JADAVPUR UNIVERSITY

"Objects Tracking from video"

By,

Debasish Samanta

(Class Roll-001910503044)
(EXAM ROLL- MCA226043)

Under the supervision of

Prof. (Dr.) Jamuna Kanta Sing

Project submitted in partial fulfilment for the degree of
Master Of Computer Application

in the

Department of Computer Science & Engineering
FACULTY OF ENGINEERING AND TECHNOLOGY

2022

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

To Whom It May Concern

I hereby recommend that the project report entitled "Objects Tracking from video " has been carried out by **Debasish Samanta** (Reg. No .: 149904 of 2019 – 20 ,Roll No: 001910503044, **EXAM ROLL-MCA226043**, under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of **MASTER of COMPUTER APPLICATION** in **DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY** during the academic year 2021-2022.

Prof. (Dr .) Jamuna Kanta Sing
Project Supervisor
Dept. of Comp. Science & Engineering
Jadavpur University, Kolkata-700032

Prof. (Dr .) Anupam Sinha
Head of the Department
Dept. of Comp. Science & Engineering
Jadavpur University, Kolkata-700032

Prof.(Dr.) Chandan Mazumdar
Dean, Faculty Council of Engineering & Technology
Jadavpur University, Kolkata-700032

<u>CERTIFICATE OF APPROVAL</u>

This is to certify that the project report entitled "Objects Tracking from video" is a bonafide record of work carried out by **Debasish Samanta,** in fulfilment of the requirements for the award of the degree of ***Master of Computer Application*** in the ***Department of Computer Science and Engineering***, ***Jadavpur University***. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

....	....
Signature of Examiner 1	Signature of Examiner 2
Date:	Date:

ACKNOWLEDGEMENT

With my most sincere and gratitude, I would like to thank **Prof.(Dr.) Jamuna Kanta Sing**, ***Department of Computer Science & Engineering***, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestion and numerous discussions have always inspired new ways of thinking. I feel deeply honoured that I got this opportunity to work under him.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all the faculty members of the ***Department of Computer Science & Engineering*** **of** ***Jadavpur University*** for their continuous support.

Last, but not the least, I would like to thank my batch mates for staying by side when I need them.

DEBASISH SAMANTA
ROLL: 001910503044
(EXAM ROLL- MCA226043)

CONTENTS

Page no.

LIST OF FIGURE

LIST OF TABLES

ABSTRACT

Object tracking is the process via which computers are able to detect, understand, and keep an eye on objects across still images or videos. It is one of the most widespread applications of artificial intelligence (AI) and computer vision (CV). With object tracking solutions, we can perform meaningful actions on visual data obtained via different types of cameras. Using suitable object detection algorithms coupled with tracking models, we can train a machine to not just recognize one or more unique objects or persons in a particular image, but also identify them in subsequent frames and follow their trajectory in a video stream.

In this project report, our the aim is to track Objects across the frames using YOLOv3 and Simple Online Real Time Tracking (SORT) on traffic surveillance video. We usually locate the target by drawing the smallest rectangle possible (the "bounding box") in which it is included. Also in this project work efficient detection and tracking on vehicle dataset is witnessed. The algorithms give real-time, accurate, precise identifications suitable for real-time traffic applications.

Keywords

Artificial Intelligence (AI);

Computer Vision (CV);

Convolution Neural Network (CNN);

You Look Only Once (YOLOv3);

On rode trafic Dataset ;

Object detection; object tracking.

I. INTRODUCTION

Object detection and tracking is one of the areas of computer vision that is maturing very rapidly. It allows us to identify and locate objects in an image or video. With this kind of identification and localization, object detection and tracking can be used to count objects in a particular scene and determine and track their precise locations, all while accurately labeling them.

Object tracking is identifying trajectory or path; object takes in the concurrent frames. Image obtained from dataset is, collection of frames. Basic block diagram of object detection and tracking is shown in Fig. 1. Data set is divided into two parts. 75 % of images in dataset are used for training and 25 % for testing. Image is considered to find objects in it by using algorithms CNN and YOLO v3. A bounding box is formed across object with Intersection over union (IoU) > 0.5. Detected bounding box is sent as references for neural networks aiding them to perform Tracking. Bounded box is tracked in concurrent frames using Multi Object Tracking (MOT). Importance of this project work is used to estimate traffic density in road. We have organise this project report in such a manner where part II covers Fundamental Concepts of Object detection and Tracking. Part III describes design, implementation details and specifications. Part IV results and analysis. Part V describes conclusions and future scope.

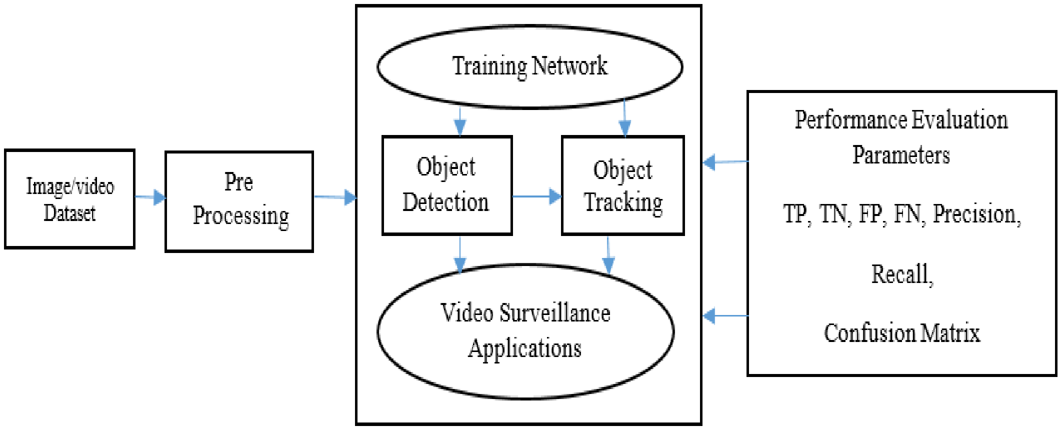


Fig. 1. Block Diagram of Object Detection and Tracking.

II. OBJECT DETECTION AND TRACKING ARCHITECTURE

We know that object detection is the process of identifying objects and finding its position .For an example we have shown various object detection tasks in Fig. 2. Classification + Localization and object detection method of identifying class of object and drawing a bounding box around it to make it distinct. The general intuition to perform the task is to apply CNN over the image. CNN works on image patches to carry out the task many such salient regions can be obtained by Region-Proposal Networks like Region Convolution Neural network (RCNN), Fast- Region Convolutional Neural

Network (Fast-RCNN), Faster- Region Convolutional Neural Network (Faster-RCNN). The efficient object detection algorithm is one which assures to give bounding box to all objects of vivid size to be recognized, with great computational capabilities, faster processing. Hence, selection of algorithm is application specific .

➢ ***<u>Convolutional Neural Networks (CNN)</u>***

We widely use Convolutional Neural Networks(CNN) architecture for computer vision related tasks. The advantage of CNN is that it automatically performs feature extraction on images i.e. important features are detected by the network itself. Here are some examples we have shown below.

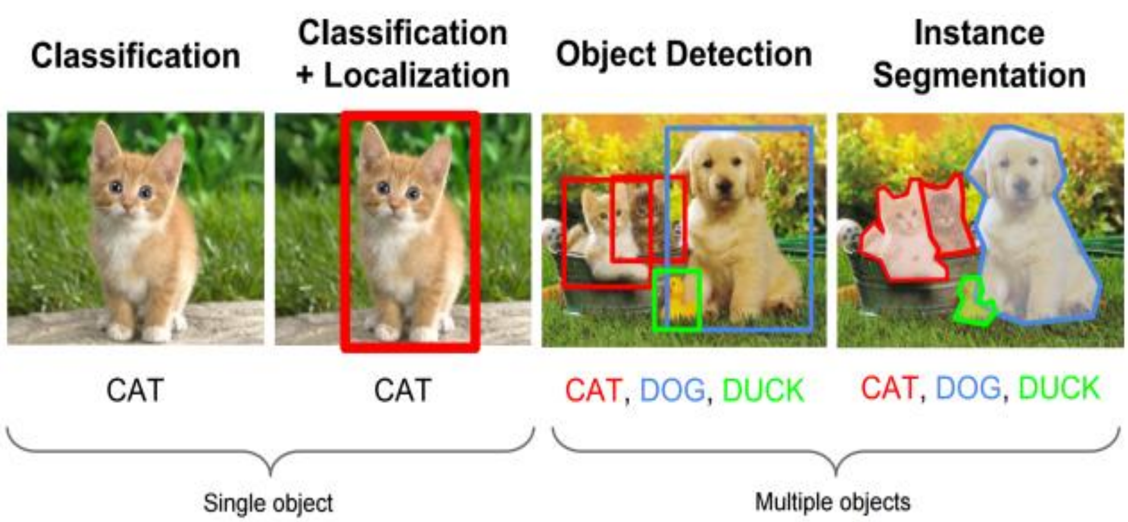


Fig. 2. Object Detection Tasks .

Convolutional Neural Network is consist of 3 important components called Convolutional Layer, Pooling layer, fully connected Layer as we have shown in Fig. 3. Considering a gray scale image of size 32*32 would have 1024 nodes in multi-layer approach.

This process of flattening pixels loses spatial positions of the image. Spatial relationship between picture elements is retained by learning internal feature representation using small squares of input data.

1) ***Convolutional layer:*** This layer encompasses filters and feature maps. Filters are the processors of a particular layer. These filters are distinct from one another and they take pixel value as input and gives out feature Map. Feature map is output of one filter layer. Filter is traversed all along the image, moving one pixel at a time. Activation of few neurons takes place resulting in a feature map.

2) ***Pooling layer:*** This layer is employed to reduce dimensionality. These are included after one or two convolutional layer to generalize features learnt from previous feature maps which helps in reducing chances of over fitting from training process.

3) ***Fully connected layer:*** This layer is used at the end to assign the feature to class probability after extracting and consolidating features from Convolutional Layer and pooling later respectively. Linear activation functions or softmax activation function are used by these layers.

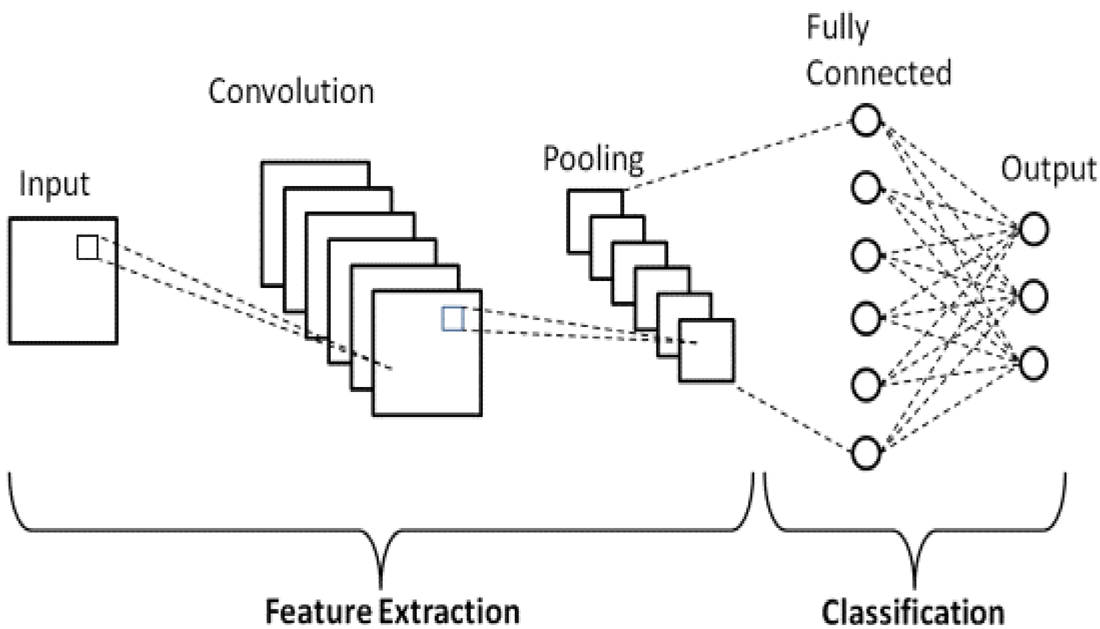


Fig. 3. Overview of CNN Architecture .

➢ *(YOLOv3)*

We know that YOLOv1 and v2 applies softmax functions to convert score into probabilities. This particular approach is feasible when objects are mutually exclusive only. YOLO v3 employs multi label classification. Independent logistic classifier is used to calculate likeliness of input belong to a specific label and loss is calculated using binary-cross entropy of each label . Here complexity is reduced since we remove the softmax function

Optimization of Bounding Boxes:

Since we use logistic, regression YOLO v3 we can predict the score of presence of object. A ground truth box is defined to all objects, if anchor box overlaps the most with ground truth box

then objectness score is said to be 1. For the anchor boxes whose overlap is greater than the preselected threshold, the anchor box incurs null cost. Every ground truth box is mapped with only one anchor box. If anchor box is not selected and assigned to bounding box then no classification and localization loss is considered therefore only confidence loss is calculated.
The anchor box is regressed to the ground truth box by gradual optimization as we have shown in Fig. 4. Coordinate parameters are now defined as

$$b_x = c_x + \sigma(t_x) \quad (1)$$

$$b_y = c_y + \sigma(t_y) \quad (2)$$

$$b_w = p_w e^{t_w} \quad (3)$$

$$b_h = p_h e^{t_h} \quad (4)$$

Where ,tx,ty,tw,th, are the predictions made by YOLO .cx,cy is top left corner of grid cell of the anchor. pw ,ph are the width and height of anchor. bx,by,bw,bh are predicted boundary box. σ (to) is box confidence score.

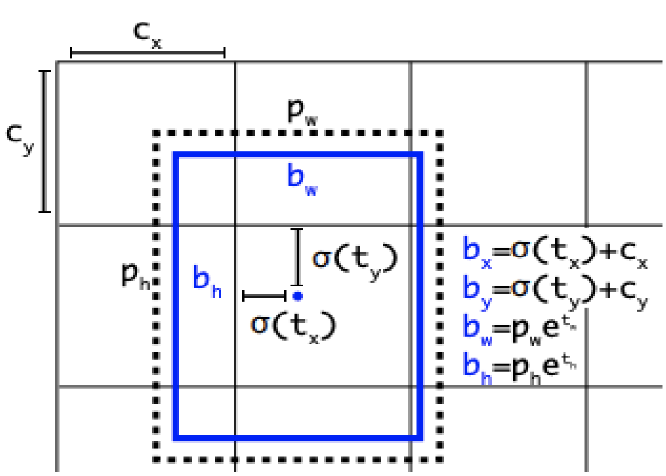


Fig. 4. Anchor Box Regression .

Predictions are made at 3 different scales which we have shown in Fig. 5.

The algorithm faces issues in identifying small objects or objects that are too close. Solution to challenge is increasing image resolutions. YOLO family upgrades its accuracy, latency. The inclusion of Feature Pyramid network helps in detecting objects that are small. FPN uses both bottom-down and a top-down pathway. Bottom-up approach is used for feature extraction. As we propagate through this approach, spatial resolution minimizes. Semantic value for each layer increases.

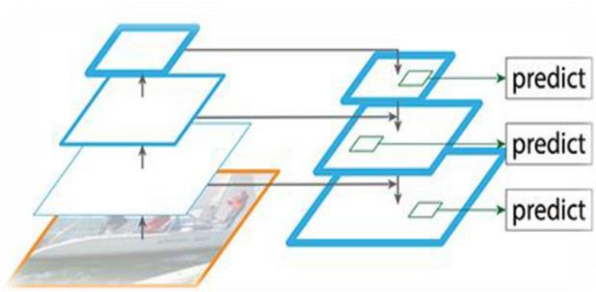


Fig. 5. Feature Pyramid Network .

➤ ***Object Tracking***

Video is collection of frames. The negligible time gap between frames makes the stream of photos looks like flow of scenes. When designing algorithm for video processing. Videos are classified into two classes. Video stream is an ongoing process for video analysis. The processor is not aware of future frames. Video sequence is video of fixed length. All the consecutive frames are obtained prior to processing of current frame. Motion is distinct factor that differentiates video form frame. Object properties and action can be realized by noticing circle in centre points of the image.

➢ ***Simple Online Real Time Tracking (SORT)***

Simple Online Real Time Tracking is a realistic approach to achieve Multi Object Tracking (MOT). Performance of SORT is enhanced by ques such as appearance; this association of appearance to SORT enhances the performance of SORT and increases performance during Scenario like longer periods of occlusion. SORT is a framework that has Kalman filtering has its crux.

❖ ***Track Handling and state estimation:***

The assignment problem maps prediction of Kalman filter to that of newly arrived measurements. The task of associating two vectors is performed by Hungarian algorithm. Adding additional information like motion and appearance parameters in conjunction with association helps in better mappings.

$$d^{(1)}(i,j) = (d_j - y_j)^T s_i^{-1}(d_j - y_i) \tag{5}$$

Unlikely association can be removed by thresholding at 95% confidence interval. The decision is given with an indicator.

$$b_{i,j}^{(1)} = 1[d^{(1)}(i,j)] \leq t^{(1)} \tag{6}$$

When the motion uncertainty is large mahalanobis distance is not suitable, hence another metric to aid in association. Metric computes appearance descriptor for each bounding box detection d_j.

$$d^{(2)}(i,j) = min\left\{1 - r_i{}^T r_k^{(i)} \middle| r_k^{(i)} \in R_I\right\} \tag{7}$$

Combination of both metrics is

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j) \tag{8}$$

III. DESIGN AND IMPLEMENTATION

We designed CNN which we have trained on Vehicle dataset, which is an Indigenous dataset for traffic surveillance applications.
We shown some of the sample images of dataset in Fig. 6. Also we have tabulated hardware and software requirements in Table I and Table II.

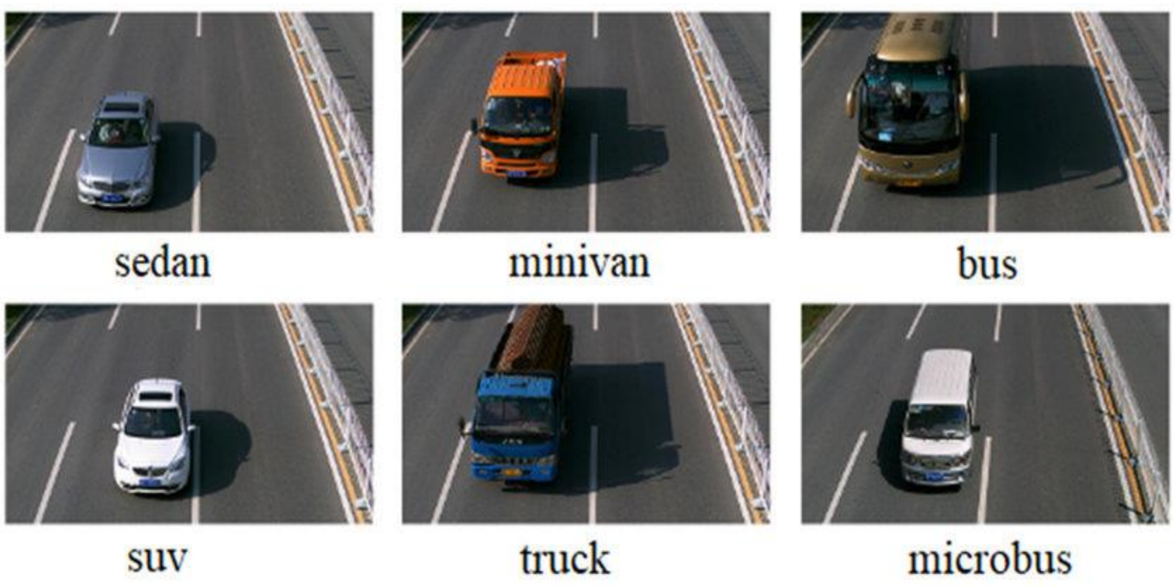


Fig. 6. Sample Images of Vehicle Dataset .

TABLE. I. HARDWARE REQUIREMENTS

Processor	Intel® Core™ i3 Processors
Clock Speed	1.8 GHz
RAM	8 GB
Storage	500 GB SSD
GPU	Intel® HD Graphics 520

TABLE. II. SOFTWARE REQUIREMENTS

Library	Tensor Flow,OpenCV
Packages	Numpy,pandas
Language	python
IDE	PyCharm
Applications	LabelImg,TensorBoard

➢ ***Neural Network Training Task***

We have shown flowchart of neural network training in Fig. 7. In first step in training a network using deep learning for an application is to prepare an appropriate dataset and make train-test Split depending on the available data. Suitable network is designed for training and validation Loss is monitored throughout the training process to produce a very less constant value after few epochs, if not then the hyper parameter tuning is performed on model to give lowest possible validation loss values. Model with best validation loss is saved and tested on real world dataset. The model is said to be good if a descent precision and recall values are obtained for new datasets else the model needs to be trained on enhanced dataset to increase performance.

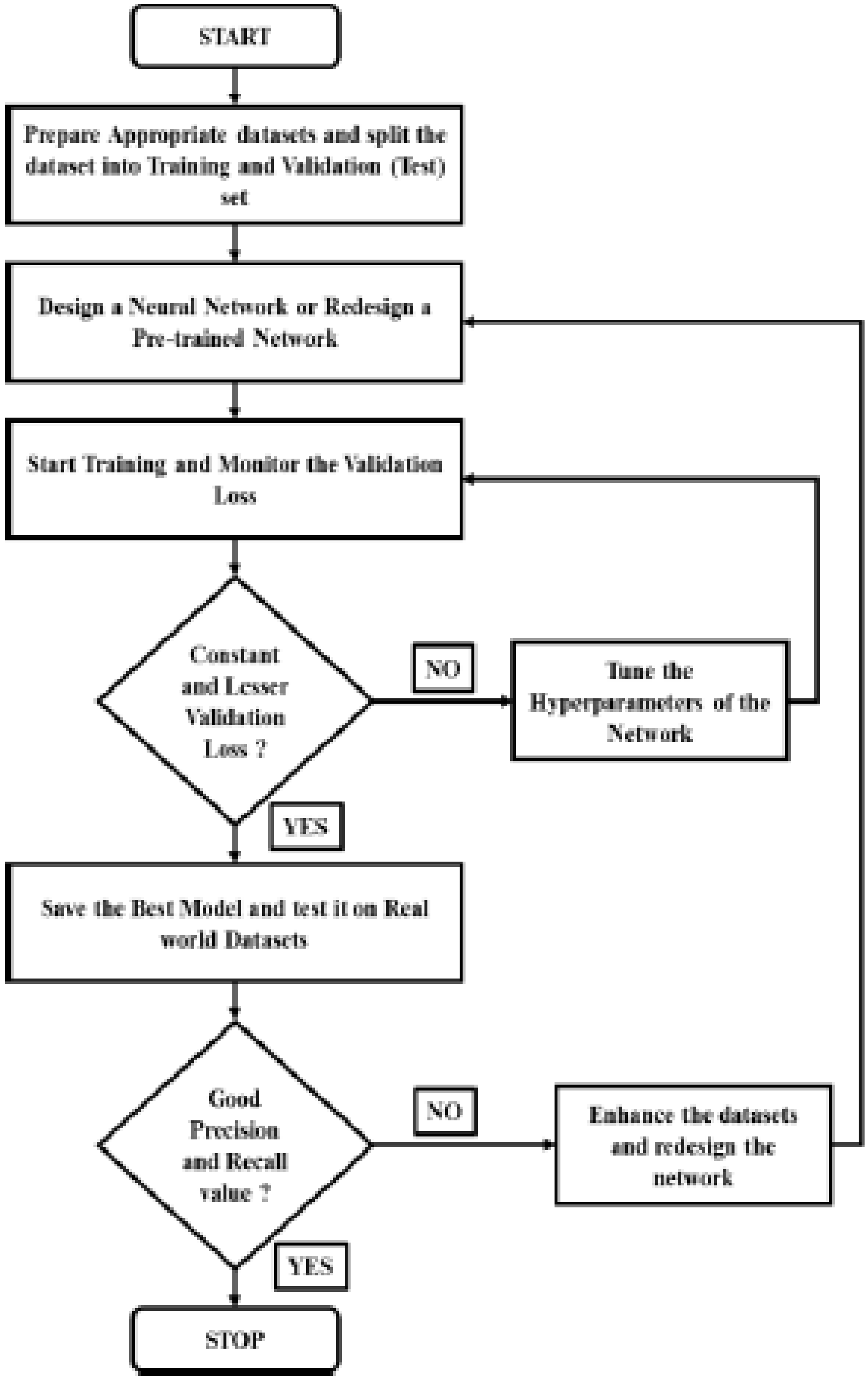


Fig. 7. Flowchart of Neural Network Training.

➢ ***Single Object Detection Task***

We have shown flow chart of single object detection in Fig. 8. Necessary libraries are imported first and we give training data as input and here we useTensor-flow algorithms . This algorithm then compiles data.

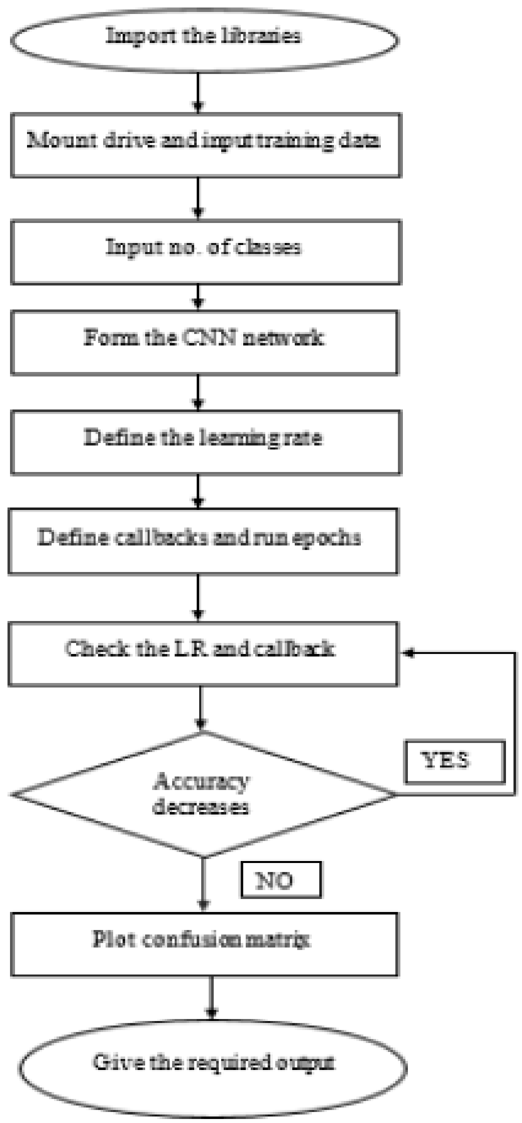


Fig. 8. Flowchart of Single Object Detection.

This algorithm can be described as supervised classification algorithm. Data flows through CNN layers and various operations are performed on data. The learning rate and callbacks are defined. Number of epochs and batch size is also defined. The epochs are then executed through which algorithm learns through training data. Training accuracy and training losses are constantly monitored. If training accuracy starts falling below a threshold, the callback function is invoked and epochs are stopped. Confusion matrix is then plotted using training and testing data. Various performance parameters can be defined and observed using the confusion matrix.

➢ ***Multiple Object Detection Task***

We have described working of YOLOv3 multiple object detection algorithm in Fig. 9 . Here an image is given as the input to algorithm and transformation is done using CNN. These transformations are done so that, input image is compatible to specifications of algorithm. Following this, flattening operation is performed. Flattening is converting data into a 1-dimensional array for inputting it to next layer. Flattening of output of convolutional layers is to create a single long feature vector and it is connected to final classification model, which is called a fully connected layer.

➢ ***Multiple Object Tracking Task***

In multiple object tracking, we have trained the vehicle tracker using YOLO v3 and deep learning methods and we optimize the detector's success rate by providing efficient vehicle detection results by testing trained vehicle detector on dataset . It consists of six phases such as loading data set, YOLO v3 design, training options configuration, object tracker training, and tracker evaluation, respectively. We have shown Flow chart of multiple object detection in Fig. 10.

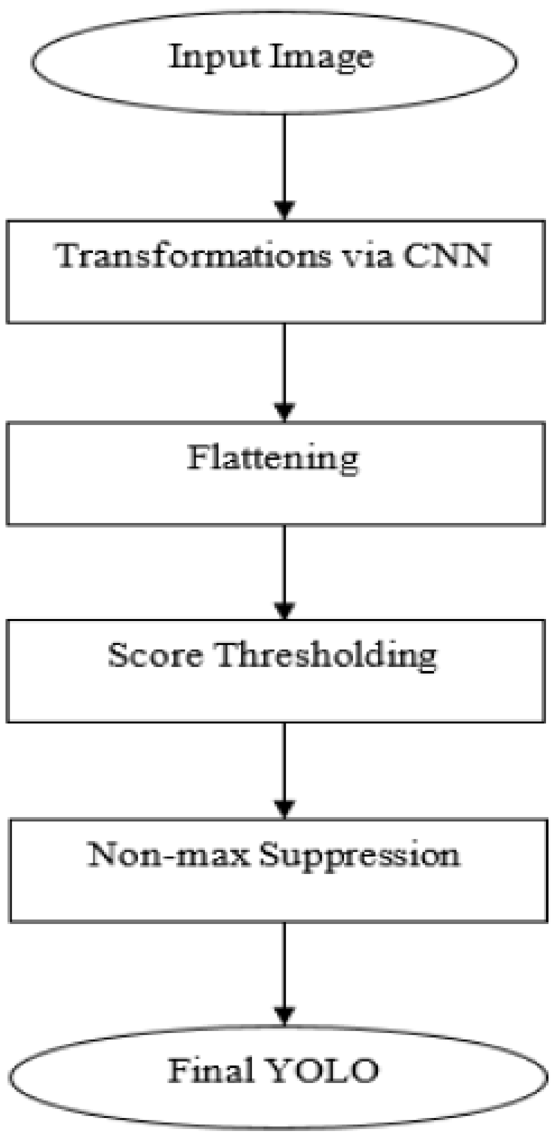


Fig. 9. Flowchart for Multiple Object Detection.

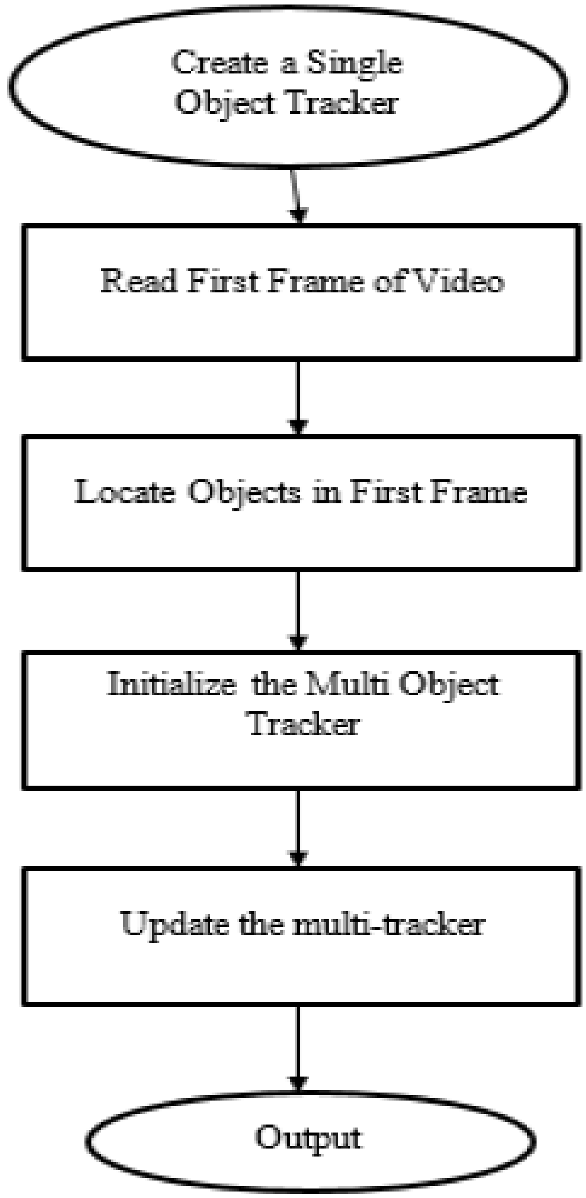


Fig. 10. Flow Chart of Multiple Object Tracking.

1) Confusion Matrix: It gives prediction information of various objects for binary classification which we have shown in Table III.

2) Accuracy and Loss: We have calculated accuracy measure by using formula (TP +TN)/(TP+TN+FP+FN). The accuracy measure, as a stand-alone measure is not reliable since it gives equal costs for both type of errors and works well for a well-balanced dataset. The loss is calculated by loss functions of used for training, and average of the loss is calculated when used batch learning that computes loss after each training each batch.

3) Precision, Recall and F1- score: Precision is the percentage of classification results that are relevant. Recall is the percentage of total relevant results that are classified correctly by algorithm. F-1 score considers both precision and recall values hence must be maximized to make the model better.

TABLE. III. CONFUSION MATRIX

	Predicted Class-1	**Predicted Class-2**
Actual Class-1	**TP - True Positive** Decision is correct	**FN - False Negative** Error – Type 1
Actual Class-2	**FP - False Positive** Error – Type 2	**TN - True Negative** Decision is correct

By these formula we calculate these metrics,which are shown below

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F - 1\, Score = 2 * \frac{Preciison*Recall}{Precision+Recall} \quad (11)$$

$$mAP = \frac{1}{No.of\ divisions} \sum_{r€(1,0.1,0.001)} p_{interp}(r) \quad (12)$$

Here the detected objects are bounded with bounding box. Tracking is performed on frames of the videos to identify objects in the successive frames using SORT. The evaluation metrics like True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) thereby Precession, Recall and hence mean Average Precession (mAP) is by us calculated using Intersection over Union (IoU).

IV. RESULTS AND ANALYSIS

In this part we have described simulation results and performance parameters and observed accuracy, precision and recall. This part underlines the confusion matrices of different datasets and convolution layers of the algorithms.

- ***Single Object Detection***

CNN is designed for single object detection and we have designed neural network and also trained and tested this. We have shown otained training accuracy and loss in Fig. 11.

```
Use tf.where in 2.0, which has the same broadcast rule as np.where
Epoch 1/10
 - 17s - loss: 7.0876 - acc: 0.4023
Epoch 2/10
 - 11s - loss: 0.7619 - acc: 0.6733
Epoch 3/10
 - 10s - loss: 0.6503 - acc: 0.7470
Epoch 4/10
 - 10s - loss: 0.5791 - acc: 0.7896
Epoch 5/10
 - 11s - loss: 0.5761 - acc: 0.7955
Epoch 6/10
 - 11s - loss: 0.5496 - acc: 0.8031
Epoch 7/10
 - 10s - loss: 0.5387 - acc: 0.8126
Epoch 8/10
 - 10s - loss: 0.5242 - acc: 0.8219
Epoch 9/10
 - 10s - loss: 0.5359 - acc: 0.8173
Epoch 10/10
 - 10s - loss: 0.5501 - acc: 0.8152
```

Fig.11. Accuracy and Loss

This encloses the parameters which are included in each step, layer progression and output image size of each layer. Each layer divides the image matrix into its components and performs an operation on image. The output image size of various layers is different due to manipulations by each layer such as initially the output image size is 28×28 which then reduces to 14×14 due to the max pooling layer which chooses the max valued pixel from the surrounding pixels. It then reduces to 7×7 due to the second max pooling layer. This pixel is then flattened into 7×7×64 that are 3136 sized vector. Hence this vector is reduced to a less sized vector by proceeding layers and displayed final calculation parameters.

Our designed neural network was trained and tested and we Obtained training accuracy and loss that we have shown in Figure 12 and 13. We obtained 82% training accuracy through training this model. Here the loss and accuracy are inversely proportional to each other. In this case when epochs is run, the model trains it itself and weights of the CNN gets updated to a more accurate value.

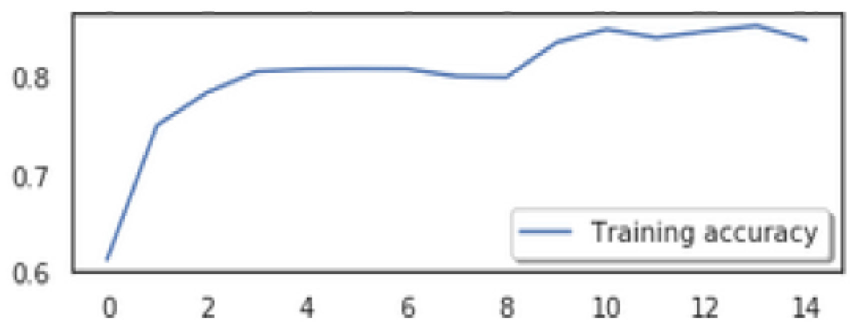


Fig. 12. Training Accuracy.

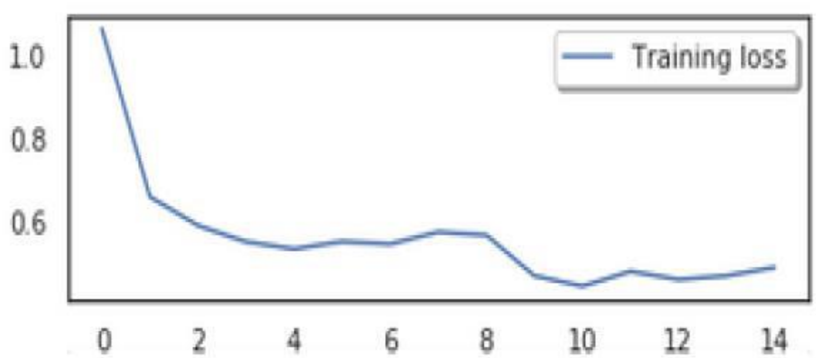


Fig. 13. Training Loss.

➢ *Multiple Object Detection:*

Here in Fig. 14 we shows that algorithm can detect objects of any size and images captured from various camera angle and distance. This attribute is because of FPN used in YOLO v3. Also intact bounding boxes as we have shown in Fig. 14 which ensures minute details and a greater IoU.

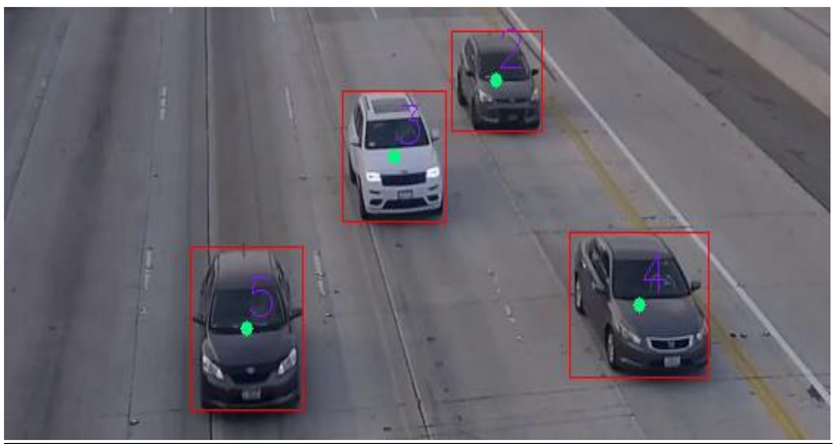


Fig. 14. YOLOv3 Results with Intact Boxes.

➢ *Multiple Object Tracking for Road Trafic Video*

Name – rode_trafic
Format - mp4
Size - 7.48Mb
Time frame - 245 sec
Video Quality – 720p
Types of objects – Cars , trucks,two wheeler bike(very less)

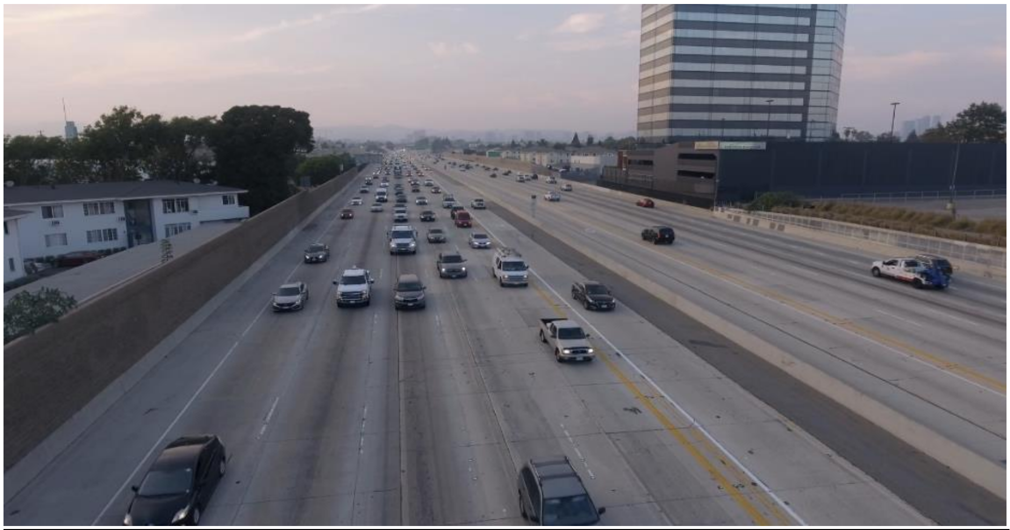


Fig. 15. Multiple Object Tracking Snippet.

RESULTS **(Multiple objects tracking at different frames)**

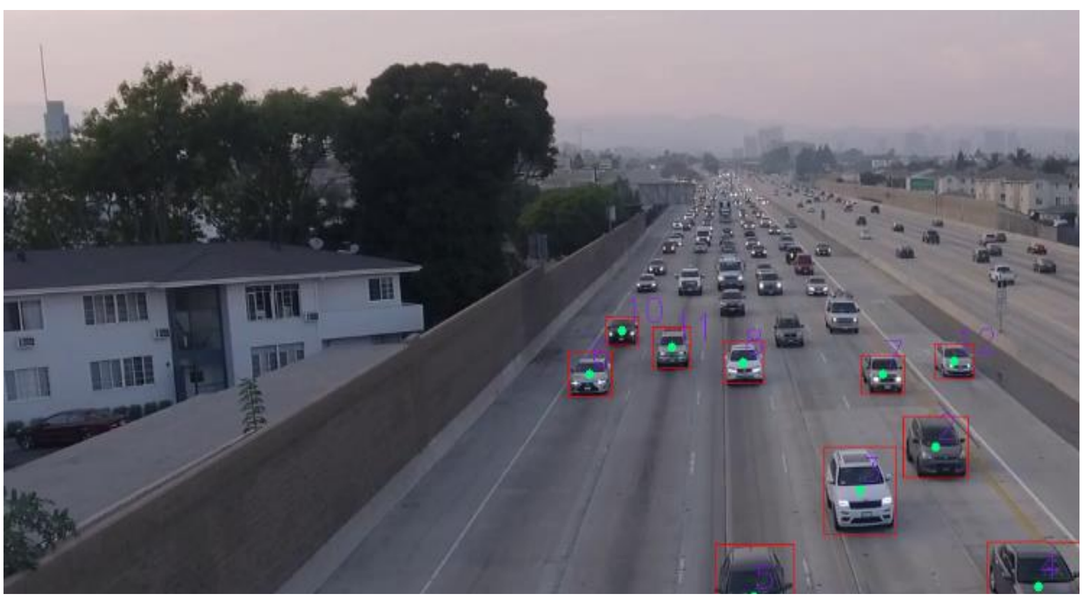
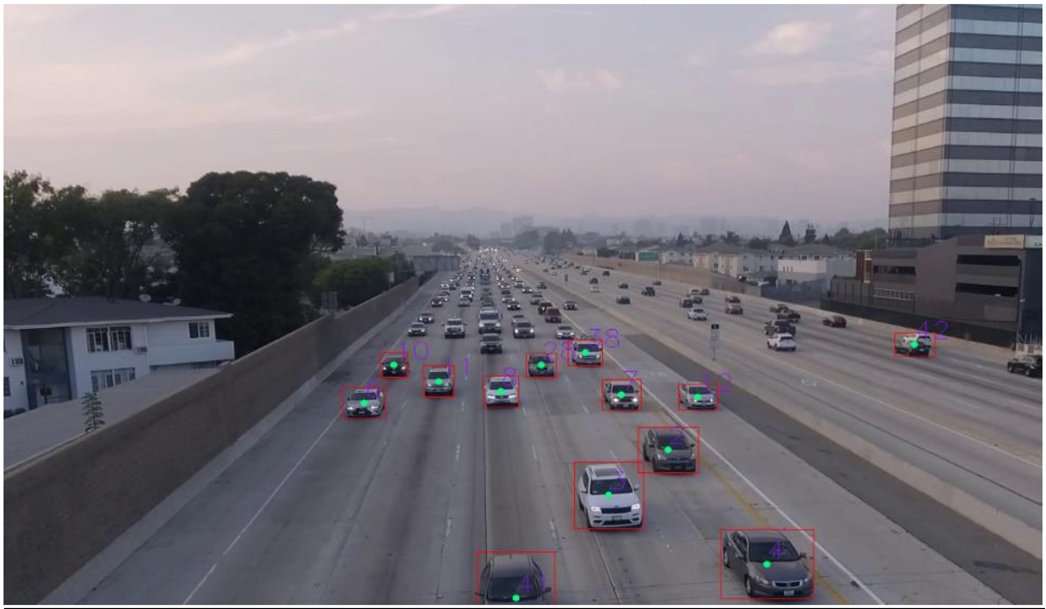
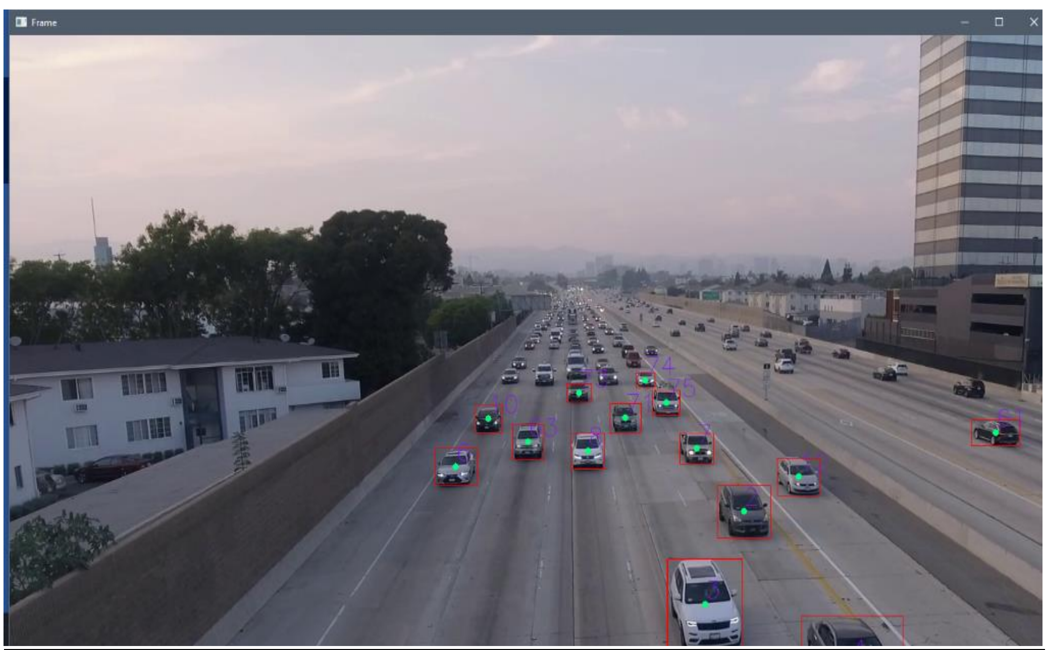


Fig. 16. Frame where objects are detected and tracked.

<u>(Other frames)</u>





In Fig. 15 we have shown the images of multiple object tracking for surveillance video, which contains cars and trucks. The vehicle tracker trained on surveillance video using YOLO V3 and deep learning methods. Here vehicle tracking process was successfully carried out by testing trained vehicle detector on test data set video. Our algorithm divided this video into frames with a rate of 25fps and performed object detection in first frame and in the latter frames, the particular detected image is tracked using its centroid position. We tracked objects in different frames at different intervals of time as shown in above figures.

Challenges in Object Tracking

While solving the object tracking problem, there arises a number of issues which can lead to a poor outcome .Properly detecting objects can be a particularly challenging task, especially since objects can have rather complicated structures and may change in shape, size, location and orientation over subsequent video frames. Various algorithms and schemes have been introduced in the few Introduction decades, that can track objects in a particular video sequence, and each algorithm has their own advantages and drawbacks. Any object tracking algorithm will contain errors which will eventually cause a drift from the object of interest. The better algorithms should be able to minimize this drift such that the tracker is accurate over the time frame of the application. In object tracking the important challenge that has to consider while the operating a video tracker are when the background is appear which is similar to interested object or another object which are present in the scene.

V. CONCLUSIONS

The inclusion of Artificial Intelligence to solve Computer vision tasks has outperformed the image processing approaches of handling the tasks. The CNN model trained to on road traffic dataset for single object detection.

Multiple object detection is implemented using YOLO v3 on dataset. Performance metrics is tabulated for YOLO v3 on considered classes of images. IoU of 0.5 is ideal for detection and tracking. Results of performance metrics is totally dependent on image data set used. Further objects are detected in video based on region of interest. Multiple object tracking is implemented for road traffic video using YOLOv3 and OpenCV. Multiple objects are detected and tracked on different frames of a video. We can use powerful GPUs to make the model more accurate and suitable for real-time applications.

Future Work

In the future, we can extend the work to detect the moving object with non-static background, having multiple cameras which we can also be used in real time surveillance applications.

REFERENCES

[1] V. D. Nguyen et all., "Learning Framework for Robust Obstacle Detection, Recognition, and Tracking", IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 6, pp. 1633-1646, June 2017.

[2] Zahraa Kain et all, "Detecting Abnormal Events in University Areas", 2018 International conference on Computer and Applications(ICCA),pp. 260-264, 2018.

[3] P. Wang et all., "Detection of unwanted traffic congestion based on existing surveillance system using in freeway via a CNN-architecture trafficnet", IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, 2018, pp. 1134-1139
.
[4] Q. Mu, Y. Wei, Y. Liu and Z. Li, "The Research of Target Tracking Algorithm Based on an Improved PCANet", 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2018, pp. 195-199.

[5] K. Muhammad et all., "Convolutional Neural Networks Based Fire Detection in Surveillance Videos", IEEE Access, vol. 6, pp. 18174-18183, 2018.

[6] D. E. Hernandez et all., "Cell Tracking with Deep Learning and the Viterbi Algorithm", International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS), Nagoya, 2018, pp. 1-6.

[7] H. S. Bharadwaj, S. Biswas and K. R. Ramakrishnan. "A large scale dataset for classification of vehicles in urban traffic scenes", Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing,ACM,2016.

[8] Mohana et.al., "Performance Evaluattion of background modeling methods for object Detection and Tracking",International Conference on Inventive systems and Control (ICISC).

[9] G. Chandan et.al., "Real Time Object Detection and Tracking Using Deep Learning and OpenCV", International Conference on Inventive Research in Computing Applications (ICIRCA), 2018.

[10] A. Raghunandan et. al., "Object Detection Algorithms for Video Surveillance Applications," International Conference on Communication and Signal Processing (ICCSP), 2018.