An Efficient Yolov7 and Deep Sort are Used in a Deep Learning Model for Tracking Vehicle and Detection

Vinod Kumar Yadav[*], Dr. Pritaj Yadav[**], Dr. Shailja Sharma[**]

[*] Research Scholar, Computer Science & Engineering Department, RNTU, Bhopal, M.P., India
[**] Associate Professor, Computer Science & Engineering Department, RNTU, Bhopal, M.P., India
[***] Associate Professor, Computer Science & Engineering Department, RNTU, Bhopal, M.P., India

Abstract- For tracking and detecting vehicles, use computer vision techniques. It is essential to traffic accident detection and intelligent transportation systems. On the highway an essential component of traffic surveillance is the detection, identification, and counting of vehicles. It takes a lot of effort to create a traffic monitoring model that performs well. Artificial intelligence-based traditional vehicle detection systems have weak detecting capability and robustness. A deep learning model for vehicle detection, tracking, and counting is proposed in this paper and is based on an efficient Yolov7 single shot detector and Deep- Sort of Multi Object Tracking algorithms. The suggested model examines the automobile detection algorithms and suggests proposed detection models using moving vehicle footage as survey data. When observed under a range of circumstances, such as high traffic, nighttime, many vehicles overlapping, and part of the vehicle missing, the suggested identification system exhibits excellent adaptability. The algorithm can accurately detect and identify automobiles based on their edge outlines, according to experimental data. YOLOv7-DeepSORT performs higher in tracking accuracy after experimental evaluation as compared to the earlier YOLOv5-DeepSORT.

Index Terms- Computer vision, deep learning, Deep-sort, Yolo, Yolov5, Yolov7 and MOT.

I. INTRODUCTION

Road When it comes to intelligent mobility, one of the most significant concerns in the development of smart cities is road safety [1]. All road users must be able to move through intersections, highways, and roads quickly, safely, and accident-free because of intelligent traffic systems. Intelligent traffic systems, which frequently have traffic video surveillance systems installed, have a particular infrastructure that is influenced by technology. Finding effective ways to predict accidents before they happen is one of the hottest subjects in road safety. Emerging technologies for mimicking the human visual system include computer vision and deep learning. For the issue of accident prevention in traffic safety contexts, a number of solutions based on computer vision and deep learning have been developed [2]. Vehicle type recognition is a vital tool for making traffic surveillance videos [20]. A non-intrusive, low-cost method to count vehicles is a software-based system that uses a video camera and computer vision algorithms. The ability to replace hardware-based systems has also been shown to be greatly enhanced by recent developments in object detection and tracking technologies and rising computing power. Generally, object tracking refers to the recognition and detection of several targets in the video, such as people, vehicles, animals, etc., without knowing the specific number of targets. In order to perform further trajectory prediction, accurate search, and other operations, various targets have separate IDs. In addition to the difficulties associated with Multiple object tracking [3], such as occlusion, deformation, motion blur, crowded environments, fast motion, changes in illumination, scale, and so forth, vehicle tracking also has to deal with more complicated issues like trajectory initialization and termination, mutual interference between similar targets, and so forth. Target detection performance has increased dramatically in recent years because of the rapid growth of deep learning and the idea of detection-based tracking has also come into being [23]. It has swiftly established itself as the foundation for multi-object tracking in use today, considerably advancing multi-object tracking activities. To track and identify the vehicles in the video frame, a convolution neural network technique based Yolov7 deep learning is use to detect vehicles and create IDs of the detected vehicles, after that frame by frame tracking vehicle by deep sort tracker in the proposed model.

II. RELATED WORK

Three categories can be used to classify the present Multi Object Tracking (MOT) framework: MOT based on joint detection and tracking [24], MOT based on attention mechanism [25], and MOT based on tracking by detection (DBT) [4]. The DBT framework's methodology begins with identifying the targets in each frame of the video sequence, cutting the targets in accordance with the bounding box, and obtaining all of the targets in the image. It therefore becomes the issue of target correlation between the front and back frames. The IoU, appearance feature, and other methods are used to generate the similarity matrix, and the Hungarian and greedy algorithms are used to solve it. This type of algorithm's object detection network performance determines how well it tracks objects. The YOLO series network is currently the most popular detection network. The YOLO algorithm, developed by Redmon et al. in 2015, uses the entire image as its input and gets the bounding box

and detection result right away [5]. This technique is quicker than existing algorithms at 45 frames per second frame detection. Instead of moving the window, the YOLO algorithm breaks the original image into small, non-conforming squares, deforms those squares, and then builds a map of those-sized objects. According to the analysis above, each feature in the feature map is likewise a little square that corresponds to the original image [6]. The target of those central points in the little square can then be predicted by using each element. When compared to YOLOv1, YOLOv2 has three key advantages: batch averaging; employing high-resolution photos, improve the categorization model [7]; the use of an a priori box In order to acquire a priori box dimensions, YOLO2 first used K-means clustering. YOLO3 continues this process by setting three priority fields for each down sampling scale, yielding a total of nine a priori sizes the box is clustered in. The most notable enhancements of YOLO3 are: alterations to the network's architecture; for object detection, multilevel function is utilized [8]. Soft-max is replaced by object classification, which uses logistic classification [9]. The YOLOv4 research introduced the top-down PAN feature fusion approach [10]. Before adding the image tensor to the backbone, a model with the majority of its parameters in YOLOv5 employs the Focus module. Focus: Subnetwork Sampling; PAN: Top-to-bottom function fusion; SPP: function fusion [19]. The two tracking algorithms for Multi Object Tracking that are most concerned with safety are SORT [11] and DeepSORT [12]. Kalman filter and Hungarian matching are features of SORT. The Kalman filter is used to anticipate the target's position [13], and Hungarian matching [17] is used to compare the prediction results of object detection networks like YOLO with those of the Kalman filter. But in reality, there are a lot of identity flips in the algorithm because of the changing target motion and frequent occlusion. as a result, the proposed DeepSORT with improved efficiency by adding cascade matching and additional functions on top of it. MOT, which combines detection and tracking framework, is based on detection and tracking [14]. This type of algorithm typically finds the two nearby frames in the video and employs a variety of techniques to assess how similar the targets are in the two frames that are being tracked and predicted.

III. METHODOLOGY

A. Architecture of yolov7

The real-time object detection model for computer vision tasks with the highest accuracy and speed is YOLOv7. In general, YOLOv7 offers a quicker and more robust network architecture that offers a better feature integration approach, more precise object detection performance, a more robust loss function, and an improved label assignment and model training efficiency. The amount of parameters and computational density of a model are the two main considerations for extended efficient layer aggregation networks is a backbone of yolov7. The input/output channel ratio and element-wise operation have an impact on the speed of network inference; according to the VovNet and CSPVNet models (CNN aims to make DenseNet more efficient by integrating all features just once in the last feature map). E-ELAN, which YOLO v7 dubbed after extending ELAN, The main benefit of ELAN (Efficient Layer Aggregation Networks) was the ability to better learn and a deeper network by managing the gradient path. A compound model scaling strategy can be used to better optimize the YOLOv7. For concatenation-based models, width and depth are scaled in this case coherently. After training, re-parameterization is a method used to enhance the model. Although the training duration is extended, the inference outcomes are enhanced. To complete models, two types of re-parameterization are used: model level and module level ensemble. These two methods can both be used to re-parameterize models at the model level. Train several models with the same parameters and different training data. To get the final model, then, take the average of their weights. A model's weights at many epochs are averaged. Re-parameterization at the module level has become very popular in research recently. The model training procedure is divided into several phases using this approach. The final model is created by ensemulating the outputs. The architecture of re-parameterized convolution in YOLOv7 makes use of RepConv [15] without identity connection (RepConvN). The goal is to prevent identity connections when re-parameterized convolution is used to replace a convolution layer with residual or concatenation. The lead head in YOLOv7 is referred to as the final output head. And the Auxiliary Head is the head that helps with middle-layer training. Label Assigner is a method that assigns soft labels after taking the ground truth and network prediction outcomes into account. The YOLOv7 network's Lead Head makes predictions about the outcome. These final results are used to generate soft labels. The crucial aspect is that the identical soft labels that are generated are used to calculate the loss for both the lead head and the auxiliary head.

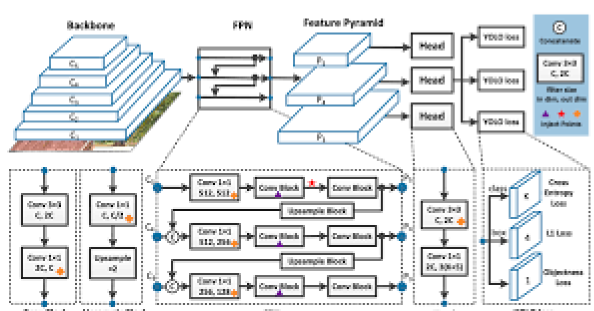

Figure 1: Yolov7 Architecture

B. DEEPSORT

Frame-by-frame data correlation is handled by the SORT method using a straightforward Kalman filter, and correlation is measured using the Hungarian technique. This algorithm has produced positive results at high frame rates. However, when SORT disregards the detected target's

appearance feature, its accuracy depends on how uncertainly the target state is estimated. Additionally, in order to increase tracking efficiency, SORT deletes targets that have not been matched in a continuous frame; however this leads to the ID switch problem, in which the ID given to the target is simple to alter on a regular basis. DeepSORT is a computer vision tracking technique that tracks objects while giving each one a unique ID. The SORT (Simple Online Real time Tracking) method has been extended by DeepSORT. In order to eliminate identity switches and improve tracking, DeepSORT incorporates deep learning into the SORT algorithm.

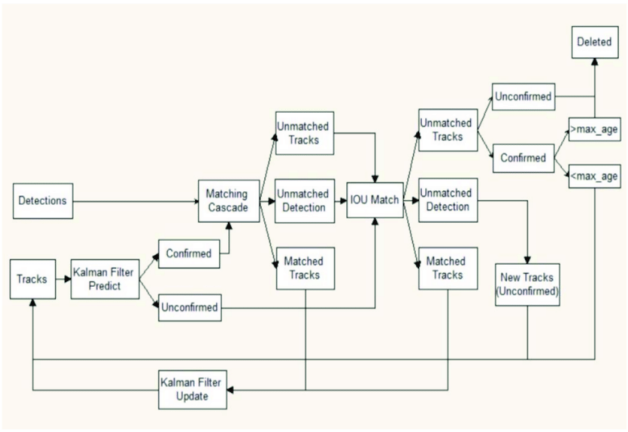

Figure 2: DeepSort tracking procedure

The Kalman filter is an essential part of deep SORT. Eight variables make up our state: (u, v, a, h, u', v', a, h'), where (u,v) are the centre's of the bounding boxes, (a) is the aspect ratio, and (h) is the height of the image. The other variables are the variables' individual velocities. The Kalman filter assist in accounting for noise in detection and makes use of prior state to forecast a suitable match for bounding boxes. For tracking, a Kalman filter and Hungarian algorithm combination is applied. Here, the Hungarian technique supports frame-by-frame data association by applying an association metric that computes bounding box overlap while Kalman filtering is carried out in image space. The final level of matching is performed by DeepSORT using IoU [16] matching, which can reduce significant alterations brought on by apparent mutations or partial blockage. Additionally, in order to extract a differentiating feature embedding from the output of the object detection network for the purpose of computing similarity, DeepSORT adapts ReID model.

Proposed DeepSortAlgorithm

1. YOLOv7 detects and records the position and position ID of the detection boxes for the vehicles it finds in the video frame.
2. Kalman filtering predicts the position of the vehicles while saving the position and the ID of the predicted boxes.
3. All newly predicted frames are kept in a temporary unit for the subsequent detection and prediction phase.
4. Position division is carried out for each newly projected frame position, and various thresholds are defined for various position areas. Calculate the distance between the newly appearing box and the previously appearing boxes by comparing the positions of all expected boxes that have already materialized. This is regarded as the subsequent position of a particular box that has previously appeared when the Euclidean distance to it less than the predetermined threshold is.
5. Update the previously predicted box's position to the recently predicted box's position and wipe the previously predicted box's ID. Compare the updated estimated position to the previous ID.
6. Coordinates of tracked bounding boxes are the output.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this research experiment, use three training datasets: the COCO128 data set, the COCO2017 data set, and the Self Custom SC_ COCO datasets—are used to examine the structure and performance of YOLOv7's network. I choose the COCO128 dataset because I don't sure if the network's training set, which consists of 128 photos of various classes, would produce the same results as the final YOLOv7 network output. In order to produce the best test results, they would produce the same results as the final YOLOv7 network output. In order to produce the best test results, the parameters and training scales are set accordingly. We validate the test results' accuracy using the COCO2017 dataset and the self-made me dataset to obtain high accuracy in vehicle identification and recognition. This article also contrasts the other sophisticated networks to show how far the network models have advanced.

A. Dataset

Both the COCO128 and COCO2017 datasets used in the experiments are from the official COCO dataset website. The Microsoft Image Recognition team's data collection, known as COCO (Common Objects in Context), is fully named. COCO records now include three types of labels: JSON-stored object instance, object key point, and caption. Training, Validation, and Testing are the three sections of the COCO2017 dataset used in this study. Total storage is about 25 GB and each part contains 118,287, 5,000, and 40,670 images. The training and validation datasets contain annotations, but the test dataset does not contain label information. Now there are 80 categories in the dataset, most of which are collections of target detection information for instance, persons, automobiles, car, trucks and buses etc.
The self-made custom dataset includes 4456 images of the front, 8914 images of the roof, 1108 images of the back, and 20,880 images of other objects (strong light, night, multiple front and rear, etc.).

B. Experiment Environment

The experiment's primary hardware and software configuration is as follows: The computer's operating system is Windows 10 64-bit, the compilation environment is python 3.8; Anaconda with Tensorflow, the visual card is an

NVIDIA GTX 3080Ti with 8GG of RAM, the processor is an Intel ® Core i5-4590 running at 3.30GHz.

C. parameter settings for model

The ir0 learning rate, an essential hyperparameter [18] in supervised learning and deep learning, controls whether and when the objective function converges to the local minimum. The objective function can reach the local minimum in a timely manner with the right learning rate. When using the gradient descent method, momentum is a typical acceleration strategy to quicken convergence. To avoid over fitting, use weight decay. Weight_decay is a coefficient in the loss function that comes before the regularization term. In general, the regularization term reveals how sophisticated the model is. In order to modify the effect of model complexity on the loss function, weight decay serves this purpose. The complex model loss function will have a high value if the weight decay is high. Box is a language created specifically to make vector drawings simpler. In computer vision, anchor has an anchor point or an anchor box [21]. A fixed reference frame is represented by the anchor box, which frequently appears in target detection. Table 1 displays the parameters of configurations used for network training.

Type	Value set
Learning rate	0.0031
momentum	0.84
Weight_dekey	0.00037
Box	0.0297
Anchor	0.90

Table 1: The parameter configurations

D. Analysis of Result

The following table 2 displays the FPS and run times for various iterations of the models on the three videos as a result of the YOLOv7 test.

Models	First Video		Second Video		Third Video	
	Inference time	FPS	Inference time	FPS	Inference time	FPS
Yolov4		14.8		14.8		14.8
Yolov5	7.0	142	7.61	126	8.20	122
Yolov5 large	29.1	35	28	35	28.81	35
Yolov7	51.2	18.75	49.5	20.2	50	20

Table 2: FPS Comparison of different Detection Models

The following evaluation metrics were employed in the experiment:

Tracking Accuracy (TA) - False positives, missed targets, and identity shifts are the three error factors combined in this measurement.

Tracking Precision (TP) - A summary of total tracking precision measured by the amount of ground-truth and reported position bounding boxes overlap.

IDF1 Score - the proportion of accurately identified detections to the usual number of computed and ground-truth detections.

Target Tracked (TT) – it is a proportion of ground-truth trajectories that a track hypothesis covers for at least 80%.

Table 3 shows that YOLOv7-DeepSORT vehicle tracking and detection is stronger to YOLOv5-DeepSORT.

Model	TA	TP	IDF1	TT
Yolov5s	39.60	80.85	52.39	15.45%
Yolov5m	39.01	81.87	51.56	17.41%
Yolov5i	40.77	81.56	52.43	20.70%
Yolov7	40.92	82.08	53.65	20.92%

Table 3: Tracking performance is compared.

Compared to YOLOv5-DeepSORT [22], this network has a greater tracking accuracy.

V. CONCLUSION AND FUTURE WORK

The objective of this work is to create an application that is speedy and precise enough to achieve the task of tracking and identifying automobiles in real-time. The DEEP-SORT algorithm will be in charge of tracking the vehicles recognized frame by frame while the YOLOV7 object-detection model is being trained and refined to detect our objects (car, truck, and bus etc). This end-of-degree application was created with the express purpose of enhancing autonomous vehicle video surveillance systems and smart city traffic management. In both a challenging situation and weather conditions, the algorithm has a high identification rate and recognition speed. In addition to running quickly, YOLOv5s significantly minimizes the model's storage requirements. To create YOLOv7-DeepSORT, we add YOLOv7 as an object detection network to Deep-SORT. Compared to YOLOv5- DeepSORT, this network has greater tracking accuracy, according to experiments. The experimental findings on a difficult data set evaluated at various camera heights, Rear View of the Vehicles, and angles showed flexibility and The suggested method's accuracy is good. The classification of automobiles is one upcoming project that will increase the system's dependability. It is beneficial to include a classification procedure because it further enhances the performance of detection.

REFERENCES

[1] G. O. Young,. Muhammad Saleem, Sagheer Abbas, Taher M.Ghazal, Muhammad Adnan Khan, Nizar Sahawneh, Munir Ahmad."Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques". Scienceirect, Egyptian Informatics Journal, Volume 23, Issue 3, September 2022, Pages 417-426.

[2] Mahmoud Abbasi, Amin Shahraki, Amir Taher kordi. "Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey". ScienceDirect Computer Communications,Volume 170, 15 March 2021, Pages 19-41.

[3] Md Zahidul Islam, Md Shariful Islam, Md Sohel Rana. "Problem Analysis of Multiple Object Tracking System: A Critical Review". International Journal of Advanced Research in Computer and Communication Engineering,Vol. 4, Issue 11, November 2015.

[4] Nikolajs Bumanis, Gatis Vitols, Irina Arhipova and Egons Solmanis. "Multi-object Tracking for Urban and Multilane Traffic: Building Blocks for Real-World Application". In Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1, pages 729-736.

[5] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai Bo Ma. "A Review of Yolo Algorithm Developments". ScienceDirect,Procedia Computer Science Volume 199, 2022, Pages 1066-1073.

[6] Viswanatha V,Chandana R K,Ramachandra A.C."Real Time Object Detection System with YOLO and CNN Models: A Review". Journal of Xi'an University of Architecture & Technology,Volume XIV, Issue 7, 2022, Page No: 144-151

[7] Kavitha N., Chandrappa D.N.."Optimized YOLOv2 based vehicle classification and tracking for intelligent transportation system". ScienceDirect, Results in Control and Optimization, volume 2, April 2021, 100008.

[8] Heng Zhang, Yingzhou Wang, Yanli Liu 1 and Naixue Xiong. " IFD: An Intelligent Fast Detection for Real-Time Image Information in Industrial IoT". Appl. Sci. 2022, 12, 7847.

[9] Reza Shakerian, Meisam Yadollahzadeh-Tabari, Seyed Yaser Bozorgi Rad. "Proposinga FuzzySoft-max-based classifier in a hybrid deeplearning architecture for human activity recognition". IET Biometricspublishedby John Wiley& SonsLtd on behalfof The Institutionof Engineeringand Technology, IET Biom.2022, page no: 171–186.

[10] Mei-Ling Huang, Yi-Shan Wu. "GCS-YOLOV4-Tiny: A lightweight group convolution network for multi-stage fruit detection".Mathematical Biosciences and Engineering Volume 20, Issue 1, September 2022, page No: 241–268.

[11] Sankar K. Pal1, Anima Pramanik, J. Maiti, Pabitra Mitra. "Deep learning in multi-object detection and tracking: state of the art". Springer Nature 2021, Applied Intelligence, page no:6400–6429.

[12] Tuan Linh Dang, Gia Tuyen Nguyen and Thang Cao. "object tracking using improved Deep sort yolov3 architecture". ICIC Express Letters, Volume 14, Number 10, October 2020.

[13] Jiankun Ling. "Target Tracking Using Kalman Filter Based Algorithms". IOP Publishing, ICAITA 2021, Journal of Physics: Conference Series 2078 (2021) 012020.

[14] Han Wu, Jiahao Nie, Zhiwei He, Ziming Zhu and Mingyu Gao. "One-Shot Multiple Object Tracking in UAV Videos Using Task-Specific Fine-Grained Features". Remote Sens. 2022, 14, 3853.

[15] Mohamed Soudy, Yasmine M. Afify, Nagwa Badr. "RepConv: A novel architecture for image scene classification on Intel scenes". IJICIS, Vol.22, No.2, page no:63-73.

[16] Di Tian, Yi Han, Shu Wang, Xu Chen, Tian Guan. "Absolute size IoU loss for the bounding box regression of the object detection". ScienceDirect, Neurocomputing Volume 500, 21 August 2022, Pages 1029-1040.

[17] Xuewen Chen, Yuanpeng Jia, Xiaoqi Tong and Zirou L. "Research on Pedestrian Detection and DeepSort Tracking in Front of Intelligent Vehicle Based on Deep Learning". Sustainability 2022, 14, 9281.

[18] Jia Wu,1a Xiu-Yun Chen,Hao Zhang, Li-Dong Xiong, Hang Lei, Si-HaoDeng. "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization". ScienceDirect, Journal of Electronic Science and Technology,Volume 17, Issue 1, March 2019, Pages 26-40.

[19] Ziwen Chen, Lijie Cao and Qihua Wang. " YOLOv5-Based Vehicle Detection Method for High-Resolution UAV Images". Hindawi,Mobile Information Systems,Volume 2022, Article ID 1828848, 11 pages.

[20] K. K. Santhosh, D. P. Dogra, P. P. Roy. "Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey". ACM Computing Surveys, Volume 53 Issue 6,November 2021, Article No.: 119,pp 1–26.

[21] Tao Zhanga, Bo Jin, Wenjing Jia. "An anchor-free object detector based on softens optimized bi-directional FPN". Computer Vision and Image Understanding, Volume 218, April 2022, 103410.

[22] Xinhua Zhao, Zheng Huang, Yongjia Lv. "Research on Real-Time Diver Detection and Tracking Method Based on YOLOv5 and DeepSORT". 2022 IEEE International Conference on Mechatronics and Automation (ICMA),August 2022.

[23] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". Springer open, Shorten and Khoshgoftaar J Big Data (2019) 6:60.

[24] Sixian Chana,Yangwei Jia,Xiaolong Zhoube, Cong Bai, Shengyong Chen, Xiaoqin Zhang. "Online multiple object tracking using joint detection and embedding network". ScienceDirect, Pattern Recognition Volume 130, October 2022, 108793.

[25] Yating Liu,Xuesong Li, Tianxiang Bai, Kunfeng Wang, Fei-Yue Wang. "Multi-object tracking with hard-soft attention network and group-based cost minimization". ScienceDirect,Neurocomputing Volume 447, 4 August 2021, Pages 80-91.

AUTHORS

First Author- Vinod Kumar Yadav, PhD (Research Scholar), Computer Science & Engineering Department, Ravindra Nath Tagore University, Bhopal, India and mail id: vinod.it210@gmail.com.

Second Author- Dr. Pritaj Yadav, Phd in Computer Science & Engineering, Assocaite Professor in Computer Science and Engineering Department, Ravindra Nath Tagore University, Bhopal, India and mail id: yadavpritaj@gmail.com.

Third Author- Dr. Shailja Sharma, Phd in Computer Science & Engineering, Professor in Computer Science and Engineering Department, Ravindra Nath Tagore University, Bhopal, India and mail id: shailja.sharma@aisectuniversity.ac.in.