Computer Vision and Image Understanding 161 (2017) 87–98

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Traffic surveillance camera calibration by 3D model bounding box alignment for accurate vehicle speed measurement

Jakub Sochor[1,*], Roman Juránek, Adam Herout

Brno University of Technology, Faculty of Information Technology, Centre of Excellence IT4Innovations, Božetěchova 2, 612 66 Brno,Czech Republic

A R T I C L E I N F O

Article history:
Received 3 December 2016
Revised 24 May 2017
Accepted 29 May 2017
Available online 1 June 2017

Keywords:
Speed measurement
Camera calibration
Fully automatic
Traffic surveillance
Bounding box alignment
Vanishing point detection

A B S T R A C T

In this paper, we focus on fully automatic traffic surveillance camera calibration, which we use for speed measurement of passing vehicles. We improve over a recent state-of-the-art camera calibration method for traffic surveillance based on two detected vanishing points. More importantly, we propose a novel automatic scene scale inference method. The method is based on matching bounding boxes of rendered 3D models of vehicles with detected bounding boxes in the image. The proposed method can be used from arbitrary viewpoints, since it has no constraints on camera placement. We evaluate our method on the recent comprehensive dataset for speed measurement BrnoCompSpeed. Experiments show that our automatic camera calibration method by detection of two vanishing points reduces error by 50% (mean distance ratio error reduced from 0.18 to 0.09) compared to the previous state-of-the-art method. We also show that our scene scale inference method is more precise, outperforming both state-of-the-art automatic calibration method for speed measurement (error reduction by 86 % – 7.98 km/h to 1.10 km/h) and manual calibration (error reduction by 19 % – 1.35 km/h to 1.10 km/h). We also present qualitative results of the proposed automatic camera calibration method on video sequences obtained from real surveillance cameras in various places, and under different lighting conditions (night, dawn, day).

1. Introduction

Surveillance systems pose specific requirements on camera calibration. Their cameras are typically placed in hardly accessible locations and optics are focused at longer distances, making the common pattern-based calibration approaches unusable (such as classical (Zhang, 2000)). That is why many solutions place markers to the observed scene and/or measure existing geometric features (Do et al., 2015; Luvizon et al., 2016; Sina et al., 2013; You and Zheng, 2016). These approaches are laborious and inconvenient both in terms of camera setup (manually clicking on the measured features in the image) and in terms of physically visiting the scene and measuring the distances.

In our paper, we focus on *precise* and at the same time *fully automatic* traffic surveillance camera calibration including scene scale for speed measurement. The proposed speed measurement method needs to be able to deal with significant viewpoint variation, different zoom factors, various roads and densities of traffic. If the method should be applicable for large-scale deployment, it needs to run fully automatically without the necessity to stop traffic for installation or for performing calibration measurements.

Our solution uses camera calibration obtained from two detected vanishing points and it is built on our previous work (Dubská et al., 2015, 2014). However, this calibration procedure only allows reconstruction of the rotation matrix and the intrinsic parameters from vanishing points, and it is still necessary to obtain the scene scale. We propose to detect vehicles on the road by Faster-RCNN (Ren et al., 2015), classify them into a few common fine-grained types by a CNN (Krizhevsky et al., 2012) and use bounding boxes of 3D models for the known classes to align the detected vehicles. The vanishing point-based calibration allows for full reconstruction of the viewpoint on the vehicle and the only free parameter in the alignment is therefore the scene scale. Fig. 1 shows an example of the 3D model and the aligned images. Our experiments show that our method (mean speed measurement error 1.10 km/h) significantly outperforms existing automatic camera calibration method by Dubská et al. (2014) (error reduction by 86 % – mean error 7.98 km/h) and also calibration obtained from manual measurements on the road (error reduction by 19% – mean error 1.35 km/h). This is important because in previous approaches, automation always compromised accuracy, forcing a trade off by the system developer. Our work shows that fully automatic calibra-

* Corresponding author.
E-mail addresses: isochor@fit.vutbr.cz (J. Sochor), ijuranek@fit.vutbr.cz (R. Juránek), herout@fit.vutbr.cz (A. Herout).

[1] Jakub Sochor is a Brno Ph.D. Talent Scholarship Holder — Funded by the Brno City Municipality.

http://dx.doi.org/10.1016/j.cviu.2017.05.015
1077-3142/

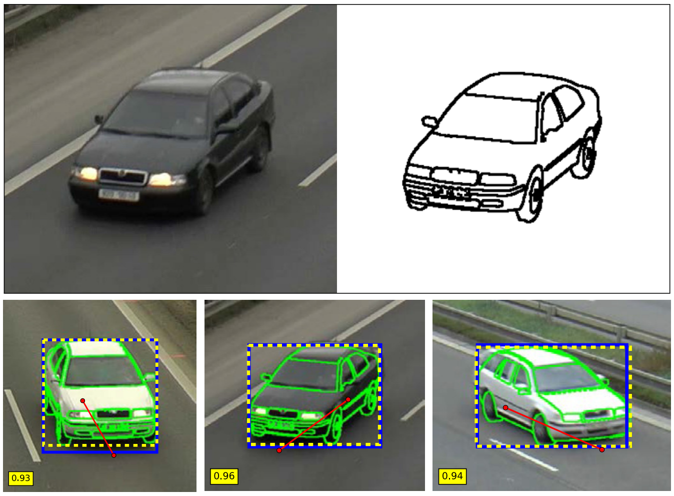

Fig. 1. Examples of detected vehicles and 3D model bounding box aligned to the vehicle detection bounding box. **Top:** detected vehicle and corresponding 3D model (edges only), **bottom:** examples of aligned bounding boxes with shown 3D model edges (green), its bounding box (yellow) and vehicle detection (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tion methods may produce better results than manual calibration (which was performed thoroughly and according to state-of-the-art approaches).

Existing solutions for traffic surveillance camera calibration (Cathey and Dailey, 2005; Dailey et al., 2000; Do et al., 2015; Dubská et al., 2015, 2014; Grammatikopoulos et al., 2005; He and Yung, 2007b; Lan et al., 2014; Luvizon et al., 2014, 2016; Maduro et al., 2008; Nurhadiyatna et al., 2013; Schoepflin and Dailey, 2003; Sina et al., 2013; You and Zheng, 2016) (see Section 2 for detailed analysis) usually have limitations for real world applications. They are either limited to some viewpoints (zero pan, second vanishing point at infinity), or they require some per-installed-camera manual work. To our knowledge, there is only one work (Dubská et al., 2014) which does not have these limitations, and therefore we compare our results with this solution. For a brief description of the method, see Section 2; a more comprehensive review can be found in a recent dataset paper BrnoCompSpeed by Sochor et al. (2016b).

The key contributions of this paper are:

- An improved camera calibration method by detection of two vanishing points. The camera calibration error is reduced by 50% – 0.18 to 0.09 mean distance ratio error.
- A novel method for scene scale inference, which significantly outperforms automatic traffic camera calibration methods (error reduced by 86% – 7.98 km/h to 1.10 km/h) and also manual calibration (error reduced by 19% – 1.35 km/h to 1.10 km/h) in automatic speed measurement from a monocular camera.
- Results show that when used for the speed measurement task, the automatic (zero human input) method can perform better than the laborious manual calibration, which is generally considered accurate and treated as the ground truth. This finding can be important also in other fields beyond traffic surveillance.

2. Related work

The camera calibration algorithm (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 (Sochor et al., 2016b), so for detailed information we refer to this review and we include only a brief description of the methods.

Several methods (Cathey and Dailey, 2005; Grammatikopoulos et al., 2005; He and Yung, 2007b) are based on the detection of vanishing points as an intersection of road markings (lane dividing lines). Other methods (Dailey et al., 2000; Dubská et al., 2015, 2014; Schoepflin and Dailey, 2003) 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 (Do et al., 2015; Lan et al., 2014; Luvizon et al., 2014, 2016; Maduro et al., 2008; Nurhadiyatna et al., 2013; Sina et al., 2013).

An important attribute of calibration methods is whether they are able to work automatically without any manual per-camera calibration input. Only two methods (Dailey et al., 2000; Dubská et al., 2014) are fully automatic and both of them use mean vehicle dimensions for camera calibration. Another important requirement for real-world deployment is whether the camera can be placed in an arbitrary position above the road, which is not true for some methods as they assume to have zero pan or other constraints.

Regarding fine-grained vehicle classification, there are several approaches. The first one is based on detected parts of vehicles (Fang et al., 2016; Krause et al., 2015; Simon and Rodner, 2015), another approach is based on bilinear pooling (Gao et al., 2016; Lin et al., 2015). There is also an approach based on Convolutional Neural Networks (CNN) and input modification (Sochor et al., 2016a). For object detection, it is possible to use boosted cascades (Dollár et al., 2014), HOG detectors (Dalal and Triggs, 2005), or Deformable Parts Models (DPMs) (Felzenszwalb et al., 2010). There are also recent advances in object detection based on CNNs (Girshick et al., 2014; Liu et al., 2016; Ren et al., 2015).

Several authors deal with alignment of 3D models and vehicles and use this technique for gathering data in the context of traffic surveillance. Lin et al. (2014) propose to jointly optimize 3D model fitting and fine-grained classification, and Hsiao et al. (2014) align edges formulated as an Active Shape Model (Cootes et al., 1995; Li et al., 2009). Krause et al. (2013) and propose to use synthetic data to train geometry and viewpoint classifiers for 3D model and 2D image alignment.Prokaj and Medioni (2009) use detected SIFT features (Lowe, 1999) to align 3D vehicle models and the vehicle's observation. They use the alignment mainly to overcome vehicle appearance variation under different viewpoints. However, in our case, as the precise viewpoint on the vehicle is known (Section 4.3), such alignment does not have to be performed. Hence, we adopt a simpler and more efficient method based on 2D bounding boxes – simplifying the procedure considerably without sacrificing the accuracy.

When it comes to camera calibration in general, various approaches exist. The widely used method by Zhang (2000) uses a calibration checkerboard to obtain intrinsic and extrinsic camera parameters (relative to the checkerboard). Liu et al. (2012) use controlled panning or tilting with stereo matching to calibrate the camera. Correspondences of lines and points are used by Chaperon et al. (2011). Yu et al. (2009) focus on automatic camera calibration for tennis videos from detected tennis court lines.

3. Traffic camera model

The main goal of camera calibration in the application of speed measurement is to be able to measure distances on the road plane between two arbitrary points in meters (or other distance units), therefore we only focus on a camera model which enables the measurement of distance between two points on the road plane.

For convenience and better comparison of the methods, we adopt the traffic camera model and notation proposed in previous papers (Dubská et al., 2015, 2014); however, to make the paper