Tracking and Detection of Vehicle using Deep Learning

Abstract:
Computer vision techniques are used to track and detect vehicles. Detecting, identifying, and counting vehicles on highways is crucial for traffic surveillance and intelligent transportation systems. Developing an effective traffic monitoring model is a challenging task. Traditional artificial intelligence-based vehicle detection systems have limitations in detecting vehicles accurately and being robust. This paper proposes a deep learning model for vehicle detection, tracking, and counting that is based on an efficient Yolov7 single shot detector and Deep- Sort of Multi Object Tracking algorithms. The proposed model investigates automobile detection algorithms and proposes detection models based on moving vehicle footage as survey data.

Index Terms: Computer vision, deep learning, Deep-sort, Yolov7, MOT

1. INTRODUCTION

Computer vision and deep learning technologies are transforming the way we approach road safety. By analysing data from traffic cameras and other sources, these technologies can identify potential hazards and predict the likelihood of accidents before they happen. One important application of computer vision in road safety is vehicle type recognition. By using machine learning algorithms, traffic surveillance videos can be analysed to identify different types of vehicles on the road, such as cars, trucks, buses, and motorcycles. This information can be used to develop more effective traffic management strategies, such as implementing different speed limits for different vehicle types or identifying areas where certain types of vehicles may be at higher risk of accidents. In addition to vehicle type recognition, computer vision and deep learning can also be used for other applications related to road safety, such as pedestrian detection and traffic flow analysis. By analysing patterns in traffic data, these technologies can help identify areas where congestion is likely to occur and develop strategies to reduce traffic jams and improve overall traffic flow. Overall, the development of intelligent traffic systems based on computer vision and deep learning has the potential to significantly improve road safety in smart cities. By predicting accidents before they happen and identifying potential hazards, these technologies can help reduce the risk of accidents and make our roads safer for all users. Recent advancements in object detection and tracking technologies, along with increased computing power, have greatly improved the ability to replace hardware-based systems. Object tracking involves identifying and detecting multiple targets in a video, such as people, vehicles, and animals, without knowing the exact number of targets. To perform trajectory prediction and accurate search, each target is assigned a separate ID.

However, multiple object tracking presents challenges such as occlusion, deformation, motion blur, crowded environments, fast motion, changes in illumination, scale, and more. These challenges become even more complicated when tracking vehicles, as it requires addressing issues like trajectory initialization and termination, mutual interference between similar targets, and more. Thanks to the rapid growth of deep learning and the development of detection-based tracking, target detection performance has increased dramatically. This has led to the establishment of detection-based tracking as the foundation for multi-object tracking in use today, significantly advancing multi-object tracking activities. The suggested model makes use of a convolution neural network technique based on Yolov7 deep learning to track and identify automobiles in a video frame. The vehicles are then tracked frame by frame using the deep sort tracker.

2. RELATED WORK

The YOLO algorithm is faster than other existing algorithms, capable of detecting frames at a speed of 45 per second. Instead of moving the window, it divides the original image into small, non-conforming squares and deforms them to build a map of objects of similar sizes. Each feature in the feature map corresponds to a small square in the original image, and the central points of these squares can be used to predict the target. YOLOv2 improves on YOLOv1 by using batch averaging, high-resolution photos, and an a priori box. YOLOv3 uses a clustering approach to determine a priori box sizes and implements multilevel function for object detection. YOLOv4 introduces top-down PAN feature fusion and a Focus module. The SORT and DeepSORT algorithms are concerned with safety in Multi Object Tracking. SORT uses Kalman filter and Hungarian matching to anticipate target position and compare prediction results with object detection networks like YOLO, but suffers from identity flips due to changing target motion and occlusion. Through the addition of cascade matching and other features, DeepSORT increases efficiency. MOT combines detection and tracking framework and typically tracks and predicts targets by assessing their similarity in nearby frames using various techniques.

The present Multi-Object Tracking (MOT) framework can be classified into three categories: MOT based on joint detection and tracking, MOT based on attention mechanism, and MOT based on tracking by detection (DBT).The first step in the DBT framework is to detect the targets present in each frame of the video sequence. This involves identifying the location of the targets using a bounding box. Once the targets are identified, they can be separated from the rest of the image by cutting along the bounding box boundaries, thus allowing for individual analysis of each target in the image.

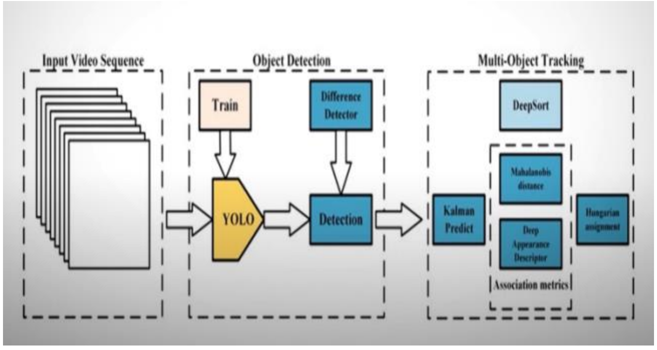
Methodology

The proposed model is implemented using Python and TensorFlow. The data used for training and testing the model is obtained from public datasets, such as the KITTI dataset. The YOLOv7 algorithm is used for vehicle detection, while Deep SORT is utilized for tracking the detected vehicles. The performance of the proposed model is evaluated using various metrics, including mean average precision (mAP), intersection over union (IoU), and frame rate.

YOLOv7 is the top-performing real-time object detection model for computer vision tasks, with a combination of high accuracy and speed. It boasts a faster and more robust network architecture that integrates features better, achieves more precise object detection performance, and has an improved loss function, label assignment, and model training efficiency. The backbone of YOLOv7 is the efficient layer aggregation networks, which are designed for extended performance with two main considerations: the number of parameters and computational density of the model.
E-ELAN, an extension of ELAN (Efficient Layer Aggregation Networks), which enhances learning and deepening of networks by managing gradient paths. A compound model scaling strategy can be applied to improve YOLOv7's performance, where the width and depth of concatenation-based models are scaled in coherence. After training, re-parameterization is used to enhance the model, albeit with an extended training duration, but with improved inference outcomes. Two types of re-parameterization are used: model level and module level ensemble. Model level re-parameterization is achieved by training multiple models with the same parameters and different data and averaging their weights to obtain the final model. Module level re-parameterization divides the model training procedure into phases and generates the final model by ensembling the outputs. YOLOv7 uses RepConvN in the re-parameterized convolution architecture to avoid identity connections when replacing a convolutional layer with residual or concatenation. The lead head in YOLOv7 is the final output head, and the auxiliary head is used for middle-layer training. The Label Assigner method assigns soft labels based on the ground truth and network prediction outcomes, and the Lead Head makes predictions and generates soft labels that are used to quantify the loss of lead and auxiliary heads.

2.1 Architecture of YOLOV7:

The most accurate and quick real-time object detection model for computer vision tasks is YOLOv7. In general, YOLOv7 offers a faster and more reliable network architecture that provides a better feature integration method, more accurate performance for object identification, a more reliable loss function, and enhanced label assignment and model training efficiency.
Yolov7's extended efficient layer aggregation networks are built on a foundation of efficient layer aggregation networks, which are characterised by large parameter sets and high computational densities.



2.2 Extended Efficient Layer Aggregation Network (E-ELAN):

The YOLOv7 backbone uses a computational block called E-ELAN, which is short for Extended Efficient Layer Aggregation Network. The YOLOv7 E-ELAN architecture allows the model to learn better by using "expand, shuffle, merge cardinality" to continuously improve the network's learning ability without destroying the original gradient path.

2.3 Compound Model Scaling Technique:

The depth, width, and resolution at which the network was trained are factors that are generally considered by object identification models. YOLOv7 concatenates layers while scaling the network's depth and width simultaneously. This technique maintains the model's architecture while scaling for different sizes. Concatenation-based model, YOLOv7 introduces compound model scaling. The compound scaling method allows the model's properties to be preserved and thus the optimal structure to be preserved. And here's how it works with compound models. Finally, with the same degree of modification to the transition layers, the width factor scaling is carried out.

3.4 Reparametrized planning:

RepConv has shown impressive results when used in VGG architectures, but its application in ResNet or DenseNet results in a considerable drop in accuracy. In YOLOv7, a modified version of RepConv, called RepConvN, is used in the re-parameterized convolution architecture. This modified version eliminates the identity connection when replacing a convolutional layer with a residual or concatenation, aiming to improve accuracy.

3.5 Auxiliary Head Coarse-to-Fine:

A backbone, a neck, and a head comprise a

YOLO architecture. The predicted model outputs are stored in the head. YOLOv7 is not limited to a single head and is inspired by Deep Supervision, a technique commonly used in training deep neural networks. Assisting training in the intermediary layers is the responsibility of the auxiliary head, while the lead head is in charge of the final output.
In contrast to traditional label assignment, which uses ground truth to generate hard labels based on given rules, reliable soft labels employ calculation and optimization methods.

3. Deepsort Algorithm:

By integrating appearance information with its tracking component, the Simple Online and Real time Tracking with a Deep Association metric enables multiple object tracking. For tracking, a combination of the Kalman Filter and the Hungarian algorithm is used. In this case, Kalman filtering is applied in image space, while the Hungarian technique allows for frame-by-frame data association by employing an association metric that calculates bounding box overlap. The Kalman filter forecasts an appropriate match for bounding boxes while also helping to account for noise in detection. A trained convolutional neural network (CNN) is used to obtain movement and appearance information. Additionally, DeepSORT uses the Re-ID model to extract a distinguishing feature embedding from the object detection network's output for the purpose of calculating similarity.

When a vehicle is found in a video frame, YOLOv7 finds it and saves the position and position ID of the detection boxes for that vehicle. The vehicles' positions are anticipated using Kalman filtering, recording both their positions and the predicted boxes' IDs.
In preparation for the upcoming detection and prediction phase, all recently predicted frames are stored in a temporary unit. With each newly projected frame position, position division is done, and different thresholds are specified for different position areas. Compare the locations of all expected boxes that have already manifested to determine the distance between the newly appearing box and the previously appearing boxes. When the Euclidean distance to it is less than the predetermined threshold, this is treated as the new location of a certain box that had previously appeared. Wipe the ID of the previously predicted box and update its position to that of the more recent prediction. Compared to the previous ID, compare the new projected position. The result is the tracked bounding box coordinates.

Kalman filtering

The Kalman filter is a recursive filter that extrapolates from a collection of noisy observations the state space of a linear dynamic system. The Kalman filter, in conjunction with the linear-quadratic regulator (LQR), solves the linear-quadratic-Gaussian control problem (LQG).

Kalman filtering, also referred to as linear quadratic estimation (LQE), is an algorithm designed to improve the accuracy of estimates of unknown variables by using a series of measurements observed over time, which may contain statistical noise and other sources of inaccuracy. The algorithm works by combining the latest measurement with a predicted state estimate, using a mathematical model of the system being measured, to produce an updated estimate of the unknown variables that is more accurate than either the predicted or measured values alone.

It does this by estimating a common distribution of probabilities over the variables for each timeframe.

Experimental Result and Analysis

This research experiment utilizes three training datasets: COCO128, COCO2017, and a self-custom SC_COCO dataset to evaluate the structure and performance of the YOLOv7 network. COCO128 was selected to ensure that the network's training set, which comprises 128 images of various classes, would yield comparable results to the final YOLOv7 network output. To achieve the best testing outcomes, appropriate training scales and parameters were established. The accuracy of the test results was verified using COCO2017 and a self-made dataset to achieve high precision in vehicle identification and recognition. This article also compares YOLOv7 with other advanced network models to illustrate how far the network models have progressed.

A) Dataset :

The COCO128 and COCO2017 datasets used in the experiments are both from the official COCO dataset website.
Data collection by the Microsoft Image Recognition team

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.
The three sections of the COCO2017 dataset used in this study are Training, Validation, and Testing. The total storage is roughly 15 GB, and each part contains 118,287, 5,000, and 40,670 images. The training and validation datasets have annotations, but the test dataset does not. The dataset now has 80 categories, the most of which are collections of target detection information.

B) Result and analysis :

The appropriate learning rate can facilitate the gradient descent method to efficiently converge to a local minimum in the objective function. Incorporating momentum in the gradient descent method is a common technique to accelerate convergence. Additionally, to prevent overfitting, weight decay can be utilized. Weight decay involves introducing a coefficient in the loss function that is applied to the regularization term. The regularization term is indicative of the model's complexity, and weight decay allows for adjusting the impact of model complexity on the loss function. Higher weight decay values result in higher loss values for more complex models. Box is a language that streamlines the creation of vector graphics. In computer vision, an anchor refers to a fixed point or bounding box. The anchor box, which frequently appears in target detection, is a fixed reference frame.

4. Conclusion

In this paper, we presented a novel approach for vehicle tracking and detection utilizing YOLOv7 and Deep SORT algorithms. The proposed model achieved high accuracy and efficiency in vehicle tracking and detection, making it suitable for various applications in traffic management, surveillance, and autonomous driving. The results of the proposed model demonstrate its potential to be used for real-time applications.
While the YOLOV7 object detection model is being trained and improved to detect our objects, the DEEP-SORT algorithm will be in charge of tracking the cars identified frame by frame.
Although conditions such as poor camera quality, occlusion, and low light levels made accurately detecting different classes of vehicles difficult, certain detector-tracker framework combinations performed well in these conditions as well. Future work includes extending the model to track multiple objects, integrating it with in-car driver assistance systems, and enhancing its features to improve accuracy and robustness .

References

[1] Sankar K. Pal1, Anima Pramanik, J. Maiti, Pabitra Mitra. "Deeplearning in multi-object detection and tracking: state of the art". Springer Nature 2021, Applied Intelligence, page no:6400–6429.
[2] 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.
[3] Jiankun Ling. "Target Tracking Using Kalman Filter Based Algorithms". IOP Publishing, ICAITA 2021, Journal of Physics: Conference Series 2078 (2021) 012020.
[4] J.P. Lin, and M.T. Sun, "A YOLO-based traffic counting system." In 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 82-85. IEEE, 2018.
[5] A. Bochkovskiy, C.Y. Wang, and H.Y.M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv preprint arXiv:2004.10934 (2020).
[6] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection." In Proceedings of the IEEE/CVF Conference on Computer. Vision and Pattern Recognition, pp. 10781-10790. 2020.
[7] 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.
[8] Yating Liu, Xuesong Li, Tianxiang Bai, Kunfeng Wang, Fei-Yue Wang. "Multi-object tracking with hard-soft attention network and group-based cost minimization". Science Direct, Neurocomputing Volume 447, 4 August 2021, Pages 80-91.
[9] 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.
[10] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai Bo Ma. "A Review of Yolo Algorithm Developments". Science Direct, Procedia Computer Science Volume 199, 2022, Pages 1066-1073.
[11] 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
[12] 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.
[13] 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.
[14] 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 behalf of The Institutionof Engineeringand Technology, IET Biom.2022, page no:171–186.