Real-time Vehicle Detection and Tracking using YOLO-based Deep Sort Model: A Computer Vision Application for Traffic Surveillance

A. Lakshmi Rishika[1], Ch. Aishwarya[1], A. Sahithi[1], M. Premchender[2]

[1]UG Student, [2]Assistant Professor, [1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Secunderabad, Telangana, India

Abstract

Intelligent vehicle detection and counting are becoming increasingly important in the field of highway management. However, due to the different sizes of vehicles, their detection remains a challenge that directly affects the accuracy of vehicle counts. To address this issue, this paper proposes a vision-based vehicle detection and counting system using You Only Look Once (YOLO-V4) based DeepSORT model for real time vehicle detection and tracking from video sequences. Deep learning based Simple Real time Tracker (Deep SORT) algorithm is added, which will track actual presence of vehicles from video frame predicted by YOLO-V4 so the false prediction perform by YOLOV4 can be avoid by using DeepSort algorithm. The video will be converted into multiple frames and give as input to YOLO-V4 for vehicle detection. The detected vehicle frame will be further analysed by DeepSort algorithm to track vehicle and if vehicle tracked then DeepSort will put bounding box across tracked vehicle and increment the tracking count. The proposed model is trained with three different datasets such as public and custom collected dataset.

Keywords: Object detection, vehicle tracking, YOLO model, Deep learning.

1. Introduction

1.1 Overview

Vehicle detection and tracking is a common problem with multiple use cases. Government authorities and private establishment might want to understand the traffic flowing through a place to better develop its infrastructure for the ease and convenience of everyone. A road widening project, timing the traffic signals and construction of parking spaces are a few examples where analysing the traffic is integral to the project. Traditionally, identification and tracking has been carried out manually. A person will stand at a point and note the count of the vehicles and their types. Recently, sensors have been put into use, but they only solve the counting problem. Sensors will not be able to detect the type of vehicle.

A fundamental source of the economic growth of any nation depends on well-planned and resilient transportation systems based on spatial information. Regardless, most cities around the world are still facing a rampant increase in traffic volume and complications in traffic management, resulting in poor quality of life in modern cities. However, recent advancements in internet bandwidth, artificial intelligence, and sensing technologies have minimized these difficulties by collaboratively bringing forward location intelligence for public safety. Automation in location intelligence in road environments using sensing technologies allow authorities to achieve resilience in road safety, controlled commutes, and assessments of road conditions.

1.2 Problem statement

Several state-of-the-art deep learning models in the domain of near real-time multi-object classification belonging to the You Only Look Once (YOLO) family (two versions of improved YOLOv3 and two versions of YOLOv5) were trained. The models were ensembled such that it tackles several of the existing challenges of real-time object detection, recognition, and classification. The challenges tackled by these YOLO models include the absence of an integrated anchor box selection process, time-taking space-to-depth conversion, the gradient descent problem, weak feature propagation, a large number of network parameters, and problems in the generalization of objects of different sizes and scales. Object detection algorithm based on deep learning is one of the most widely used and challenging tasks in the field of computer vision. It is widely used in many fields such as medical image, unmanned vehicle, security system and robot research. The object detection task is to distinguish the target object in the image from the background information. It usually consists of two parts namely locating and marking the bounding box of the object to be detected in the image and completing the task of classifying the target in the bounding box. One of the most common methods for object detection is YOLO. The algorithm was introduced and outperformed other detection methods in speed and accuracy. Since then, it has been under continuous improvements to enhance its performance. YOLOv4 achieved high performance compared with state-of-the-art object detection methods. Naturally, such performance encouraged researchers to exploit its potentials in transportation. Therefore, this work implements YOLOv4 for object detection and DeepSORT for tracking the detected vehicles. Three different variations of the deep learning models are used and compared their performance: a pre-trained model with the COCO dataset, and two custom-trained models with different datasets. The three different datasets; the COCO dataset, the Berkeley DeepDrive dataset, and our custom developed dataset obtained by a Dash Cam installed onboard vehicle driven on city streets and highways in the Kingdom of Saudi Arabia (KSA).

2. Literature Survey

Wang et al. [1] proposed an algorithm to detect abnormality in vehicles behaviour such as stalled cars and cars speeding up or slowing down. They used YOLO algorithm for detection and Kalman filter for tracking. They tested their framework on videos from traffic cameras. The combination of YOLO and Kalman filter is applicable, despite some scenarios where farther contextual knowledge is needed to improve detection results. Kumar et al. [2] trained a sentiment classification model to detect negative sentiment about a road hazard from Twitter. The data is collected using search filtering with specific terms that relate to traffic. Then, naive Bayes, K-nearest-neighbor and the dynamic language model (DLM) are used to build models to classify the tweets into a hazard and not hazard. Song et al. [3] considered small vehicles on highways in their proposed detection and counting system. They published a new high-definition dataset containing annotations of small objects. They developed a segmentation method to extract and divide roads into remote and proximal areas. They used YOLOv3 as their detection method and the ORB algorithm as a feature extractor. They analysed the trajectories of detected objects for counting purposes. Their proposed method provide good performance as can replace the traditional ways of counting vehicles without any new hardware equipment.

Tejaswin et al. [4] also used random forest classifier to predict traffic incidents. The traffic incidents are clustered and predicted using spatio-temporal data from Twitter. The location information is extracted using NLP and background knowledge by using Freebase API, which is a community-curated structured database containing large number of entities and each one defied by multiple properties and attributes that helps in entity disambiguation. Suma et al. [5] built a classification model using logistic regression with stochastic gradient descent to detect events related to road traffic

from English tweets using Apache Spark. They used the latent Dirichlet allocation (LDA) topic modeling module to filter traffic messages. In addition, they used the Spark MLib library and trained classifiers using SVM, KNN and NB to detect traffic events. Chen et al. [6] aimed at detecting objects by generating 3D object proposals. Their proposed work utilized stereo imagery. They based the method on minimizing an energy function which encodes object size prior, object placement and some context depth information. Then, they used convolutional neural network to use appearance, context and depth information for object detection. The method predicted 3D bounding box coordinate and object pose. The approach proved to be superior to previously published object detection work on the KITTI benchmark.

Sudha and Priyadarshini [7] designed an approach for multiple vehicles detection and tracking. Their work utilized an enhanced YOLOv3 algorithm with an improved visual background extractor for detection. For tracking, Kalman filtering with particle filter technique were deployed. Authors tested the proposed solution on two private datasets, KITTI and DETRAC benchmark datasets. Sang et al. [8] came up with an improved detection model based on YOLOv2, which used the k-means++ algorithm to train a dataset to cluster vehicle bounding boxes. To improve the loss due to different scales of vehicles, normalization was introduced. Moreover, repeated convolution layers were removed to improve feature extraction.

Du et al. [9] proposed the real-time detection of vehicles and traffic lights with the YOLOv3 network, an improved version of YOLOv2, by detecting small objects with balanced speed and precision using a new, high-quality dataset named the Vehicle and Traffic Light Dataset (V-TLD). They used YOLOv3 to detect and classify the vehicles and used an ORB algorithm to obtain driving directions. Mahto et al. [10] used a fine-tuned YOLOv4 for vehicle detection using the UA-DETRAC dataset, which was faster than previous iterations. YOLOv5, despite being produced by a different author than its predecessors, has higher performance in terms of accuracy and speed among the YOLO family. Liu et al., in [11], proposed 3-D constrained multiple kernels, facilitated with Kalman filtering, to track objects detected by a YOLOv3 network. These recent but sophisticated tracking algorithms have improved the accuracy of object tracking, but they require heavy computational power. Here, they propose a simple object-centroid tracking algorithm to track the detection provided by YOLO-based DL networks in multiple lanes of the road in real time. Furthermore, this study compares the use of two YOLO variants, YOLOv3 and YOLOv5, to obtain a real-time vehicle tracking method that can process multiple video streams with a single GPU, using multi-threading techniques.

Ali et al. [12] used inductive loop sensors to detect and count diverse vehicles in lane-less roads. They developed a multiple loop system with a new structure for inductive loop sensors. Their solution was able to sense vehicles and divide them by type. During testing, the system provided accurate counting of vehicles despite the heterogenous traffic conditions. Neupane, et al. [13] proposed a multi-vehicle tracking algorithm that obtains the per-lane count, classification, and speed of vehicles in real time. The experiments showed that accuracy doubled after fine-tuning (71% vs. up to 30%). Based on a comparison of four YOLO networks, coupling the YOLOv5-large network to our tracking algorithm provided a trade-off between overall accuracy (95% vs. up to 90%), loss (0.033 vs. up to 0.036), and model size (91.6 MB vs. up to 120.6 MB). The implications of these results are in spatial information management and sensing for intelligent transport planning. A. H. Abdel-Gawad, et al. [14] proposed a detection-based tracking approach for Multiple VRU Tracking of video from an inside-vehicle camera in real-time. YOLOv4 scans every frame to detect VRUs first, then Simple Online and Realtime Tracking with a Deep Association Metric (Deep SORT) algorithm, which is customized for multiple VRU tracking, is applied. The results of our experiments on both the Joint Attention in Autonomous

Driving (JAAD) and Multiple Object Tracking (MOT) datasets exhibit competitive performance. Tao et al. [15] constructed vehicle identification for images on road using an optimized YOLO method. In this method, the last two fully connected layers are detached and added an average pool layer. The simulation results shows that the accuracy and precision rate is higher than single object detection. This algorithm will compute the intersection-over-union (IOU) distance between each detection and every predicted bounding box from the existing targets. The last algorithm work in SORT algorithm will assign either create unique identities or destroyed it. In DeepSORT there will be deep learning algorithm which helps reduces high number of identity switches and improve efficiency of tracking through occlusions in SORT algorithm.

3. Proposed System

Intelligent vehicle detection and counting are becoming increasingly important in the field of highway management. However, due to the different sizes of vehicles, their detection remains a challenge that directly affects the accuracy of vehicle counts. To address this issue, this paper proposes a vision-based vehicle detection and counting system. A new high-definition highway vehicle dataset with a total of 57,290 annotated instances in 11,129 images is published in this study. Compared with the existing public datasets, the proposed dataset contains annotated tiny objects in the image, which provides the complete data foundation for vehicle detection based on deep learning.

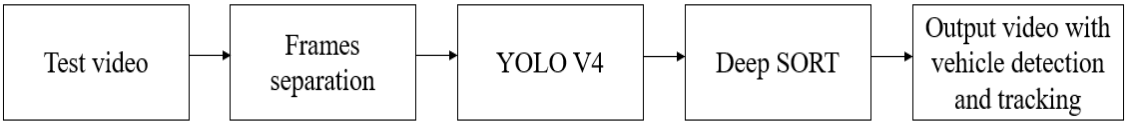

Fig. 1. Block diagram of proposed method.

Figure 1 shows the block diagram of proposed method. The use of deep convolutional networks (CNNs) has achieved amazing success in the field of vehicle object detection. CNNs have a strong ability to learn image features and can perform multiple related tasks, such as classification and bounding box regression. The detection method can be generally divided into two categories. The two-stage method generates a candidate box of the object via various algorithms and then classifies the object by a convolutional neural network. The one-stage method does not generate a candidate box but directly converts the positioning problem of the object bounding box into a regression problem for processing.

Therefore, this work is focused on implementation of You Only Look Once (YOLO-V4) based DeepSORT model for real time vehicle detection and tracking from video sequences. Deep learning based Simple Real time Tracker (Deep SORT) algorithm is added, which will track actual presence of vehicles from video frame predicted by YOLO-V4 so the false prediction perform by YOLOV4 can be avoid by using DeepSort algorithm. The video will be converted into multiple frames and give as input to YOLO-V4 for vehicle detection. The detected vehicle frame will be further analysed by DeepSort algorithm to track vehicle and if vehicle tracked then DeepSort will put bounding box across tracked vehicle and increment the tracking count. The proposed model is trained with three different datasets such as COCO, Berkeley and Dash Cam dataset. Here, Dash Cam dataset is the custom collected dataset.

3.1 Dataset

COCO dataset: The dataset contains car images with one or more damaged parts. The img/ folder has all 80 images in the dataset. There are three more folders train/, val/ and test/ for training, validation

and testing purposes respectively. Berkeley dataset: Berkeley DeepDrive(link is external) (BDD) and Nexar announced the release of 36,000 high frame-rate videos of driving, in addition to 5,000 pixel-level semantics-segmented labeled images, and invited public and private institution researchers to join the effort to develop accurate automotive perception and motion prediction models.

3.2 Preprocessing and frame separation

Digital image processing is the use of computer algorithms to perform image processing on digital images. As a subfield of digital signal processing, digital image processing has many advantages over analogue image processing. It allows a much wider range of algorithms to be applied to the input data — the aim of digital image processing is to improve the image data (features) by suppressing unwanted distortions and/or enhancement of some important image features so that our AI-Computer Vision models can benefit from this improved data to work on. To train a network and make predictions on new data, our images must match the input size of the network. If they need to adjust the size of images to match the network, then they can rescale or crop data to the required size.

they can effectively increase the amount of training data by applying randomized augmentation to data. Augmentation also enables to train networks to be invariant to distortions in image data. For example, they can add randomized rotations to input images so that a network is invariant to the presence of rotation in input images. An augmented Image Datastore provides a convenient way to apply a limited set of augmentations to 2-D images for classification problems.

They can store image data as a numeric array, an ImageDatastore object, or a table. An ImageDatastore enables to import data in batches from image collections that are too large to fit in memory. they can use an augmented image datastore or a resized 4-D array for training, prediction, and classification. They can use a resized 3-D array for prediction and classification only.

There are two ways to resize image data to match the input size of a network. Rescaling multiplies the height and width of the image by a scaling factor. If the scaling factor is not identical in the vertical and horizontal directions, then rescaling changes the spatial extents of the pixels and the aspect ratio.

Cropping extracts a subregion of the image and preserves the spatial extent of each pixel. They can crop images from the center or from random positions in the image. An image is nothing more than a two-dimensional array of numbers (or pixels) ranging between 0 and 255. It is defined by the mathematical function f(x,y) where x and y are the two co-ordinates horizontally and vertically.

3.2.1 Resize image: In this step-in order to visualize the change, they are going to create two functions to display the images the first being a one to display one image and the second for two images. After that, they then create a function called processing that just receives the images as a parameter.

Need of resize image during the pre-processing phase, some images captured by a camera and fed to our AI algorithm vary in size, therefore, they should establish a base size for all images fed into our AI algorithms.

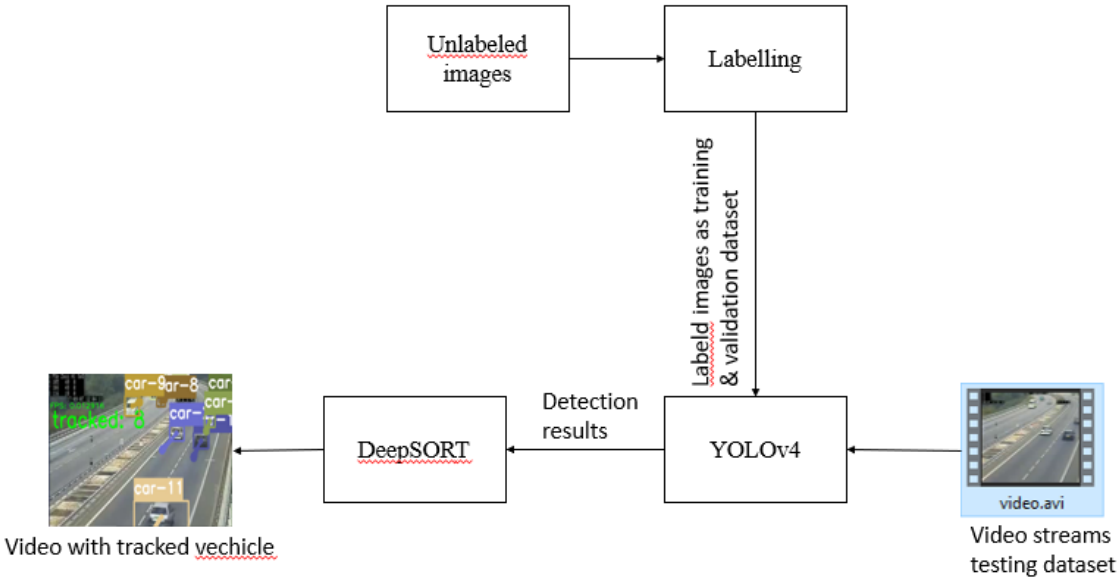

Fig. 2. The methodology of proposed system.

3.3 YOLO V4

The test video from the datasets is taken to apply the algorithms and splitted into different frames. The spitted frames are applied to YOLOV4 to detect the vehicles and to track the vehicles, frames are applied to DeepSort. Finally, the output video is generated with vehicle detection and tracking.

Tracking vehicles is another aspect of research in transportation. DeepSORT is a recent tracking algorithm, extending SORT (Simple Online and Real-Time) tracking algorithm. The original algorithm was developed considering MOT task. With the main goal of supporting online and real-time applications. This means that the tracker associates detected objects from previous and current frames only.

YOLO is a Convolutional Neural Network object detection system, that handles object detection as one regression problem, from image pixels to bounding boxes with their class probabilities. Its performance is much better than other traditional methods of object detection, since it trains directly on full images. YOLO is formed of 27 CNN layers, with 24 convolutional layers, two fully connected layers, and a final detection layer.

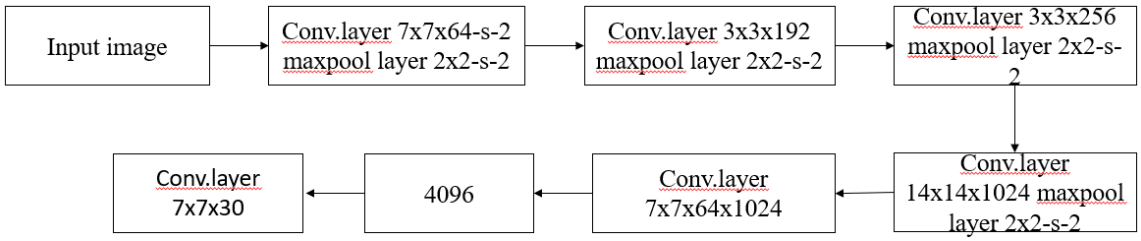

Fig. 3. Structure of YOLOV4 algorithm.

YOLO divides the input images into an N-by-N grid cell, then during the processing, predicts for each one of them several bounding boxes to predict the object to be detected. Thus, a loss function has to be calculated. YOLO calculates first, for each bounding box, the Intersection over Union (IoU); It uses then sum-squared error to calculate error loss between the predicted results and real objects. The final loss being the sum of the three loss functions:

1) classification loss: related to class probability.
2) localization loss: related to the bounding box position and size.
3) confidence loss measuring the probability of objects in the box.

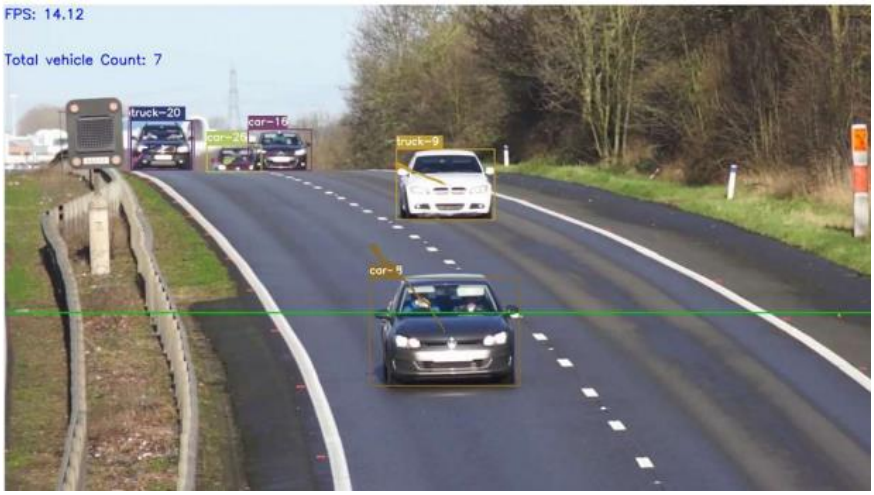

Fig. 4. Detection of cars using YOLOV4.

YOLO framework though is the one-stage methods which directly converts the object bounding box positioning issue into a regression issue for processing without generate a candidate box. The YOLO network breaks the picture into a defined number of grids. Each grid is accountable for estimating objects within the grid whose central points are. Then after several years the YOLO framework has been developed it algorithm to version 4 which improve speed and accuracy of object detection. YOLO V4 extracts the residual network part of the future entreats each region in the entire future map equally considering that each region contributes the same to the final detection however in real life scenes complex and rich contextual information often exist around the object to be detected in the image and each region in the feature map is treated equally resulting in a lack of network feature expression ability inaccurate bounding box position poor robustness and other problems. To solve these problems a channel attention mechanism module is introduced into the YOLO V4 object detection algorithm.

3.4 DeepSort

DeepSORT algorithm is deployed for the purpose of tracking. The enhanced version of the algorithm where the association metric is substituted by an informed metric integrating motion and appearance information using Convolutional Neural Network. The algorithm takes the detection outputs from the previous stage and run tracking for each detected object. In tracking by detection scheme, the accuracy of tracking is based on the quality of detection results. The Kalman filter an important role in deep sort. It identifies noise in detecting and uses previous states to predict the closed frame surrounding the object best suited. Each time it detects an object it creates a track containing all the necessary information of that object it also tracks and deletes track with detection time exceeds a given threshold due to objects are out of frame. In addition to eliminate duplicates they set a minimum threshold value for detection in the first frame the next problem lies in association between new objects and new predictions from the Kalman filter.

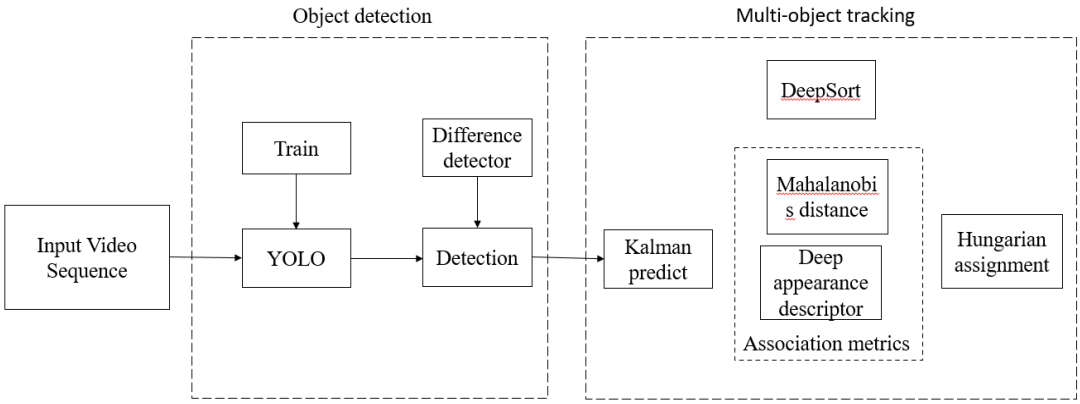

Fig. 5. DeepSort algorithm for multi-object Tracking.

DeepSORT which is an improved version of SORT is one of the most popular state-of-the-art objects tracking frameworks today. DeepSORT has integrated a pre-trained neural network to generate feature vectors to be used as a deep association metric. Since DeepSORT was developed focusing on the Motion Analysis and Re-identification Set (MARS) dataset, which is a large-scale video-based human reidentification dataset, it uses a feature extractor trained on humans which does not perform well on vehicles. Several state-of-the-art object detection and tracking algorithms including SORT and DeepSORT were deployed detect and track different classes of vehicles in their region of interest and it has been stated that the trackers did not perform ideally at predicting vehicle trajectories which resulted in ID switches during occlusions. A vehicle tracking fuses the prior information of the Kalman filter to solve the problem of vehicle tracking under occlusion. But it has been stated that the proposed method does not perform well if the target is lost for a longer period.

4. Results and Discussion

To implement this project, we have designed following modules.

1) Generate & Load YOLOv4-DeepSort Model: using this module we will generate and load YOLOV$-DeepSort model.
2) Upload Video & Detect Car & Truck: using this module we will upload test video and then apply YOLOV4 to detect vehicle and this detected vehicle frame will be further analyse by DeepSort to track real vehicles.

The pre-trained YOLOv4 model is obtained by training the model on the COCO dataset.

Fig. 6. Detection and tracking of cars and trucks using YOLOV4 and DeepSort algorithms.

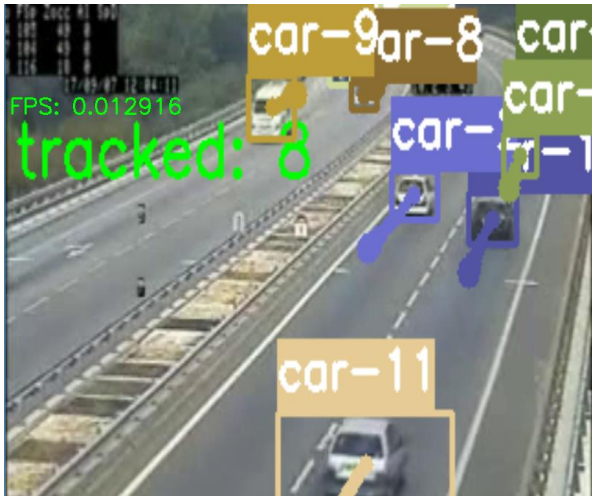

Fig. 7. Tracking of vehicles using Frame Per Second and application till the end of the video.

6. Conclusion

In conclusion, the vehicle detection and tracking method presented uses TensorFlow library with DeepSORT algorithm based on YOLOv4 model. It can be proven that using YOLOv4 and YOLOv4-tiny is acceptable and faster than previous one. It can be use in Realtime surveillance camera in the highway or recording video to evaluate the number of vehicles pass by according to what time it started recorded to last recorded. This data then can be used for traffic management by implementing answer if the place proven a lot of congestion or not. It is the best to use YOLOv4 model than previous model YOLOv3 if the system wants the highest accuracy with acceptable speed. If the system wants the best accuracy with the highest speed as possible because limitation in hardware or to process it in real-time, it is recommended to use YOLOV4-tiny model which it can achieve higher accuracy. This system can be improved to be more adaptable for vehicle detection if using several suggestion ideas. A vehicle tracking algorithm based on the framework suggested in DeepSORT which is capable of tracking the nonlinear motion of vehicles with a high level of accuracy. The proposed algorithm utilizes YOLOv4 with Darknet, an open-source neural network framework, for vehicle localization and identification. The number of detection errors was minimized by optimizing the training of the detector through hyperparameter optimization and data augmentation.

References

[1] C. Wang, A. Musaev, P. Sheinidashtegol, and T. Atkison, “Towards Detection of Abnormal Vehicle Behavior Using Traffic Cameras,” in Big Data -- BigData 2019, 2019, pp. 125–136.

[2] Kumar, A.; Jiang, M.; Fang, Y. Where not to go? In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval; ACM: New York, NY, USA, 2014; Volume 2609550, pp. 1223–1226.

[3] H. Song, H. Liang, H. Li, Z. Dai, and X. Yun, “Visionbased vehicle detection and counting system using deep learning in highway scenes,” Eur. Transp. Res. Rev., vol. 11, no. 1, p. 51, Dec. 2019.

[4] Tejaswin, P.; Kumar, R.; Gupta, S. Tweeting Traffic: Analyzing Twitter for generating real-time city traffic insights and predictions. In Proceedings of the 2nd IKDD Conference on Data Sciences; ACM: New York, NY, USA, 2015; pp. 1–4.

[5] Suma, S.; Mehmood, R.; Albeshri, A. Automatic Event Detection in Smart Cities Using Big Data Analytics. In Proceedings of the Communications and Networking; Metzler, J.B., Ed.; Springer: Cham, Switzerland, 2018; Volume 224, pp. 111–122.

[6] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1259–1272, 2018.

[7] D. Sudha and J. Priyadarshini, "An intelligent multiple vehicle detection and tracking using modified vibe algorithm and deep learning algorithm," Soft Comput., vol. 0123456789, 2020.

[8] Sang, J.; Wu, Z.; Guo, P.; Hu, H.; Xiang, H.; Zhang, Q.; Cai, B. An improved YOLOv2 for vehicle detection. Sensors 2018, 18, 4272.

[9] Du, L.; Chen, W.; Fu, S.; Kong, H.; Li, C.; Pei, Z. Real-time detection of vehicle and traffic light for intelligent and connected vehicles based on YOLOv3 network. In Proceedings of the 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, UK, 14–17 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 388–392.

[10] Mahto, P.; Garg, P.; Seth, P.; Panda, J. Refining YOLOv4 for Vehicle Detection. Int. J. Adv. Res. Eng. Technol. (IJARET) 2020, 11, 409–419.

[11] Liu, T.; Liu, Y. Deformable model-based vehicle tracking and recognition using 3-D constrained multiple-Kernels and Kalman filter. IEEE Access 2021, 9, 90346–90357

[12] S. Sheik Mohammed Ali, B. George, L. Vanajakshi, and J. Venkatraman, "A multiple inductive loop vehicle detection system for heterogeneous and lane-less traffic," IEEE Trans. Instrum. Meas., vol. 61, no. 5, pp. 1353– 1360, 2012.

[13] Neupane, Bipul, Teerayut Horanont, and Jagannath Aryal. "Real-Time Vehicle Classification and Tracking Using a Transfer Learning-Improved Deep Learning Network." Sensors 22.10 (2022): 3813.

[14] A. H. Abdel-Gawad, A. Khamis, L. A. Said and A. G. Radwan, "Vulnerable Road Users Detection and Tracking using YOLOv4 and Deep SORT," 2021 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC), 2021, pp. 140-145, doi: 10.1109/JAC-ECC54461.2021.9691441.

[15] Tao J, Wang H, Zhang X, Li X, Yang H (2018) An object detection system based on YOLO in traffic scene. In: 6th International conference on computer science and network technology, pp 315–319.