Graph@FIT Submission to the NVIDIA AI City Challenge 2018

Jakub Sochor  Jakub Špaňhel  Roman Juránek  Petr Dobeš  Adam Herout
Graph@FIT, Brno University of Technology, Czech Republic
{isochor,ispanhel,ijuranek,idobes,herout}@fit.vutbr.cz

Abstract

In our submission to the NVIDIA AI City Challenge, we address speed measurement of vehicles and vehicle re-identification. For both these tasks, we use a calibration method based on extracted vanishing points. We detect and track vehicles by a CNN-based detector and we construct 3D bounding boxes for all vehicles. For the speed measurement task, we estimate the speed from the movement of the bounding box in the 3D space using the calibration. Our approach to vehicle re-identification is based on extraction of visual features from “unpacked” images of the vehicles. The features are aggregated in temporal domain to obtain a single feature descriptor for the whole track. Furthermore, we utilize a validation network to improve the re-identification accuracy.

1. Introduction

In this submission, we address the tasks of vehicle speed measurement and re-identification of the NVIDIA AI City Challenge (i.e. Task1 and Task3). We used our previously published approach to camera calibration [6, 31]. Also, for the re-identification task, we use the approach proposed in our paper [34], which is currently under review. In the paper, we show that using “unpacked” versions of vehicles [30, 33] improves the re-identification performance. Furthermore, in the paper, we propose a method for feature aggregation in temporal domain LTFD and Weighted Euclidean distance which we also use in our approach to this re-identification challenge task.

The vehicles are detected by a CNN-based detector, tracked in time and a 3D bounding box is constructed for every detected vehicle. These 3D bounding boxes are used for different purposes in the tasks. In the re-identification task, we use the 3D bounding box to produce “unpacked” version of the vehicle and normalize the input image. In the speed measurement task, the 3D bounding box is used to compute the center of the vehicle’s base, as it is a point which lies in the road plane; therefore, it can be used for speed measurement.

To put our approach to a larger context, we include a brief overview of the state of the art in camera calibration for speed measurement and vehicle re-identification. After that, we describe the used methods for both speed measurement and re-identification in detail.

2. Related Work

2.1. Camera Calibration for Vehicle Speed Measurement

Camera calibration (obtaining intrinsic and extrinsic parameters of the surveillance camera) is critical for the accuracy of vehicle speed measurement by a single monocular camera, as it directly influences the speed measurement accuracy. There is a very recent comprehensive review of the traffic surveillance calibration methods [32], so for detailed information we refer to this review and we include only a brief description of the methods.

Several methods [11, 2, 9] are based on the detection of vanishing points as an intersection of road markings (lane dividing lines). Other methods [6, 7, 27, 3] use vehicle motion to calibrate the camera. There is also a set of methods which use some form of manually measured dimensions on the road plane [24, 25, 29, 21, 22, 4, 17].

Finally, in our previous paper [31], we proposed an algorithm for camera calibration by detecting vanishing points using edgelets and scene scale estimation based on alignment of 3D models to recognized vehicles.

2.2. Vehicle Re-Identification

There are mainly two types of methods – methods based on automatic license plate recognition [5, 15, 38], which are not anonymous and require zoomed-in cameras. The other type of methods is based on vehicles’ visual appearance [1, 8, 42, 19] or on a combination of both approaches [20].

Formerly, different types of hand-crafted features were used. For example, authors used PCA-SIFT [1], HOG descriptors and color histograms [42], SIFT-BOW and Color Names model [19], or just information about date, time,
color, speed and vehicles’ dimensions [8]. Recently, deep features learned by CNNs [18, 28, 37, 39, 43] are being used for this task. Liu et al. [20] combine the hand-crafted and deep features.

Improvements were also made by exploiting spatio-temporal [20, 37] or visual-spatio-temporal [28] properties. Some of them benefit from Siamese CNNs for license plate verification [20] or vehicle image similarities [28]. Moreover, introduction of triplet loss [43] or Coupled Cluster Loss (CCL) [18] led to accuracy improvements and faster convergence. Recently, Yan et al. [39] propose to use Generalized Pairwise Ranking or Multi-Grain based List Ranking for retrieval of similar vehicles, which performs even better than CCL.

Also, in our paper [34], which is currently under review, we propose a method for feature aggregation in temporal domain of multiple observations of the vehicle in one track.

2.3. Vehicle Re-Identification Datasets

Datasets of vehicles are available [16, 41, 33], which are created for fine-grained recognition with annotations on several attributes such as type, make and color. However, the identities of the vehicles in the datasets are not known; thus, the datasets are not directly applicable for vehicle re-identification, especially for evaluation.

When it comes to genuine vehicle re-identification, Liu et al. [20] constructed a rather small VeRi-776 dataset containing 50,000 images of 776 vehicles. Liu et al. [18] collected VehicleID dataset containing 26,267 vehicles in 220k images taken from a frontal/rear viewpoint above road. Recently, Yan et al. [39] published two datasets VD1 and VD2 for vehicle re-identification and fine-grained classification with over 220k of vehicles in total, with make, model, and year annotation. However, both datasets are limited to frontal viewpoints only.

Recently, we collected dataset CarsReId30k [34], which contains ∼30k of vehicle tracks from various viewpoints with precise ground truth identity acquired from a zoomed-in camera by license plate recognition.

3. Used Approach

In our submission to the NVIDIA AI City Challenge 2018, we focused on vehicle speed measurement (Track1) and vehicle re-identification (Track3). In the following text,
our approach to both these tasks is described. However, we describe common processing steps of our approach to both these problems first.

### 3.1. Vehicle Detection and Data Preprocessing

As a first step of processing of the video, it is necessary to detect vehicles in all frames. We used Faster-RCNN\(^{26}\) with ResNet101\(^{10}\) backbone. The detector was trained on UA-DETRAC\(^{23}\) and COD20K\(^{12}\) datasets. The detections were merged to tracks using Kalman filter\(^{13}\). Examples of the detections can be found in Figure 1 (left).

Furthermore, in every video, we detected vanishing points using our recent algorithm\(^{31}\). Visualization of the detected vanishing points can be found in Figure 2 (top left).

For each detected vehicle, we used general object contour detector\(^{40}\) to estimate the contours\(^{33}\) of the vehicles (Figure 1 – center left and center right) and then constructed 3D bounding boxes of the vehicles\(^{6}\) (Figure 1 right). The 3D bounding box is used for two main purposes. First, it is possible to use it to normalize the image for vehicle fine-grained recognition\(^{30,33}\) and re-identification\(^{34}\). Second, it is possible to use it for estimation of a point on the road plane which can be used for speed measurement of the vehicles\(^{6,31}\).

### 3.2. Vehicle Speed Measurement

As we obtained the camera calibration in the form of 3 vanishing points, we are able to estimate intrinsic camera parameters (i.e. focal length) and rotation of the road plane with respect to the camera. However, one unknown remains and that is the scale of the scene (i.e. distance of the road plane from camera). Therefore, it is necessary to estimate this last parameter.

To estimate the scale, we used measurements on the road plane. We used Google Earth to measure the real world distance of two points (Figure 2 – top right). However, as we assume that the measurements will be imprecise we used a large number of measurements (\(~ 40\)) on the road plane to reduce the error.

The set of such measurements \(\mathcal{M}\) was divided into two groups. The first group \(\mathcal{M}_u\) contains measurements in the direction to the first vanishing point \(u\) (represented by red arrows in Fig. 2, top left). The other group \(\mathcal{M}_v\) is computed as \(\mathcal{M}_v = \mathcal{M} \setminus \mathcal{M}_u\).

As the second vanishing point \(v\) (green arrows in Fig. 2 – top left) is sometimes detected imprecisely (see horizon line in Figure 2 – top left), we further refine its position. To
achieve that, we optimize the following term

\[ v^* = \arg \min_v \sum_{p_1, p_2, m \in M_v} |m - d_{u,v,M_v}(p_1, p_2)|, \quad (1) \]

where \( d_{u,v,M_v}(p_1, p_2) \) represents the distance of two image points \( p_1 \) and \( p_2 \) in meters. It should be noted that the distance depends on the vanishing points and on the scene scale (which is computed using measurements \( M_v \)); therefore, we include all these variables in the lower index of the function. The final scene scale \( \lambda \) is computed using all measurements \( M \). For further details about the optimization and scene scale computation see our paper [31].

**Final calibrations** represented by the regular orthogonal grid can be found in Figure 3.

Using this calibration, it is possible to measure distance of two image points in the road plane in meters. Furthermore, with known video framerate, it is possible to measure elapsed time and thus compute the speed of the vehicles. To make the speed measurement stable, we measure the speed between detections 10 video frames apart.

As we need to measure the speed of a point in the road plane, we construct the center of the base of the 3D bounding box as the intersection of its diagonals. In order to further eliminate noise in the speed measurement, we represent short tracks by the median value. Longer tracks are filtered to suppress high frequencies in the speed signal. All measured speeds in all videos are shown in Figure 4.

### 3.3. Acquisition of Training Data for Vehicle Re-Identification

As the vehicle re-identification task contains vehicles observed from various viewpoints in video, it is necessary to acquire a similar dataset. We used our dataset CarsReId30k [34] which contains 8,343 unique vehicles, 29,676 observed tracks, and 92,846 positive pairs.

The dataset was collected using 8 cameras recording at the same time. Four cameras always observed the same location from different viewpoints (left, center, right), and one
The CarsReId30k dataset [34] contains vehicles from different viewpoints but only from the frontal sides of vehicles. Therefore, we expanded the dataset and added videos from viewpoints which observe the rear side of the vehicles from different viewpoints in the same manner. The final dataset contains 17,681 unique vehicles, 73,976 observed tracks, and 292,226 positive pairs. For examples of positive and negative pairs, see Figure 6.

![Figure 6. Examples of queries, positive, and negative samples. The negatives are sorted by difficulty from left to right (hard to easy) based on distances obtained from our re-identification feature vectors. It should be noted that the hardest negative sample has usually subtle differences (e.g. missing a small spoiler in the first row). Image source: [34].](image)

3.4. Approach to Vehicle Re-Identification

Following the methodology from our previous paper [34], we first **fine-tuned the CNN** on vehicle identification task. We used InceptionResNetV2 [36] with “unpacked” version of vehicle images [33] and the input image size was $331 \times 331$. The fine-tuning was done with Adam [14] optimizer and learning rate $1e^{-4}$.

Afterwards, we cached the identification features and trained **LFTD network** [34] to aggregate the features in temporal domain as there are multiple observations for the vehicle as they pass in front of the cameras. The LFTD network contains one fully connected layer with 1,024 output features and tanh non-linearity. Furthermore, the network contains feature weighting mechanism which weights different elements of the feature vectors by different weights. The network is trained as a Siamese network.

Furthermore, we used **Weighted Euclidean distance** which is expressed as

$$d_{WE}(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{D} w_i (u_i - v_i)^2}, \quad (2)$$

where $\mathbf{u}, \mathbf{v}$ are feature vectors and $\mathbf{w} = [w_1, w_2, \ldots, w_D]$ are learned weights.

We kept queries from two recording sessions (nine recording sessions for training) for validation. The Hit@1 with identification features and average temporal pooling was 69%. It increased to 78% with the LFTD network and standard Euclidean distance. Finally, using the weighted
Euclidean distance pushed the performance on validation data to 79%.

These 1,024-dimensional features and learned weighted Euclidean distance were used for computation of **pairwise distance** between all detected and tracked vehicles at every location. We used these distances to construct quadruplets with one vehicle track from every location. These quadruplets are supposed to represent vehicle tracks with the same identity of the vehicles. As the vehicle needs to be observed at every location, we used the maximal distance of pairs within the quadruplet to measure the total similarity of the quadruplet. Therefore, we are interested only in the quadruplets with low maximal distance of pairs in the quadruplet.

To further improve the vehicle re-identification accuracy, we trained a **validation neural network** which takes two images of vehicles and produces the probability that they have the same identity. We used the two validation sessions from our dataset for the training and the hardest negative pairs were used as negative samples during the training. Schematic design of the validation network can be found in Figure 7.

We used this validation network with every pair in the quadruplets and used median and minimal probability values as a threshold for the quadruplets and removed the ones with the probabilities below the thresholds.

Finally, we used these sorted and filtered quadruplets as the identities. We took only five best scoring quadruplets to limit false positives. As the vehicles could be observed multiple times at every location, we merged these five quadruplets with any high scoring quadruplet containing at least once the same vehicle. This way, we acquire sets of vehicles with assumed same identity as required by the task.
rules. For examples of such sets see Figure 8.

4. Conclusions

We participated in two tasks of the NVIDIA AI City Challenge 2018: the speed measurement task and vehicle re-identification task. Our solution is using camera calibration by extracting vanishing points and scene scale calibration by measuring distances in orthophoto maps. The camera calibration is used for speed measurement but also for constructing 3D bounding boxes around the vehicles (their contours) which help normalizing and “unpacking” the images for further processing by neural nets. The re-identification task is using these preprocessed images, makes pair-wise comparisons of detected vehicles, constructs quadruplets of vehicles appearing at all the given locations and verifies these sets by another neural net. The neural networks have been trained on our previously collected data sets collected for vehicle surveillance.

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II); project IT4Innovations excellence in science - LQ1602. Also, this work was supported by TACR project “SMARTCarPark”, TH03010529.

References


