Deep Highway Multi-Camera Vehicle Re-ID with Tracking Context

Xiangdi Liu[1], Yunlong Dong[1], Zelin Deng[2]
1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
Wuhan, China
2. College of Business, City University of Hong Kong
Hong Kong, China
dengzelinhust@gmail.com

***Abstract*—While object detection and re-identification has become increasingly popular in computer version, The growing explosion in the use of surveillance cameras on highway highlights the importance of intelligent surveillance.multi-camera vehicle Tracking, aiming to seek out all images of vehicle of interest in different cameras, can provide abundant information such as vehicle movement for highway supervision department. This paper focus on a interesting but challenging problem, building a real-time highway vehicle tracking framework.We design a two-stage deep learning-based algorithm framework, including vehicle detection and vehicle re-identification.Vehicle re-identification is the most significant part in this tracking framework, however, the most existing methods for vehicle Re-ID focus on the appearance or texture of single vehicle image and achieve limited performance. In this paper, we propose a novel deep learning-based network named VTC (Vehicle Tracking Context) to extract features from vehicle tracking context. Extensive experimental results demonstrate the effectiveness of our method, furthermore, intelligent surveillance system based on proposed tracking framework has been successfully use in Beijing-Hong Kong-Macao Expressway.**

***Keywords*—Deep Learning; Multi-Camera Vehicle Re-ID; Tracking Context; LSTM**

I. Introduction

Vehicle is the most significant object class in highway video monitoring, attracts more and more focuses in computer vision research field. Multi-camera vehicle tracking is a task given multi-camera video stream as inputs, searching the vehicle of interest in the multiple cameras image gallery. Tracking vehicle of interest in different cameras on the highway can highly improve work efficiency of highway supervision department and emergency response capacity.

Different from object detection, classification, multi-camera vehicle tracking is a high-level application of computer version technology, highly connected with vehicle detection, classification and object Re-ID. For example, if police wants to find a suspect car in a highway with hundreds cameras, it is difficult to monitor all the surveillance video with hundreds of cameras and thousands of hours video. With computer vision technology, we can detect vehicle location and filter large amount vehicle by color, shape, type and appearance. Furthermore, we can measure similarity of target vehicle image and images in database with a re-identification algorithm. In those computer vision technologies, vehicle detection can extract vehicle images with no background noise and vehicle Re-ID can search the most similar images in feature space, are the core method for constructing a progressive multi-camera vehicle tracking system, Fig.1 gives a straightforward description of vehicle detection.

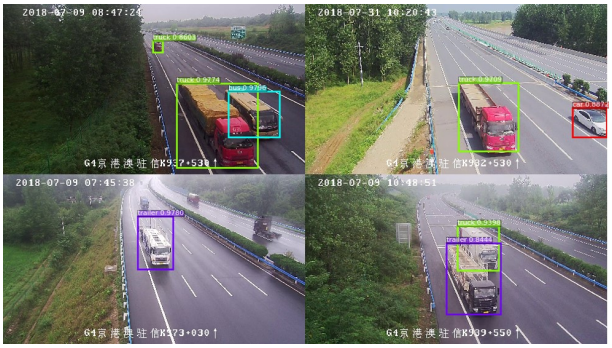

Fig. 1. The typical output of vehicle detection, extracting the bounding box and class of vehicle of interests

However, multi-camera vehicle tracking in real-world highway still face several challenges.First, most research focus on single aspect, e.g. vehicle Re-ID or vehicle detection, it is important to build an integrated system framework from video stream to vehicle tracking images across all cameras.Second, existed Re-ID algorithm use legible and high resolution image as input, but in real-world highway surveillance video, vehicle is always in low resolution and always blurred and noisy. Third, existed Re-ID algorithm neglect spatio-temporal relationship of images, which can provide rich additional information for vehicle tracking.

In this paper, we propose an integrated two-stage multi-camera vehicle tracking framework, which can track vehicle in real-time and had been deployed in Beijing-Hong Kong-Macao Expressway.To get better performance of vehicle tracking, we propose VTC(vehicle tracking context), a novel spatio-temporal feature fusion vehicle Re-ID approach. Moreover, we present a new highway vehicle Re-ID dataset named "HighwayID". Section II provides a brief introduction of related work on vehicle multi-camera tracking while section III describe the construction of the dataset "HighwayID". In section IV, the two-stage vehicle multi-camera tracking

978-1-7281-4390-3/20/$31.00 ©2020 IEEE

framework and the novel Re-ID algorithm VTC are presented in detail. Finally experiments and conclusion are presented in section V and VI.

II. Related Work

In this section we discuss recent work of two related core field, object detection and Re-ID(re-identification). Recently, deep convolutional network is introduced into object detection, [1] is the milestone in object detection, proposed a two stage learning-based object detection method, [2][3] improve detecting accuracy and inferring speed, but still can not satisfy real-time requirement. [4] use a simple but well-designed deep neural network for both object location and classification, presents a classic one-stage approach for objection detection. [5] introduce a multi-scale training method based on YOLOv1, can predict more than 9000 different object categories in real-time.[6] rebuild YOLOv2 network structure with residual block and FPN-like framework for multi-scale detection, achieve better accuracy and higher speed. Particularly, for the detection problem of vehicle, [7] design a two-stage approach employing simple data augmentation tricks to get better detecting performance in vehicle.

Vehicle re-identification is a subtopic of re-identification.In Re-ID field, the most popular topics are re-identification of human face or person.It can be described as below: give a image of interest, point out the same object in candidate gallery.However, there is not much work on vehicle Re-ID.[8] describe the re-identification problem and apply a fundamental neural network, and success apply in signature Re-ID.Google propose a ingenious loss function named "Triplet Loss", widely use in face re-identification.[9] describe a modified Triplet-like loss function which accelerate the training convergence, they also design a two branch network structure to use additional feature for improving Re-ID performance.[10] introduce a approach based on 3-D bounding boxes built around the vehicles, and has better performance when image inputs are in different viewpoint.

III. HighwayID dataset construction

In this section, the dataset construction is introduced in detail:the environments of the surveillance system setting and the construction procedure.

A. Environments of surveillance system setting

To collect a high-quality dataset for vehicle detection and vehicle Re-ID. We select 27 cameras on Beijing-Hong Kong-Macao Expressway. The distance of adjacent camera is 1km to 2km and effective stadia of those camera is 1km. All camera resolution is set to 1920x1080 and frames per second is set to 25. The cameras are installed at about 30 degree and orientation along the road. The scene of camera includes four-lane highway road, subgrade, bridge and farmland.

B. Dataset Construction

With 27 cameras on Beijing-Hong Kong-Macao Expressway, we select videos of several days in different weather condition including sunny day, rain day and foggy day. We collect 2.41 TB highway surveillance videos in total. The vehicle detection dataset and vehicle Re-ID dataset are built based on those videos.

1) Vehicle Detection Dataset: Considering dataset quality and variety, we select three days video then sample those video every 15 frame. A car in 80Km/h can be captured about 10 different snapshots in camera. We label vehicles of four class: car, truck, bus and trailer, then we annotate bounding box of vehicles of all snapshots. We use software labelimg for bounding box labeling, which labelling every vehicle with rectangle bounding box around the whole body and class. We drop labels with bounding box smaller than 48x48 and obtain 14.8GB vehicle detection dataset of 21800 bounding box.

2) Vehicle Re-ID Dataset: The most import task in building vehicle Re-ID dataset is to label vehicle with unique ID, which is also the mast most time-cost as we need recognize multi-camera vehicle ID in thousand image gallery. We select two hour videos of the 27 cameras, extracting pure vehicle images based on vehicle detection dataset. Then we label every car a unique ID with highway entry camera, and vehicle in other cameras will be labelled according to vehicle ID in entry camera. Every vehicle in single camera will reserve three images, which are from far to near.For vehicle may leave on the half way, every vehicle with unique ID has 30 to 81 different snapshots. Finally, we obtain multi-camera snapshots of 845 unique vehicles on the highway.

IV. Vehicle Multi-camera Tracking

In this section, we propose an integrated vehicle tracking framework which has been applied to real-world highway successfully. To achieve better re-identification performance, we introduce a novel spatio-temporal feature fusion network named vehicle tracking context, "VTC".

A. Vehicle Detection with YOLOv3

Object detection is a popular topic in computer version[11] and has been applied to many different .YOLO is a popular one-stage object detection approach for small object detection, in the meantime, YOLO is also high speed and can run in real time. We use YOLOv3 as our vehicle detection frame to detect vehicles from video stream.Input of network is image resized to 618x618 and outputs are Bounding boxes, class and confidence scores. Pure vehicle image will be extracted for further vehicle tracking.

B. Vehicle Re-ID with Tracking Context

The core idea of vehicle Re-ID is compare L2 distance of images in feature space, it is significant to map image into features which represent similarity. Convolutional Neural Network is a popular feature extracting approach because of its powerful feature expression ability. Triplet loss is proposed by google, successfully used in face recognition. Triplet loss

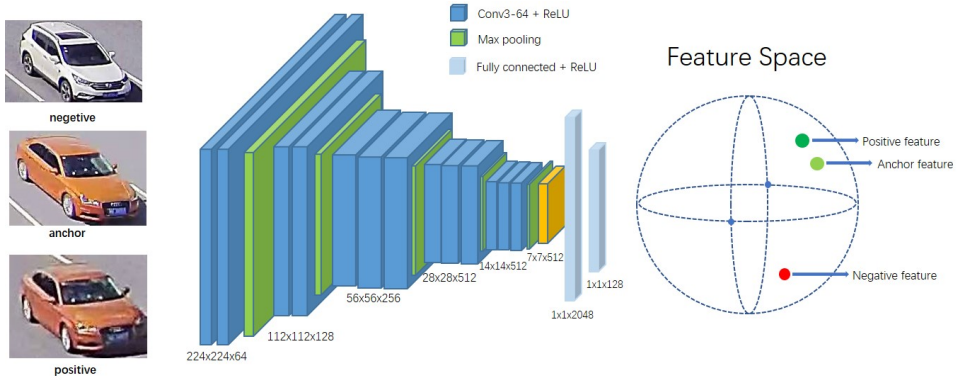

Fig. 2. Distribution of vehicle types

divide images to anchor image, negative image and positive image, aims to minimize the similarity distance between anchor image and positive image, and maximizes the distance between the anchor and a negative of a different identity. This is described in Fig.3. The loss formulation is written below,

$$L = \sum_{i}^{N} ||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)|_2^2|| + \alpha$$

where α is the margin to force distance in same class smaller than distance in different class.

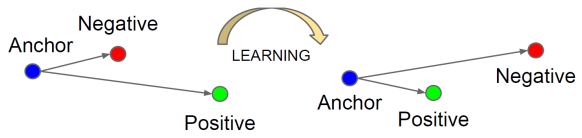

Fig. 3. Triplet loss for re-identification

The baseline vehicle Re-ID network is described in Fig.2, which use CNN as feature extractor and Triplet loss for similarity feature reshape.

In highway, vehicle is small and blurry when it is far away from camera, then will become bigger and more clear as it drive close to camera. In additional, the viewpoint and orientation of vehicle is changing from far to near. Previous method treat vehicle images separately[12] and ignore the relevance of vehicle images in same camera. It is reasonable to fuse from-far-to-near feature in spatio-temporal domain, based on this principle, we exploit different feature fusion method to further improve Re-ID performance and propose the vehicle tracking context network.

A direct method for feature fusion in concatenating feature of vehicles in same camera. The procedure includes extracting vehicle feature with CNN, concatenating features of same vehicle, seteing loss function to Triplet to train a fully connected network. With this method, we can get better performance than baseline method, however, this approach also neglect the far-to-near relationship of those features. Recurrent neural network has been proved to be effective in sequential inputs, play a significant role used in natural language processing. LSTM(Long Short-Term Memory) is specific recurrent neural network (RNN) architecture, designed to model temporal sequences and their long-range dependencies more accurately than RNNs. Inspired by sequential feature presentation ability of LSTM, we design the network structure, using vehicle context images as input, named VTC(vehicle tracking context). Different from concatenating features directly, our network structure fuze vehicle feature and its context image feature with a view to spatio-temporal relationship by making inputs in sequential.

C. Training details

The YOLOv3 network in our experiments is trained with Paddle-Paddle, a new deep learning platform. Without pre-trained model, we set SGDM with momentum of $\mu = 0.8$ as optimizer, learning rate increase from 0 to 1×10^{-4} at the first epoch for warm-up. Training is done in mini-batches of size 16. To combat overfitting we augment training data by flipping , random-crop , color-Jitter.It takes two days for training Yolov3 with Titan RTX to converge. The VTC network is trained with the widely-used framework ”pytorch”. The size of both positive and negative sets are set to 5 and batch size is set to 10.Network optimizer is SGDM . We start with a base learning rate of $\lambda^{(0)} = 0.001$and then drops by repeatedly multiply 0.7 every epoch.

D. System construction

Technically, there are mainly two core components we need to consider for running the vehicle multi-camera tracking system: how to assure sufficient tracking accuracy and how to search vehicle in real-time. The key to solve these problem is reduce the image number in vehicle gallery, inspired by this principle, we design the system in Fig.4. Yolov3 Extract all vehicles on the road and transmit to VTC network as inputs. VTC network extract features of vehicle and store in Database with timestamp and geographical location. When you want to track a vehicle, you need select a vehicle and tracking system will search the car according to its occurrence time and geographical location.

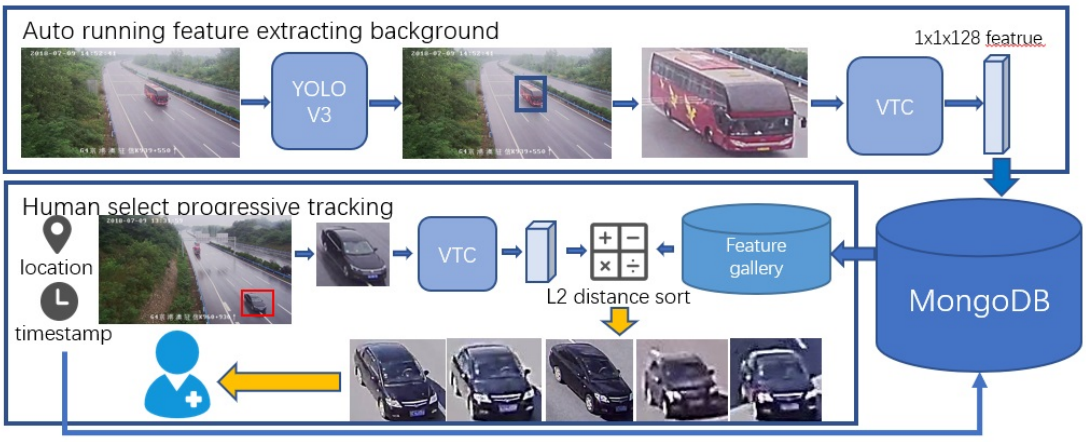

Fig. 4. Vehicle tracking framework

V. Experiments

To evaluate performance of multi-camera tracking, we compare our VTC method with base triplet loss network and feature concatenating network, the result listed in table.1. The vehicle tracking system has been successfully deployed in Beijing-Hong Kong-Macao Expressway.

TABLE I
Experiments Result

algorithm	Top1 Acc	Top5 Acc	FPS
Base Triplet Network	0.57	0.76	67
Concatenating Network	0.67	0.83	64
Vehicle Tracking Context	0.74	0.87	63

VI. Conclusion

We introduce VTC, a novel Re-ID approach for vehicle tracking. This method fuse features of same vehicle in camera, get better performance than previous method.

For real-time and active vehicle tracking, we propose an integrated tracking system including vehicle detection, vehicle Re-ID and vehicle search, The vehicle tracking system has been successfully deployed in Beijing-Hong Kong-Macao Expressway.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant 91748112. The authors would like to thank Wei Li, Xiuchuan Tang and Linan Deng in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology for helping in the experiments.

References

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[2] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[3] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.
[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
[5] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[6] Laroca R, Severo E, Zanlorensi L A, et al. A robust real-time automatic license plate recognition based on the YOLO detector[C]//2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-10.
[7] Laroca R, Severo E, Zanlorensi L A, et al. A robust real-time automatic license plate recognition based on the YOLO detector[C]//2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-10.
[8] Bromley J, Guyon I, LeCun Y, et al. Signature verification using a" siamese" time delay neural network[C]//Advances in neural information processing systems. 1994: 737-744.
[9] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.
[10] Sochor J, Špaňhel J, Herout A. Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 20(1): 97-108.
[11] Dong Y, Liu X, Huang B, et al. Deep Grasping Prediction with Antipodal Loss for Dual Arm Manipulators[C]//International Conference on Intelligent Robotics and Applications. Springer, Cham, 2019: 470-480.
[12] Liu X, Liu W, Ma H, et al. Large-scale vehicle re-identification in urban surveillance videos[C]//2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2016: 1-6.