangzhou, Jincang, Liuzhou, Nanjing, Shaoxing, Shenyang, Nanyang, Zhangjiakou, Suzhou and Xuzhou. The dataset covers various weather and lighting conditions, representing diverse scenarios in our daily life. The maximal resolutions of video clips and static images are 3840 × 2160 and 2000 × 1500 , respectively. VisDrone2018 benchmark and perform evaluation of the four tasks, i.e., (1) object detection in images, (2) object detection in videos, (3) single object tracking, and (4) multi-object tracking.
- **Alexnet** Alex et al. [4] present a CNN based architecture for object recognition. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, they used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers they employed a regularization method called "dropout" that proved to be very effective. The network takes between five and six days to train on two GTX 580 3GB GPUs. All of their experiments suggest that results can be improved simply by waiting for faster GPUs and bigger

datasets to become available.
One of the drawback of this model is that it has restricted the input image size of 224 x 224 only.

- **Network In Network** Min Lin et el. [5] propose a strategy called global average pooling to replace the traditional fully connected layers in CNN. The idea is to generate one feature map for each corresponding category of the classification task in the last conv layer. Instead of adding fully connected layers on top of the feature maps, we take the average of each feature map, and the resulting vector is fed directly into the softmax layer. One advantage of global average pooling over the fully connected layers is that it is more native to the convolution structure by enforcing correspondences between feature maps and categories. Thus the feature maps can be easily interpreted as categories confidence maps. Another advantage is that there is no parameter to optimize in the global average pooling thus overfitting is avoided at this layer. Furthermore, global average pooling sums out the spatial information, thus it is more robust to spatial translations of the input. We can see global average pooling as a structural regularizer that explicitly enforces feature maps to be confidence maps of concepts (categories).
- **DeepSORT** Wojke et el. [6] propose a deep neural based metric learning tracker based on SORT(simaple online and real time) tracker. Here, they place much of the computational complexity into an offline pre-training stage where they train a deep association metric on a large-scale person re-identification dataset. During online application, they establish measurement-to-track associations using nearest neighbor queries in visual appearance space. Experimental evaluation shows that their extensions reduce the number of identity switches by 45%, achieving overall competitive performance at high frame rates. This framework has two main techniques one is kalman filter and other is hungarian algorithm. Kalman filter is used to model motion of objects and Hungarian algorithm is used to associates objects of current frame to previous frames. This association is based on a kind of similarity metric such as IoU, cosine similarity.
- **Vehicle re-identification model based on VeRI 776 dataset** Liu et el. [7] present a vehicle re-identification model that is trained on VeRI 776 dataset for appearance matching. This trained model is further used with deepSORT framework for tracking vehicles.

## III. IMPLEMENTATION

All experiments are carried out on VisDrone 2019 dataset. VisDrone 2019 is slightly different from it's 2018 version but for our vehicle detection task category it is same. For mentioned dataset the distribution of sample images shown in section A, Alexnet with global average pooling layer is described in section B.

### A. Data set analysis

Distributions of sample images of VisDrone 2019 is shown in below figure 1. As we can see in figure 2, most of the samples have resolution below 50 x 50. We have six different classes named 'car, 'bus', 'bicycle', 'truck', 'van' and, 'motorcycle'. Given dataset's distribution is not uniform

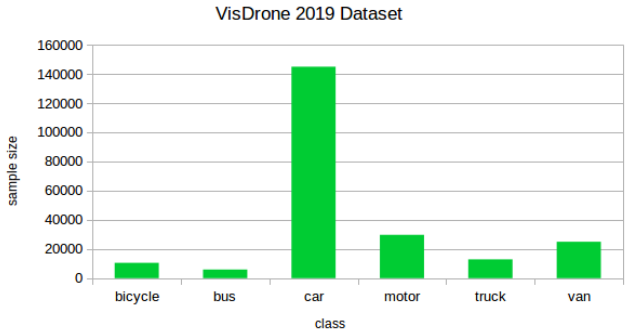


Fig. 1. VisDrone 2019 dataset

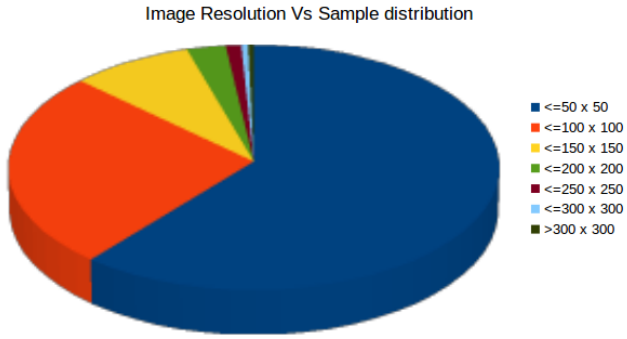


Fig. 2. VisDrone 2019 dataset Samples Vs Resolution

over all classes so, we are also providing weights for loss function to get better accuracy. Class weights are inversely proportional to their frequencies. VeRI 776 dataset is used to train a metric model of deepSORT. It contains a frames of objects that show how an object moves sequentially 3.

### B. Alexnet with GAP

Conventional Alexnet uses 224 x 224 input images but as we seen our sample images have variable resolution so, we have modified existing parameters such as size of filters and stride size. Summary of modified architecture for shown in Table II. "ReLU" activation is used at every layer except the output layer. "Softmax" is used at output layer. Instead of using fully connected layers at the end, we are using global max pooling which makes the architecture independent of input image size.

Fig. 3. VeRI 776 dataset

| Layer | Kernel Size/Neurons | Stride size |
|---|---|---|
| Input (conv1) | 11 x 11 x 96 | 4 x 4 |
| MaxPool1 | 2 x 2 x 96 | 2 x 2 |
| Conv2 | 5 x 5 x 256 | 1 x 1 |
| MaxPool2 | 2 x 2 x 256 | 2 x 2 |
| Conv3 | 3 x 3 x 384 | 1 x 1 |
| Conv4 | 3 x 3 x 384 | 1 x 1 |
| Conv5 | 3 x 3 x 256 | 1 x 1 |
| GlobalMaxPool | 1 x 1 x 256 | - |
| Output (FC3) | 6 | - |

TABLE II
SUMMARY OF ALEXNET+GAP MODEL

### C. Object detection with modified Alexnet

Here, we are using sliding window approach to generates regions that might contains our objects. We choose different sized windows 32 x 32, 64 x 64, and 128 x 128. The result shown in below figure 4. It is observed during testing that this mechanism giving us low bounding box accuracy and more number of false positives. So, we have to look for better solution such as Fast-RCNN, Faster-RCNN, SSD, yolo family detectors.

Fig. 4. Object detector using modified alexnet

### D. Yolov5s Object detector

We have seen that sliding window with alexnet classifier is not giving us required accuracy so, now we are looking at yolo family detector. We trained yolov5s detector on VisDrone data set. Below figures shows results and some parameters of yolo output.

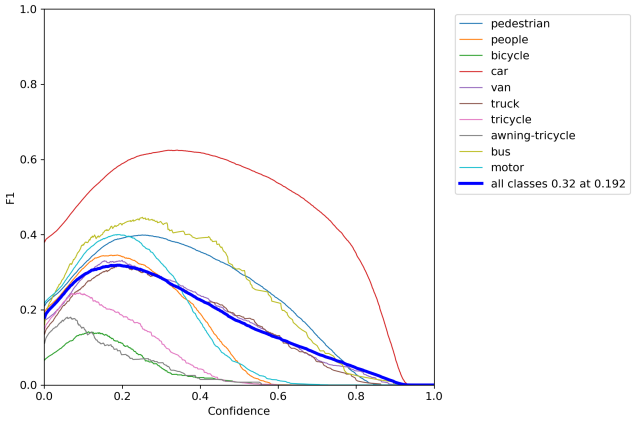


Fig. 5. Yolov5s F1 score

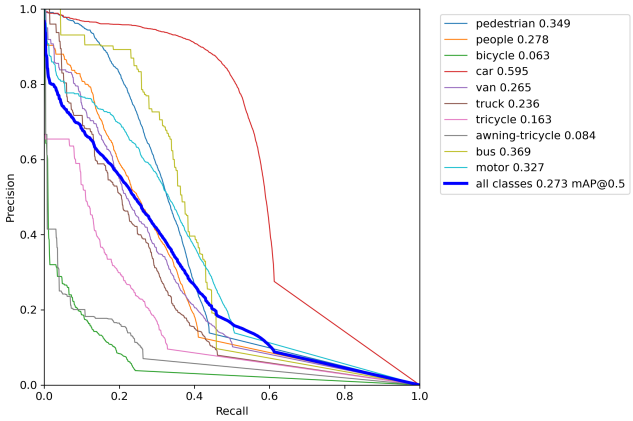


Fig. 6. Yolov5s PR curve

### E. DeepSORT object tracker with VeRID

Here, we have trained two vehicle re-identification models on VeRI 776 data set that helps to figure out similarity between objects.

*1) Resnet50 Model with IoU score:* This model is just matching the appearance of objects and to do association it uses IoU score of object. So, this model has very poor accuracy. This model can further improved if we add other similarity metric.

*2) Resnet Model with cosine score:* This model uses cosine similarity metric to compare object's appearance. Also, it uses IoU score thus it is providing better accuracy than previous one.

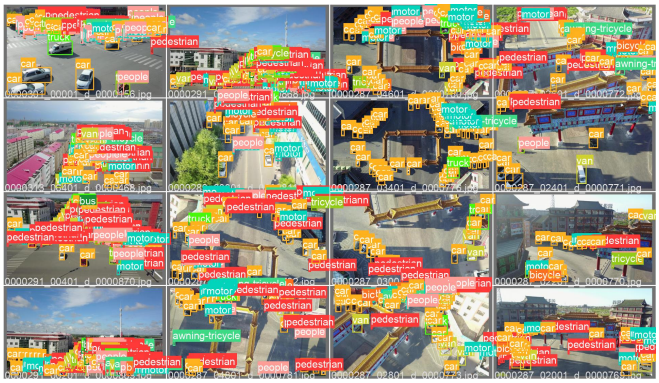

Fig. 7. Yolov5s Validation result

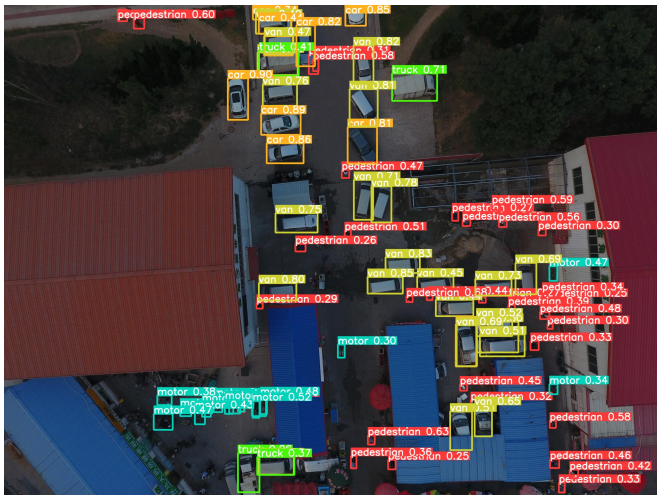

Fig. 8. Yolov5s Test result

## IV. RESULT

We have carried out experiments on VisDrone 2019 data set for above both the models. VeRID model with cosine metric is giving us very less ID switching during tracking. So, it will be useful in trajectory prediction.

## V. CONCLUSION

To get better accuracy we can further use Byte track algorithm which uses lower accuracy bounding boxes for tracking instead of removing it from predictions.

## REFERENCES


[1] M. Haris and A. Glowacz, "Road object detection: A comparative study of deep learning-based algorithms," *Electronics*, vol. 10, no. 16, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/16/1932

[2] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[3] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[5] M. Lin, Q. Chen, and S. Yan, "Network in network," 12 2013.

[6] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *CoRR*, vol. abs/1703.07402, 2017. [Online]. Available: http://arxiv.org/abs/1703.07402

[7] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *ECCV*, 2016.