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. To overcome these issues, a robust vehicle recognition system which is suitable for UAV aerial video is proposed under this project.**

***Index Terms*— UAV, ITS, deep learning, computer vision, object detection, object classification, Alexnet, SPP, GAP, VisDrone 2019, CNN, VAID, AU-AIR dataset**

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 recognition is prominent step. 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 model and convolution neural network(CNN) is used as the underlying model architecture.

II. RELATED WORK

There are different UAV based datasets available for on road vehicles and various deep learning algorithms used for vehicle recognition. Literature survey shows that mostly variants of CNN architecture are widely used for such task.

- **AU-AIR dataset** Ilker Bozcan et el. [2] present a multi-purpose aerial dataset that has multi-modal sensor data (i.e., visual, time, location, altitude, IMU, velocity) collected in real-world outdoor environments. The AU-AIR dataset includes meta-data for extracted frames (i.e., bounding box annotations for traffic- related object category) from recorded RGB videos. The videos are recorded at different flight altitudes from 5 meters to 30 meters and in different camera angles from 45 degrees to 90 degrees. The whole dataset includes 32,823 labeled video frames with object annotations and the corresponding flight data. Eight object categories are annotated including person, car, van, truck, motorbike, bike, bus, trailer. The total number of annotated instances is 132,034. The dataset is split into 30,000 training-validation samples and 2,823 test samples. They have used a quadrotor (Parrot Bebop 2) to capture the videos and record the flight data. An on-board camera has recorded the videos with a resolution of 1920 × 1080 pixels at 30 frames per second (fps). The sensor data have been recorded for every 20 milliseconds.
- **VAID dataset** Lin et al. [3] present a VAID (Vehicle Aerial Imaging from Drone) dataset. They collect about 6,000 aerial images under different illumination conditions and viewing angles from different places in Taiwan. The images are taken with the resolution of 1137 × 640 pixels in JPG format. Our VAID dataset contains seven classes of vehicles, namely 'sedan', 'minibus', 'truck', 'pickup truck', 'bus', 'cement truck' and 'trailer'. The images in the dataset are taken by a drone (DJI's Mavic Pro). To keep the sizes of the vehicles consistent in all images, the altitude of the drone is maintained at about 90 – 95 meters from the ground during video recording. The output resolution is 2720 × 1530 at 2.7K and the frame rate is about 23.98 fps. For an average sedan with the length of 5 meters and the width of 2.6 meters, the apparent size in the image is about 110 × 45 pixels. In the VAID dataset, the images are scaled to the resolution of 1137×640, and a sedan in the images is about the size of 40 × 20 pixels.
- **VisDrone 2018 dataset** Zhu et al. [4] 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 differ-

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

ent 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. [5] 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.
- **SPPnet** Zhang et el. [6] present a spatial pyramid pooling(SPP) in deep convolutional networks for visual recognition. As we seen in Alexnet a drawback is the restriction of input image size. Here, they introduced a pooling strategy, "spatial pyramid pooling", to eliminate the above requirement. The new network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. With these advantages, SPP-net should in general improve all CNN-based image classification methods. SPP layer is placed in between last convolution layer of neural network and first fully connected layer. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, SPPnet rank 2 in object detection and 3 in image classification among all 38 teams.
- **Network In Network** Min Lin et el. [7] 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).
- **Vehicle recognition Based on CNN-SVM** Karungaru et el. [8] present a improved Alexnet with SPP and SVM to effectively recognise vehicles. Here, CNN works as feature extractor and SVM plus dense layers as classifier. Overall architecture is shown in below figure 1.

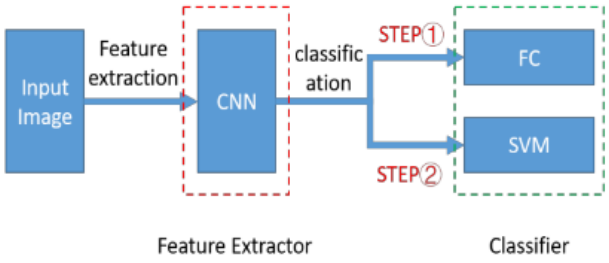

Fig. 1. Compound Network, [8]

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 all mentioned dataset the distribution of sample images shown in section A, modified architecture of CNN network(the alexnet) described in section B.

A. Data set analysis

Distributions of sample images of AU-AIR, VAID and, VisDrone 2019 are shown in below figures 2, 3, 4. As we can see in figure 5, 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 over all classes so, we are also providing weights for loss function to get better accuracy. Class weights are inversely proportional to their frequencies.

B. Architecture of CNN

Conventional Alexnet uses 224 x 224 input images but as we seen our sample images have resolution nearly 50 x 50 so, we have modified existing parameters such as size of filters and stride size. Summary of modified architecture for 50 x 50 images shown in Table I. "ReLU" activation is used at every layer except the output layer. "Softmax" is used at output layer. To remove the fixed input size restriction, we introduced global average pooling (GAP) before dense layer. With GAP we tried two variants of alexnet one is Alexnet + GAP and, second is CONV layers + GAP.

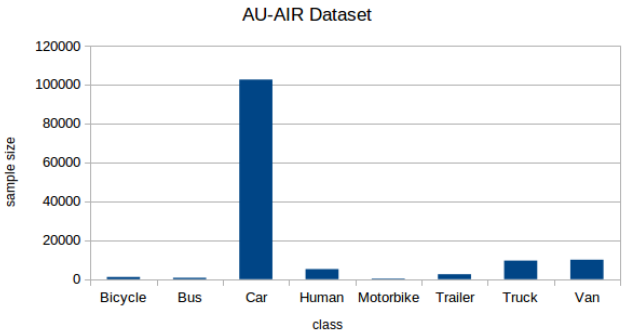

Fig. 2. AU-AIR dataset

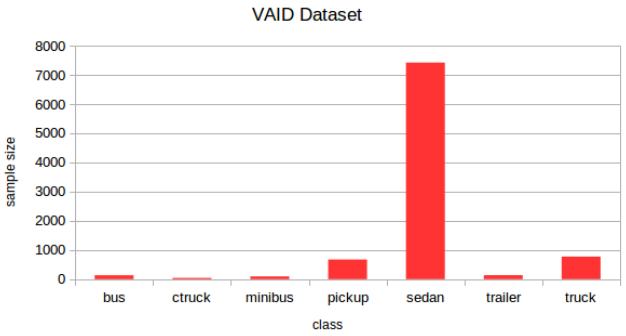

Fig. 3. VAID dataset

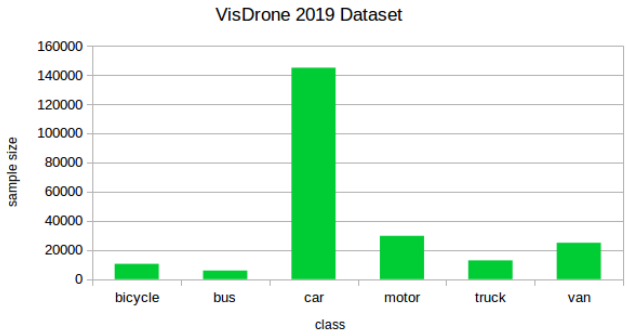

Fig. 4. VisDrone 2019 dataset

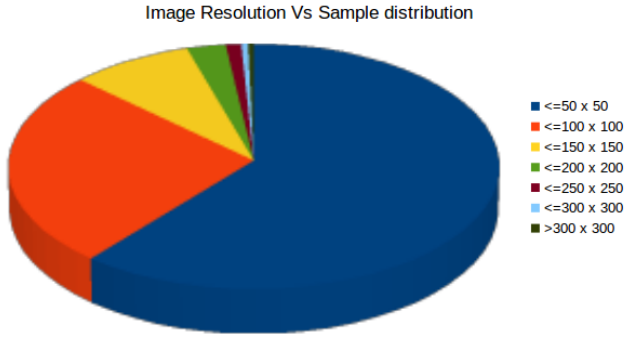

Fig. 5. VisDrone 2019 dataset Samples Vs Resolution

IV. RESULT

We have carried out experiments on VisDrone 2019 data set for different variants of alexnet. Outputs are summarised in below Table II:

V. CONCLUSION

After bunch of experiments, we have concluded that alexnet is sufficient for vehicle recognition. For further improvement, we can use R-CNN, YOLO, fast R-CNN and, Faster R-CNN.

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
MaxPool3	2 x 2 x 256	2 x 2
FC1	4096	-
FC2	4096	-
Output (FC3)	6	-

TABLE I

SUMMARY OF 50 X50 ALEXNET MODEL

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] I. Bozcan and E. Kayacan, "AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," *CoRR*, vol. abs/2001.11737, 2020. [Online]. Available: https://arxiv.org/abs/2001.11737

[3] H.-Y. Lin, K.-C. Tu, and C.-Y. Li, "Vaid: An aerial image dataset for vehicle detection and classification," *IEEE Access*, vol. 8, pp. 212 209–212 219, 2020.

[4] 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.

[5] 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

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, vol.

CNN	Val-Accuracy	Test-Accuracy
Alexnet (50 x50)	0.80	0.74
Alexnet + GAP	0.82	0.78
Conv + GAP	0.78	0.77

TABLE II

PERFORMANCE OF VARIANTS OF ALEXNET

abs/1406.4729, 2014. [Online]. Available: http://arxiv.org/abs/1406.4729

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

[8] S. Karungaru, L. Dongyang, and K. Terada, "Vehicle detection and type classification based on cnn-svm," *International Journal of Machine Learning and Computing*, vol. 11, pp. 304–310, 08 2021.