Practical Camera Calibration From Moving Objects for Traffic Scene Surveillance

Zhaoxiang Zhang, *Member, IEEE,* Tieniu Tan, *Fellow, IEEE,* Kaiqi Huang, *Senior Member, IEEE,* and Yunhong Wang, *Member, IEEE*

***Abstract*—We address the problem of camera calibration for traffic scene surveillance, which supplies a connection between 2-D image features and 3-D measurement. It is helpful to deal with appearance distortion related to view angles, establish multiview correspondences, and make use of 3-D object models as prior information to enhance surveillance performance. A convenient and practical camera calibration method is proposed in this paper. With the camera height *H* measured as the only user input, we can recover both intrinsic and extrinsic parameters of the camera based on redundant information supplied by moving objects in monocular videos. All cases of traffic scene layouts are considered and corresponding solutions are given to make our method applicable to almost all kinds of traffic scenes in reality. Numerous experiments are conducted in different scenes, and experimental results demonstrate the accuracy and practicability of our approach. It is shown that our approach can be effectively adopted in all kinds of traffic scene surveillance applications.**

***Index Terms*—3-D recovery, camera calibration, object classification, traffic scene, vanishing point, visual surveillance.**

I. Introduction

CAMERA CALIBRATION, as a fundamental topic in computer vision, is meant to recover both intrinsic and extrinsic parameters of camera models. Camera calibration is not only essential for classical computer vision problems, such as stereo, metrology, reconstruction, and virtual reality, but also can benefit other practical applications, such as traffic scene surveillance. First, camera calibration supplies a connection between image plane and 3-D measurement. It can help to deal with appearance distortion related to view angles, which is a very difficult problem to solve for 2-D image feature-based methods [1]. Second, with cameras calibrated, we can establish correspondences among different cameras so that multiview information can be fused to enhance surveillance performance [2]. Furthermore, calibrated cameras make it possible to adopt prior information of 3-D models to make object detection or tracking more robust to noise and occlusions [3]–[5].

The importance of camera calibration has motivated much work in the past decades, with all kinds of approaches proposed. The common practice to achieve camera calibration in traffic scenes is to collect a large set of correspondences between 3-D points and their projections on image plane [6], [7]. However, a time-consuming wide site survey is always required, and it is difficult to measure 3-D points that do not lie on the ground plane in wide surveillance scenes. The often planar distribution of sampling points around the ground plane inevitably leads to inaccuracy of camera calibration. Alternative strategies are to adopt additional reference objects for calibration. Tsai [8] made use of 3-D known metric structures to undergo a known translation to obtain constraints. Zhang [9], [10] further weakened the conditions to achieve calibration by 2-D or even 1-D template. However, reference object-based calibration is also not suitable for traffic scene surveillance. The requirement of templates limits the practicability and applicability of calibration algorithms. Furthermore, since surveillance cameras always cover a very wide traffic scene, projections of reference objects are of very small size in images, which would definitely effect the accuracy of camera model estimation.

Autocalibration methods seem to be a more suitable way to recover camera models for traffic scene surveillance. Since surveillance applications prefer to make use of static cameras, autocalibration cannot be achieved from camera motion but may be done from inherent structures of monocular scenes. Linear or planar patterns of scene components can be adopted to derive constraints for calibration. Li *et al.* [11] made use of planar information in scenes to compute five intrinsic parameters linearly. Triggs [12] achieved camera calibration by involving absolute quadric and collineations between scene plane and image plane. Caprile and Torre [13] described methods for using vanishing points to recover intrinsic parameters from a single camera but extrinsic parameters from multicameras. Cipolla *et al.* [14] proposed a method for recovering both intrinsic and extrinsic parameters by three vanishing points and two reference points from two view angles. Liebowitz *et al.* [15] developed a method for estimating intrinsic parameters by Cholesky decomposition and applied it to a scene reconstruction problem. Deutscher *et al.* [16] made use of vanishing points in a Manhattan world to recover camera parameters for

Manuscript received February 1, 2012; revised May 6, 2012; accepted June 12, 2012. Date of publication July 27, 2012; date of current version March 7, 2013. This work was supported in part by the National Basic Research Program of China, under Grant 2010CB327902, and by the National Natural Science Foundation of China, under Grant 61005016. This paper was recommended by Associate Editor S. Battiato.

Z. Zhang and Y. Wang are with the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: zxzhang@buaa.edu.cn; yhwang@buaa.edu.cn).

T. Tan and K. Huang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: tnt@nlpr.ia.ac.cn; kqhuang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TCSVT.2012.2210670

visual tracking. Overall, prior information of scene structures is definitely useful in simplifying camera calibration. However, scene structures are variant to different scenes and not always available in reality.

In the absence of inherent scene structures, the methods described above may not work. Researchers further make use of object motions in videos to replace scene structures. Most existing work in this direction focuses on shape and motion information of pedestrians in videos. Krahnstoever *et al.* [17] proposed methods for tracking pedestrians in surveillance videos and for recovering camera models in a Bayesian framework. Junejo *et al.* [18] recognized harmonic homologies from observing pedestrians to obtain linear constraints on camera parameters. Lv *et al.* [19] obtained three orthogonal vanishing points by extracting head and foot positions of humans in videos under the assumptions of constant human height and planar human motion. Micusik and Pajdla [20] proposed an approach to achieve camera calibration from foot-head homology, which has achieved convincing performance. A similar strategy has been adopted recently by Evans and Ferryman [21] to achieve camera calibration from observation of a pedestrian. However, most of the above methods require not only an accurate pedestrian detection, but also precise foot-head localization. As we know, precise pedestrian detection is very difficult in surveillance videos due to noise and shadows. Foot-head detection would be even more challenging in low resolution traffic scene surveillance videos. The inaccuracy of foot-head detection and homology may affect the accuracy of camera calibration.

In addition, there is much work focusing on rectification or normalization of the ground plane since most objects of interest in traffic scene surveillance move on the ground plane. Junejo *et al.* [22], [23] achieved trajectory rectification and path modeling for video surveillance. Bose *et al.* [24] tracked vehicles and detected constant velocity linear paths to realize ground plane rectification instead of recovering intrinsic and extrinsic camera parameters. Stauffer *et al.* [25] normalized the appearance features of objects on the ground plane under an inaccurate linear assumption. These methods are useful for surveillance applications such as classification or tracking, but cannot recover accurate intrinsic and extrinsic parameters of camera models.

On one hand, camera calibration is urgently necessary for traffic scene surveillance applications, such as view-independent object classification, object matching or tracking among multicameras, and 3-D model based object recognition. On the other hand, most of the existing algorithms are often not suitable and practical in wide-area surveillance scenes. It is much desired to develop more practical camera calibration methods applicable to wide-area surveillance scenes. We think to explore the particular and inherent properties of traffic scenes should be the key to solve this problem.

In this paper, we propose a novel method for accurate and practical camera calibration from traffic scene surveillance videos based on the above consideration. With moving objects extracted from videos using motion information, prior information of traffic scenes is adopted to estimate three vanishing points corresponding to three orthogonal directions in the 3-D world based on motion and appearance information of objects in videos. With camera height H measured as the only user input, we can recover both intrinsic and extrinsic camera parameters, which is of great help for many surveillance applications. Furthermore, our method is proved to be applicable to all kinds of traffic scene layouts. Extensive experiments are conducted to evaluate the performance of the calibration algorithm in traffic scenes of different view angles. We also show potential applications of our approach in 3-D recovery from images, automatic object classification, and 3-D model-based object recognition. Experimental results demonstrate the accuracy and practicability of our approach. An earlier and much shorter version of this paper appeared in [26].

(a) (b) (c)

Fig. 1. Motion detection results with shadows removed. (a) One frame of the video. (b) Background recovered. (c) Detected moving objects.

The remainder of this paper is organized as follows. In Section II, we introduce our method for extracting accurate foreground areas with shadows removed. The strategies to estimate three orthogonal vanishing points from motion and appearance of video objects for different traffic scene layouts are proposed in Section III. In Section IV, we introduce our method for realizing calibration from three orthogonal vanishing points and the camera height H as the only one user input. Experimental results and analyses are given in Section V. We draw our conclusions in Section VI.

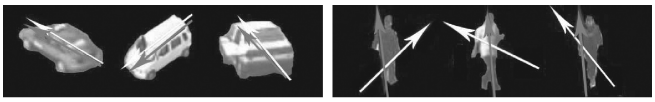
II. Motion Detection

Motion and appearance of moving objects in surveillance videos supply plentiful information for camera calibration. In this section, we introduce our method for the extraction of accurate foreground areas with shadows removed. As we know, Gaussian mixture model [27] is a popular method in the field of motion detection due to its outstanding ability to deal with slow lighting changes, periodical motions in cluttered background, slow moving objects, and long-term scene changes. However, this method still has disadvantages as it cannot deal with fast illumination changes and shadows very well, which are very common in traffic scene surveillance. In our work, we adopt the method described in [28] to deal with the disadvantages mentioned above.

Experimental results of background maintenance and motion detection are shown in Fig. 1. As we can see, foreground objects are detected accurately with cast shadows removed.

III. Vanishing Point Estimation

A vanishing point is defined as the intersection of a series of projected parallel 3-D lines in image plane, which is very useful for camera autocalibration [7]. In this section, we



(a) (b)

Fig. 2. Illustrations of the two orientations at different view angles. White arrowhead: velocity orientation. Gray arrowhead: main axis orientation. (a) Illustration of two orientations for vehicles. (b) Illustration of two orientations for pedestrians.

propose our method for estimating three orthogonal vanishing points from the appearance and motion of moving objects in outdoor traffic scenes.

As we know, road traffic scenes have a number of helpful general properties for the estimation of vanishing points, which are summarized as follows:

1) almost all moving objects including vehicles and pedestrians are moving on the ground plane;
2) vehicles always run along the roadway that is often straight or has a series of straight segments in the field of view in most cases;
3) projections of vehicles on images are rich in line segments along two orientations, which correspond to the symmetrical axis direction of the vehicle and its perpendicular direction in most view angles;
4) in most cases, pedestrians are walking with their trunks perpendicular to the ground plane.

These four properties are found in most road traffic scenes and they can be used to estimate three orthogonal vanishing points, which are described in detail as follows. Here, we first consider the simplest case, in which there is only one straight roadway in the view field. Of course, the roadway layouts in reality have large variance and are not always of this simple case. Solutions of different road layouts will be extended from the simplest case and given at the end of this section.

A. Coarse Moving Object Classification

As we know, pedestrians and vehicles are two types of most common objects of interest in surveillance videos. They have different appearance and motion properties, which can be applied for classification. Two kinds of orientations are extracted from every moving object detected from videos to achieve coarse classification. The first one is the velocity orientation in the image plane, which is calculated as follows. Assuming the position of one moving object in image plane is (X_1, Y_1) in Frame $(t - \Delta t)$ and (X_2, Y_2) in Frame t, the velocity orientation θ_V in Frame t is estimated as follows:

$$\theta_V = \arctan\left(\frac{Y_2 - Y_1}{X_2 - X_1}\right) \quad \theta_V \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \tag{1}$$

In practice, we select Δt from [10, 15]. The second orientation is the main axis direction θ_M of silhouettes of the detected object, which can be estimated from moment analysis as follows:

$$\theta_M = \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \quad \theta_M \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \tag{2}$$

where μ_{pq} is the central moment of order (p, q). The difference between these two orientations is then calculated as follows:

$$\Delta\theta = \begin{cases} |\theta_V - \theta_M|, & \text{if } |\theta_V - \theta_M| \le \frac{\pi}{2} \\ \pi - |\theta_V - \theta_M|, & \text{else.} \end{cases} \tag{3}$$

The orientation difference supplies coarse category information. As we know, the two orientations have quite significant difference for pedestrians moving in videos, while they are very close to each other for vehicles in most cases of camera views, as illustrated in Fig. 2. As a result, we take $\Delta\theta$ as the discriminant feature for coarse classification. K-mean clustering seems to be a good method for unsupervised classification. However, large view angle variance in wide-area videos may lead to serious misclassification. In fact, we do not need to care about the classification accuracy of all objects appearing in videos, but can select a subset in which we can achieve reliable classification. Based on this consideration, we can set two rigorous thresholds θ_1 and θ_2 ($\theta_1 < \theta_2$) to achieve this aim. The object is labeled as a vehicle if $\Delta\theta < \theta_1$ and as a pedestrian if $\Delta\theta > \theta_2$. Those objects whose orientation difference is between θ_1 and θ_2 are discarded. With our collected moving objects in all kinds of traffic scenes shown in Section V, the true positive rates of vehicles and pedestrians are shown in Fig. 3. Based on the analysis of this curve, we assign θ_1 as $\frac{1}{36}\pi$ (5°) and θ_2 as $\frac{1}{9}\pi$ (20°) to ensure that the true positive rates of pedestrians and vehicles are more than 97%. The subsequent estimation of vanishing points benefits from this strict selection and classification strategy.

B. Line Equation Estimation

Vanishing points should be estimated from images in two steps. First, we should detect a series of image-projected parallel lines. Second, we estimate the intersections of these projected parallel lines in the image plane. The four general properties in road traffic scenes, we summarized before, supply important information for projected parallel line detection.

In the simplest case of only one straight roadway in 3-D scenes, most vehicles are running along the roadway (general property 2) so that they should be running in the same or inverse 3-D direction. In this case, it can be derived that the symmetrical axes of these vehicles should be parallel to each other, which are also parallel to the ground plane. Of course, those lines perpendicular to the symmetrical axes of respective vehicles should also be parallel to each other.

As we have described before, image projections of vehicles are rich in line segments that correspond to the symmetrical axis direction and its perpendicular direction in the 3-D world (general property 3). As a result, these line segments can be extracted and gathered into two line sets of image projected parallel lines. Each set can be applied to estimate one horizontal vanishing point. Furthermore, the two estimated vanishing points from these two sets should be orthogonal to each other.

For accuracy, we only estimate two line equations for every object labeled as a vehicle. One is parallel to the symmetrical

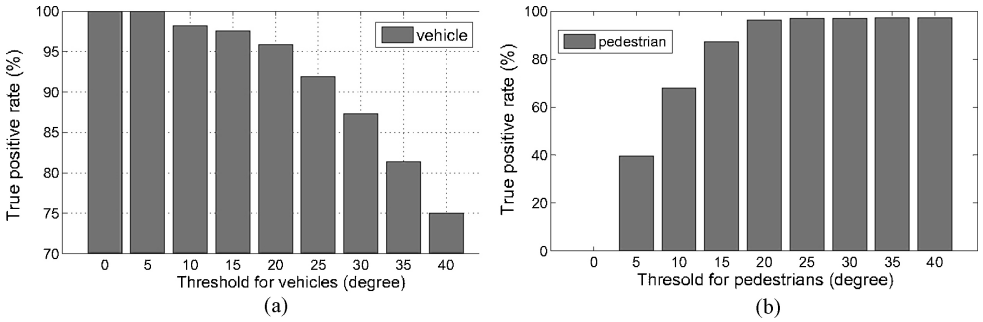


Fig. 3. True positive rate curve of (a) vehicles and (b) pedestrians.

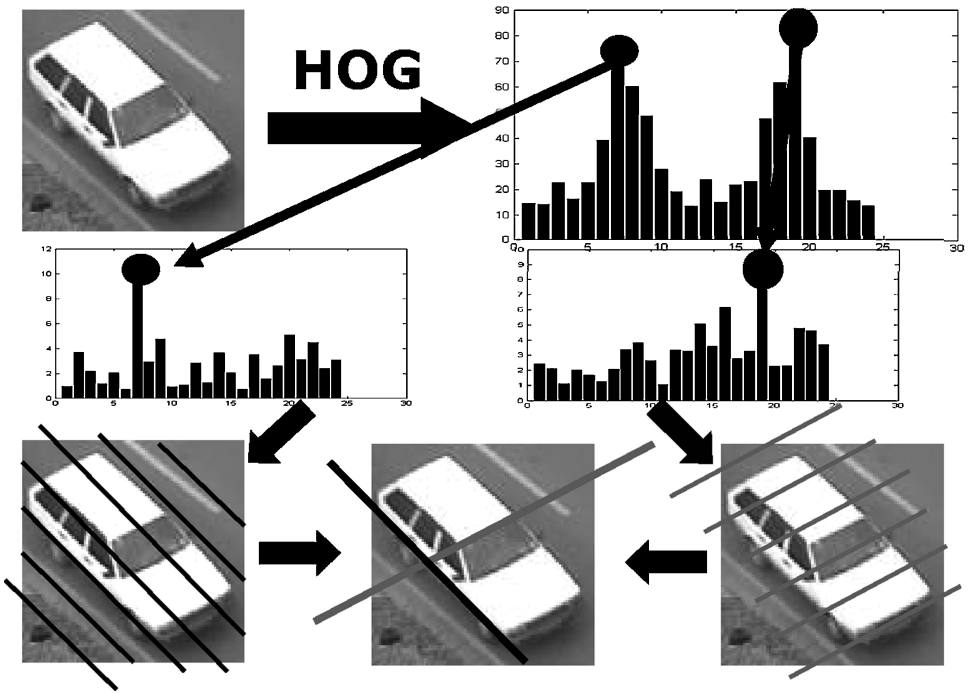
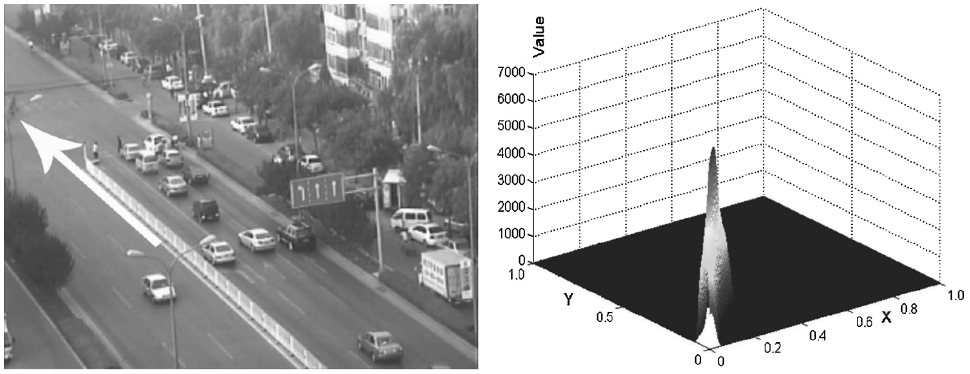


Fig. 4. Flowchart of estimation of line equations for vehicles [26].

axis, while the other is perpendicular to it. Instead of sensitive edge point detection and combination into edge lines, we first estimate the two orientations of estimated lines based on histogram of orientated gradient (HOG) [29] in two stages. For every moving region labeled as vehicles detected from videos, the gradient magnitude and orientation are computed at every pixel within it. The orientation is divided into N bins and the histogram is formed by accumulating orientations within the region, weighted by the gradient magnitude. The two bins with the largest values are chosen as coarse line orientations and an N bin HOG is calculated in each bin again to determine accurate line orientation, respectively. For every orientation accurately estimated, the line with this orientation slides from the top to the bottom of the region to determine its position in which the line most fits image data by correlation. An example is illustrated with line equations determined, as shown in Fig. 4.

Every detected object labeled as a vehicle can supply two line equations corresponding to two orthogonal 3-D directions. Motion direction can be adopted here to distinguish these two estimated lines. The line with its orientation closer to the motion direction corresponds to the symmetric axis direction of the vehicle, while the other corresponds to the perpendicular direction.

Vehicles in surveillance videos supply two sets of projected parallel lines to estimate two orthogonal horizontal vanishing lines. Pedestrians do not have so significant gradient orientations as vehicles. However, as we know, most pedestrians are walking with their trunks perpendicular to the ground plane in most situations (general property 4). Instead of localizing head and foot position in a small region, we use line equations to describe pedestrian trunks. With moving blobs labeled as pedestrians, the line is determined by its orientation as the main axis orientation of the silhouettes and requiring the line passing by the centroid of the silhouettes. This strategy is more accurate and robust to noise and view angle changes. All these extracted lines from pedestrians correspond to the orientation perpendicular to the ground plane so that they should be parallel to each other in the 3-D world. In this way, another set of projected parallel lines is estimated from pedestrians in videos, which can be applied to estimate one vertical vanishing point.



(a) (b)

Fig. 5. Illustration of estimating vanishing points from traffic scenes [26]. (a) Illustration of a traffic scene. (b) Voting surface.

C. Intersection Estimation

As we have described above, there are three line sets estimated from detected objects in videos in the simplest case. The first set contains lines corresponding to the symmetric axis direction of vehicles in reality. The second set contains lines corresponding to the perpendicular direction of the symmetric axis. The third set contains lines corresponding to the perpendicular direction of the ground plane.

With abundant objects detected from videos and labeled as vehicles and pedestrians, the three line sets are becoming larger and larger. Ideally, lines of every set should intersect at the same point called vanishing points. However, due to the large portion of outliers and noise in videos, lines do not intersect at the same point at all. Various approaches can be adopted here for robust estimation of the intersection point from redundant line equations. The simplest way is to solve simultaneous line equations based on the least-squares strategy. Also, the problem can be transformed to estimate a point the sum of whose distances to all lines is minimal. This is a classical optimization problem which can be solved by the Levenberg–Marquardt method. In addition, RANSAC is another strategy to solve this problem, which has been adopted in [19].

In spite of the accuracy and robustness supplied by the above methods, they are not suitable to our case. With videos processed frame by frame and more and more lines collected into line sets, we hope that our algorithm should dynamically estimate the vanishing points on-the-fly without repeated calculations.

We adopt a voting-based strategy to achieve this aim, which is a widely used choice in all kinds of applications. Tan *et al.* [4] made use of voting to estimate the most reasonable pose of 3-D models. Cham *et al.* [30] made use of a voting strategy to achieve optimal camera pose estimation. Battiato *et al.* [31] proposed a robust alignment method based on voting in parameter spaces to recover geometric transformations. Here, we make use of voting based on the fact that every point on the line is a possible candidate of the intersection point. The possibility is assumed to satisfy a Gaussian distribution on the neighborhood due to the distance to the point. As a result, for every line l extracted from objects in videos, each point $s(x, y)$ lying on l generates a Gaussian impulse in the voting space with (x, y) as its center. With time accumulated, a voting surface can be generated and the position of its global extreme corresponds to the estimated intersection of these lines. Compared to other estimation methods, this strategy can estimate the positions of vanishing points on-the-fly without repeated calculations. Compared to traditional voting-based methods, this strategy supplies more conspicuous global extreme, smoother surface, and is more robust to noise and outliers. One example of estimation of the vanishing point from voting surface is shown in Fig. 5.

Two line sets generated from vehicles are taken to estimate two orthogonal horizontal vanishing points, while the other set generated from pedestrians is taken to estimate one vertical vanishing points. In this way, we can extract three orthogonal vanishing points (u_1, v_1), (u_2, v_2), and (u_3, v_3) from appearance and motion information of moving objects in traffic scene surveillance videos in the case of only one straight roadway in the view field.

D. Special Cases

We have discussed solutions to estimate three orthogonal vanishing points from traffic scene surveillance videos under the assumption of only one straight roadway in the view field. However, this assumption is not always true in reality. There may be more than one roadway in the view field, like a crossroad. The roadway may not be straight at all, like a bend. In these cases, the method described above cannot be applied directly to estimate the two horizontal vanishing points. Most roadway layouts in reality can be seen as combinations of several primitive layouts. In the following, we will describe solutions to all these primitive layouts, so that our work is applicable to almost all kinds of layouts in reality.

1) *Straight Segments:* In reality, it is very common that the whole roadway is not straight, but contains a series of straight segments. Fig. 6(a) shows a roadway containing one inflexion which divides the whole roadway into two straight segments. Fig. 6(b) shows the case of two inflexions dividing the whole roadway into three straight segments. Fig. 6(c) shows a roadway which is composed of a straight segment and a bend. The roadway in all these cases contains at least one straight segment, which is enough to be applied for horizontal vanishing points estimation.

The detection of straight segments can be simply realized by motion information. With objects extracted by motion detection and classified as vehicles, conventional tracking can

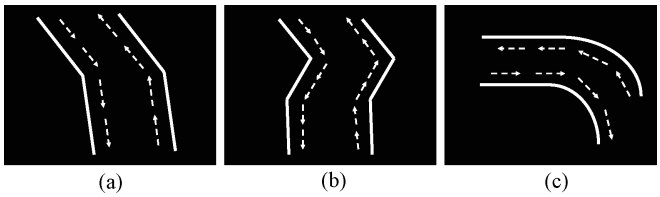


Fig. 6. Cases of straight segments. (a) Case 1. (b) Case 2. (c) Case 3.

Algorithm 1 Straight segment detection

```
Motion detection and vehicle tracking
for all Tracked vehicles do
    if motion direction of subseries with divergence < δ1 then
        mark passing regions as straight segment candidate once;
    end if
    track direction = avg (motion directions)
end for
if track direction variance < δ2 and marked as straight segment candidate more than N times then
    merge neighbor regions as straight segments
end if
straight segment direction = avg (motion directions within the region)
```

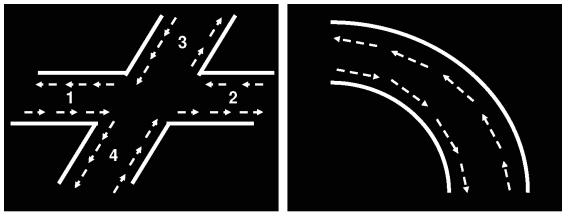


Fig. 7. (a) Case of a crossroad. (b) Case of a bend.

help us to monitor the change trend of velocity directions. As we know, the velocity direction of a vehicle changes little in a straight segment, but has evident change at the inflexion or the bend part. Regions with similar vehicle motion directions are merged to generate straight segments. We give detailed algorithms to detect straight segments in Algorithm 1.

In this way, we can detect all straight segments from scenes and each straight segment can estimate two orthogonal horizontal vanishing points as discussed before, which can be combined with the estimated vertical vanishing point to form a triple of orthogonal vanishing points.

2) *Crossroad:* The case of crossroads is more complicated, as illustrated in Fig. 7(a). The activities of vehicles in the crossroad contain running ahead, turning left, turning right, and turning around. For the crossroad shown in Fig. 7(a), we can detect four straight segments [Segments 1–4 as shown in Fig. 7(a)] by velocity direction information.

For every straight segment, we can estimate a pair of orthogonal horizontal vanishing points with the strategy similar to the above case. Evidently, the pairs estimated from Segments 1 and 2 should coincide with each other, while the pairs estimated from Segments 3 and 4 should coincide with each other.

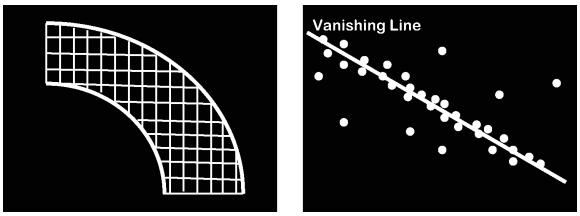


Fig. 8. (a) Division of roadway into pieces. (b) Estimation of horizontal vanishing line.

In fact, we can still use the strategy of the simplest case of one straight road for vanishing points estimation of crossroad. In this case, there will be two evident peaks in the voting spaces for horizontal vanishing points estimation. One peak corresponds to the direction of Segments 1 and 2, while the other peak corresponds to the direction of Segments 3 and 4. Evidently, traffic flow should always focus on one of the two directions to avoid accidents. As a result, the two peaks should be of different heights so that they can be distinguished from each other. This strategy can also recover two pairs of orthogonal horizontal vanishing points.

3) *Bend:* Bends should be the most complicated case, as shown in Fig. 7(b), which is an interesting topic in single-view camera calibration [32]. No straight segment can be detected from this roadway. Instead, based on the roadway detection described in Algorithm 1, we can divide the roadway into pieces, as illustrated in Fig. 8(a). If the centroid of the vehicle is in the piece, we consider that the whole vehicle is passing by the piece. We assume that most vehicles passing one piece should move in the same direction. As a result, we can estimate a pair of horizontal orthogonal vanishing points by a strategy of the simplest case based on motion and appearance information of those vehicles passing by every piece. Since we divide the whole roadway into many small pieces, each piece with enough vehicles passing by will supply us a pair of estimated horizontal vanishing points so that we can obtain a large set of pairs of estimated vanishing points. However, the assumption of vehicles moving in the same direction is not always accurate in all pieces. We should select the most accurate pair from the collected large set.

As we know, all horizontal vanishing points should lie on the horizontal vanishing line. For all the horizontal vanishing points estimated from those pieces, Hough transform can be applied to estimate this vanishing line, as illustrated in Fig. 8(b). The pair with the smallest sum of distances to the estimated vanishing line is selected for the following processing. In practice, we prefer to divide the roadway into equal rectangular regions for convenient computation, which are adopted in our experiments.

E. Summary

For one straight roadway, we can estimate a triple of orthogonal vanishing points from appearance and motion

information of moving objects in videos. Since other cases of roadway layouts can be seen as combinations of the above three primitive layouts, we can estimate at least one triple of orthogonal vanishing points in almost all kinds of roadway layouts in reality. The method of complete calibration from one or more than one triple of orthogonal vanishing points will be described in the next section.

IV. Camera Calibration

In this section, we introduce our approach to recover camera models from vanishing points.

For a pin-hole camera, perspective projection from the 3-D world to an image can be conveniently represented in homogeneous coordinates by the projection matrix $\mathbf{P}$

$$\lambda_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} = \mathbf{K} \quad [\mathbf{R} \quad \mathbf{T}] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}. \tag{4}$$

Here, $(X_i, Y_i, Z_i, 1)^T$ are homogeneous coordinates of the 3-D point, while $(u_i, v_i, 1)^T$ are homogeneous coordinates of the projected point on image plane. λ_i is the unknown scale parameter.

As we know, matrix $\mathbf{P}$ can be further decomposed into the 3×3 rotation matrix $\mathbf{R}$, the 3×1 translation vector $\mathbf{T}$, and the intrinsic parameter matrix $\mathbf{K}$ that has the form

$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

Here, α_u and α_v represent the focal length of cameras in terms of pixel dimensions in the u and v directions, respectively. s is referred to as the skew parameter and $(u_0, v_0)^T$ are the coordinates of the principal point [7].

For surveillance cameras, we can usually make the assumption of zero skew ($s = 0$) and unit aspect ratio ($\alpha_u = \alpha_v = f$) so that intrinsic matrix $\mathbf{K}$ is simplified to have only three degrees of freedom.

A. Recovery of K and R

The three vanishing points correspond to the three orthogonal directions in the 3-D space, which are chosen to set up the world coordinate system (WCS). Due to the fact that points at infinity correspond to the three orthogonal directions, we can derive the constraints as follows:

$$\begin{bmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \lambda_3 u_3 \\ \lambda_1 v1 & \lambda_2 v_2 & \lambda_3 v_3 \\ \lambda_1 & \lambda_2 & \lambda_3 \end{bmatrix} = \mathbf{P} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{KR} \tag{6}$$

where $(u_1, v_1)^T$, $(u_2, v_2)^T$, and $(u_3, v_3)^T$ are coordinates of the three orthogonal vanishing points in the image plane. λ_1, λ_2, and λ_3 are unknown scale parameters.

Since the rotation matrix $\mathbf{R}$ satisfies $\mathbf{R} \cdot \mathbf{R}^T = \mathbf{I}$, (6) can be rearranged to derive constraints on $\mathbf{K}$ as follows:

$$\begin{bmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ 0 & \lambda_2^2 & 0 \\ 0 & 0 & \lambda_3^2 \end{bmatrix} \begin{bmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{bmatrix}^T = \mathbf{K}\mathbf{K}^T = \mathbf{W}^{-1}. \tag{7}$$

Under the assumption of unit aspect ratio ($\alpha_u = \alpha_v = f$) and zero skew ($s = 0$), (7) can be solved to recover three intrinsic camera parameters (f, u_0, v_0) and the three unknown factors (λ_1, λ_2, λ_3).

With $\mathbf{K}$ and (λ_1, λ_2, λ_3) solved, they can be substituted into (6) to solve the rotation matrix $\mathbf{R}$.

For more than one triple of orthogonal vanishing points estimated from traffic scene surveillance videos, we can adopt the method described in [33] to combine constraints for accurate camera calibration. Based on (7), it can be derived that every two orthogonal vanishing points $(u_i, v_i)^T$ and $(u_j, v_j)^T$ satisfy

$$\begin{bmatrix} u_i & v_i & 1 \end{bmatrix} \mathbf{W} \begin{bmatrix} u_j \\ v_j \\ 1 \end{bmatrix} = 0 \tag{8}$$

where

$$\mathbf{W} = \begin{bmatrix} w_1 & w_2 & w_4 \\ w_2 & w_3 & w_5 \\ w_4 & w_5 & w_6 \end{bmatrix}. \tag{9}$$

It gives a linear equation of $\mathbf{W}$ as follows:

$$u_i u_j w_1 + (v_i u_j + u_i v_j) w_2 + v_i v_j w_3 + (u_i + u_j) w_4 + (v_i + v_j) w_5 + w_6 = \mathrm{g}^T \mathbf{W} = 0. \tag{10}$$

All the estimated tripes of vanishing points can generate linear equations and can be collected together to form $\mathbf{AW} = 0$. The $\mathbf{W}$ is solved based on a least-squares strategy to minimize $\|\mathbf{AW}\|$. We should further normalize the rows of $\mathbf{A}$ to give all constraints the same weight [33]. Then, $\mathbf{K}$ can be obtained from Cholesky decomposition.

B. Recovery of T

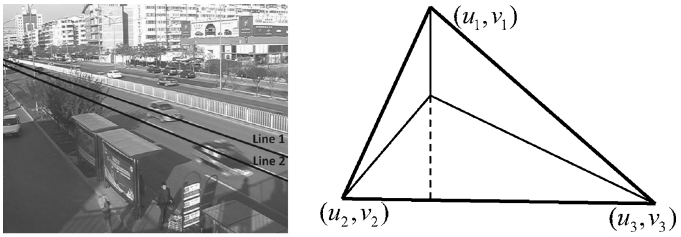
Translation matrix $\mathbf{T}$ describes the pose of cameras in the WCS. As mentioned in [32], an arbitrary reference point (u_4, v_4) corresponding to the origin of WCS on the ground plane can be chosen from image planes, but not lying on any of the three vanishing lines determined by the three orthogonal vanishing points to avoid degeneracies. It can be deduced that

$$\lambda_4 \begin{bmatrix} u_4 \\ v_4 \\ 1 \end{bmatrix} = \mathbf{K} \quad [\mathbf{R} \quad \mathbf{T}] \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \mathbf{KT}. \tag{11}$$

This supplies two constraints about $\mathbf{T}$, which is not sufficient to completely solve $\mathbf{T}$ because $\mathbf{T}$ has three degrees of freedom.

The other property is that the optical center of the camera lies on the $z = H$ plane so that

$$-\mathbf{R}^{-1}\mathbf{T} = \begin{bmatrix} x_c \\ y_c \\ H \end{bmatrix} \tag{12}$$



(a) (b)

Fig. 9. Illustration of estimating camera parameters. (a) Illustration of a frame of Video 1. (b) Triangle of vanishing points.

where (x_c, y_c) are the coordinates of the optical center in WCS. So, another linear equation about **T** can be derived from (12). The above-derived simultaneous equations are sufficient to recover the translation matrix **T**.

In this section, we have proposed our method of complete calibration of surveillance cameras with one or more than one triples of three estimated orthogonal vanishing points and the measured camera height H. In the next section, experiments are conducted to evaluate the performance of our calibration method.

V. Experimental Results and Analysis

Experiments are conducted in different scenes and experimental results are presented in this section to test the performance of the proposed approach.

A. Illustration of the Procedure in the Simplest Case

The procedure of camera calibration is first illustrated in the simplest case of only one straight roadway in the view field. The testing video (Video 1) is captured by a Panasonic NV-MX500 digital video, with one illustrated frame (720×576 pixels in size) shown in Fig. 9(a). With the accumulating of moving objects, the three orthogonal vanishing points are finally estimated as $(u_1, v_1) = (-217, 70)$, $(u_2, v_2) = (1806, 31)$, and $(u_3, v_3) = (427, 4906)$, as shown in Fig. 9(b). We further manually mark two lines from landmarks on the ground plane along the roadway orientation, as shown in Fig. 9(a), and calculate the intersection as $(-220, 68)$. The intersection is close to the (u_1, v_1), which further illustrates the effectiveness of our approach to estimate vanishing points from moving objects.

Using the methods described in Section IV, we can recover the intrinsic camera parameters $\alpha_u = \alpha_v = 884$, $(u_0, v_0) = (336, 226)$. With the camera height measured as 7420 mm and the center of the image taken as the corresponding point of the origin of WCS on the ground plane, the rotation matrix **R** and the translation matrix **T** can be recovered as follows:

$$\mathbf{R} = \begin{bmatrix} -0.5244 & 0.8512 & 0.0190 \\ -0.1484 & -0.1134 & 0.9824 \\ 0.8384 & 0.5124 & 0.1858 \end{bmatrix} \tag{13}$$

$$\mathbf{T} = \begin{bmatrix} 781\,\text{mm} \\ 2020\,\text{mm} \\ 29\,180\,\text{mm} \end{bmatrix}. \tag{14}$$

TABLE I
Recovered Intrinsic Parameters of the Digital Camera (Pixels)

Parameter	f	u_0	v_0
Video 1	884	336	226
SIFT [34]	880	332	231

Fig. 10. Correspondence of labeled points on an image plane.

To test the effectiveness of the intrinsic parameter estimation, we have further captured two images from different view angles with overlap in the laboratory. SIFT corresponding points [34] are obtained to substitute the template in Zhang's method [9] to estimate the intrinsic parameters of the camera. The intrinsic parameters estimated based on these two methods are listed in Table I. As we can see, the estimated parameters are close to each other, which illustrates the effectiveness of our approach to estimate intrinsic camera parameters.

To further test the performance of parameter estimation, we compare our results with the well-known direct linear transformation (DLT) method [7]. As we know, in traffic scene surveillance, the most conventional method for camera calibration is based on point correspondences between 3-D real scenes and 2-D images [6]. In order to improve the accuracy of projection matrix estimation, we need to survey the whole scene and label points distributed uniformly within the whole scene. To compare with our approach, we manually labeled more than 60 points and mark the corresponding points in the image plane, as shown in Fig. 10.

Twenty points of them are selected to recover the projection matrix **P** based on DLT as follows:

$$\mathbf{P} = \begin{bmatrix} -195.67 & 912.46 & 83.77 & 9\,429\,158 \\ 50.74 & 17.79 & 897.79 & 9\,270\,693 \\ 0.68 & 0.70 & 0.15 & 32\,739 \end{bmatrix}. \tag{15}$$

In comparison, the projection matrix recovered by our approach is as follows:

$$\mathbf{P} = \begin{bmatrix} -181.94 & 925.3 & 79.3 & 10\,504\,946 \\ 58.87 & 15.8835 & 911.18 & 8\,403\,957 \\ 0.84 & 0.51 & 0.19 & 29\,180 \end{bmatrix}. \tag{16}$$

TABLE II
EVALUATION OF PERFORMANCE OF ALL KINDS OF TRAFFIC SCENES (PIXELS)

Video	(u_1, v_1)	(u_2, v_2)	(u_3, v_3)	f	(u_0, v_0)	f_{sift}	$(u_0, v_0)_{\text{sift}}$	Error	$\text{Error}_{\text{DLT}}$
Video 2	(1815, 21)	(−1912, 636)	(2122, 7645)	1594	(946, 521)	1548	(924, 541)	7	11
Video 3	(2004, −5)	(−3491, −92)	(821, 8925)	2090	(954, 497)	2214	(978, 548)	8	14
Video 4	(2050, 251)	(−1243, 215)	(931, 9836)	1501	(1033, 481)	1544	(968, 542)	15	13
Video 5	(916, 107)	(−436 712, 57 116)	(2512, 12 475)	2520	(970, 636)	2219	(961, 539)	34	14
Video 6	(3470, −70)	(−397, 319)	(1883, 9294)	1818	(1004, 556)	1823	(946, 545)	11	12
Video 7	(−8971, −308)	(1750, 471)	(−3536, 62 145)	2823	(940, 541)	2743	(961, 534)	7	15
Video 8	(−844, −1369)	(2732, −128)	(459, 1934)	1344	(927, 583)	1493	(983, 532)	13	11
Video 9	(1739, 103)	(−3169, 429)	(1429, 8580)	1831	(897, 578)	1719	(941, 552)	19	13

On one hand, the given **K**, **R**, and **T** can fully determine a projection matrix **P**. On the other hand, a given projection matrix **P** can be decomposed into **K**, **R**, and **T**. It makes sense to compare calibration accuracy by comparing estimated projection matrices. Here, comparison of DLT and our method is conducted by using the other 40 corresponding pairs for testing. For a 3-D point in the testing set, we take the manually labeled 2-D points as the "ground truth" of its projection on image plane. With **P** estimated, we can calculate the projected point in the image plane. Then, the accuracy of one correspondence is measured by the Euclidean distance between the backprojected point and the ground truth on the image plane. The accuracy of the estimated projection matrix is measured by the sum of all distances between the calculated point and the ground truth. Beyond the manual collection of correspondences of DLT, we find that our approach performs better, giving a more than 3% higher accuracy in this accuracy measurement. The possible reason is that our approach makes use of redundant moving objects to achieve camera calibration rather than the 20 points in the DLT-based method.

B. Performance Evaluation in All Kinds of Traffic Scenes

In the above, we have taken one traffic scene as an example and evaluated the performance of our approach for camera calibration in the simplest case of only one straight roadway in the view field. However, most of traffic scenes in reality do not satisfy this property. We have captured the other eight videos (Videos 2–9) of typical traffic scenes by Panasonic HDC-HS700 of resolution (1920 × 1080), as illustrated in Fig. 11. Experiments are conducted to test the performance of our approach to deal with all kinds of traffic scene layouts.

Video 2 is another case of one straight roadway in the view field, which is processed similarly to Video 1. Video 3 contains more than one straight segment. Video 4 is a combination a straight segment and a bend. Video 5 contains a straight roadway and a bend. Based on motion detection and roadway extraction described in Algorithm 1, one strongest straight segment is extracted, respectively, from each scene and adopted to estimate a couple of horizontal vanishing points. Videos 6 and 7 are two typical crossroads. We select the straight segment of the largest number of vehicles to estimate two horizontal vanishing points. Video 8 has only one bend in the view field, which is divided as shown in the Fig. 11(g), and one couple with the smallest distance to the estimated vanishing line is selected for further processing. Video 9 is the complicated case of combinations of two bends and a straight segment. One bend of the heaviest traffic is selected, as shown in Fig. 11(h), to estimate one couple of horizontal vanishing points. In all these scenes, each couple of horizontal vanishing points are combined with the vertical vanishing point estimated from pedestrians to achieve camera calibration.

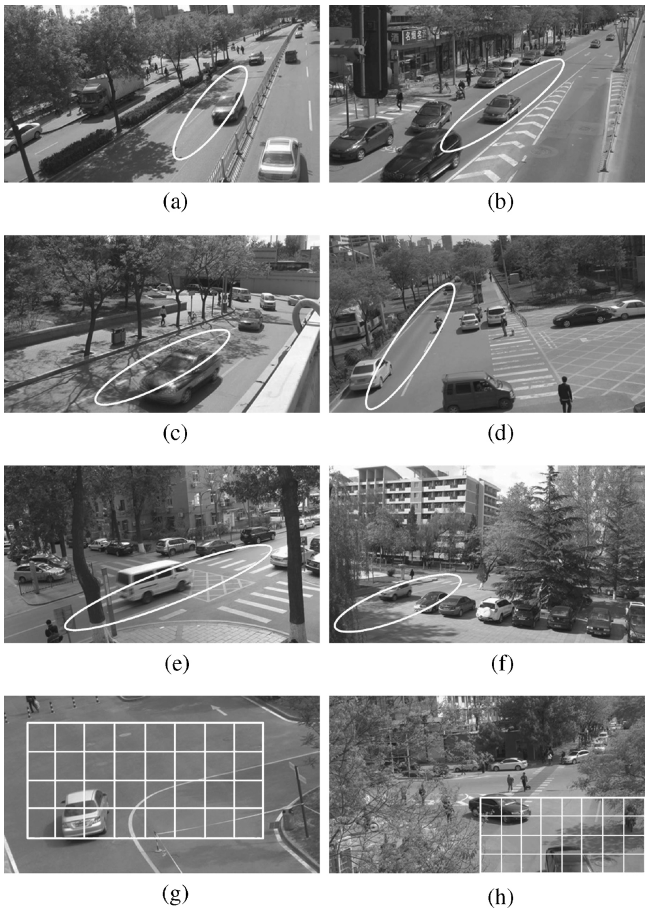


Fig. 11. Illustration of all kinds of typical traffic scenes. (a) Video 2. (b) Video 3. (c) Video 4. (d) Video 5. (e) Video 6. (f) Video 7. (g) Video 8. (h) Video 9.

The estimated three orthogonal vanishing points and the recovered intrinsic parameters of all the videos are listed in Table II. We also compare the estimated intrinsic parameters with the SIFT corresponding based method described in the first experiment. As we can see, the intrinsic parameters

(a) (b) (c)

Fig. 12. Illustration of frames from videos of different view angles. (a) Video 1 (first view). (b) Video 10 (second view). (c) Video 11 (third view).

TABLE III
RECOVERED INTRINSIC PARAMETERS OF DIFFERENT VIEW ANGLES (PIXELS)

Parameter	f	u_0	v_0
Video 1	884	336	226
Video 10	872	325	234
Video 11	893	342	238
SIFT [34]	880	332	231

estimated from our method are close to the SIFT-based method in all the videos, which demonstrate the effectiveness of our approach to achieve intrinsic camera parameter estimation in all kinds of traffic scenes. Furthermore, we manually measure the 3-D coordinates of 60 points in each scene. Twenty of them are chosen to estimate projective matrix based on DLT. We compare the backprojection errors in pixels of the remaining 40 points between our approach and the DLT-based method. The average backprojection error of our approach (13.7 pixels) is comparable to the DLT-based method (12.9 pixels), but supplies more practicability and convenience. We should notice that our approach performs much worse in Video 5 with large backprojection error in pixels. It is because one of the estimated horizontal vanishing points in Video 5 is much far from the visible image plane, which leads to instability and larger errors. We should admit that our approach may perform worse or even fail in some of these extreme degenerated cases, but performs well in most surveillance scenes.

C. Stability Evaluation to Different Factors

In the above, we have shown the effectiveness of our approach to deal with all kinds of traffic scenes. Further experiments are conducted to test the stability of our approach to different factors.

1) *Stability Testing to Different View Angles:* The first experiment is conducted to test the stability of our approach to different view angels. As shown in Fig. 12, the three videos are captured in different view angles without changing the intrinsic parameters of cameras. The respective recovered parameters from these three videos are listed in Table III. As we can see, the intrinsic parameters recovered, respectively, from three videos of different view angles vary in a less than 2% small range and are comparable to the method based on interest point correspondence by SIFT [34], which demonstrates the stability of our approach to different view angles.

2) *Stability Testing to Different Straight Segments Selection:* The second experiment is conducted to test the stability of our approach to different straight segments selection

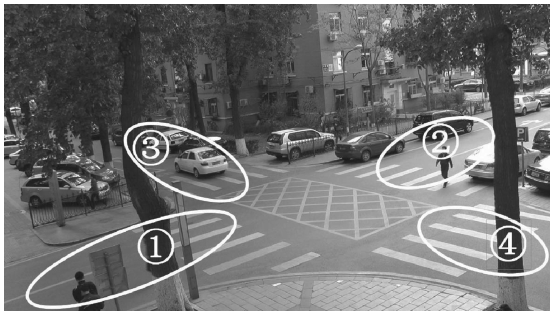


Fig. 13. Illustration of a crossroad in Video 6.

TABLE IV
RECOVERED INTRINSIC PARAMETERS OF DIFFERENT STRAIGHT SEGMENTS (PIXELS)

Parameter	f	u_0	v_0
Segment 1	1821	1013	559
Segment 2	1804	1009	557
Segment 3	1784	985	542
Segment 4	1792	997	533
Fusion	1817	1007	554
SIFT [34]	1823	946	545

TABLE V
COMPARISON OF BACKPROJECTION ERRORS IN IMAGES (PIXELS)

Segment	Segment 1	Segment 2	Segment 3	Segment 4	Fusion
Error	12	13	17	15	11

in the view field. As shown in Fig. 13, there is a crossroad in the view field. Based on motion detection and straight segment estimation, we can extract four straight segments, as illustrated in Fig. 13. For each straight segment, vehicles can be accumulated to estimate a couple of horizontal vanishing points, which are then combined with the vertical one to achieve camera calibration. We can further make use of all estimated four triples of vanishing points to obtain a more stable result based on the least-squares strategy described in Section IV. The recovered intrinsic parameters from each of the straight segment and all segments are listed in Table IV. It is shown that the estimated parameters are close to each other within a 3% variance, which demonstrates the stability of our approach for different selections of straight segments in the view field.

We further compare the back-projection errors in pixels of 40 known 3-D points in the scene of Video 6, with average backprojection errors shown in Table V. As we can see, all the backprojection errors are acceptable for each straight segment, and are comparable to the DLT-based method. The least-squares-based strategy has not improved the accuracy evidently, which is because the orientations are close to each other for all these four segments and the estimated triple of orthogonal vanishing points are closely distributed. It should be noted that the backprojection error based on Segment 1 or 2 is less than Segment 3 or 4. The reason may be that there is heavier traffic in Segment 1 or 2 to accumulate more redundant information.

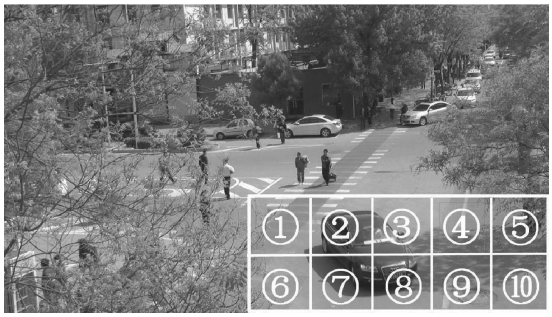


Fig. 14. Illustration of a bend in Video 9.

3) *Stability Testing to Different Subregion Selection:* The third experiment is conducted to test the stability of our approach in different selections of subregions in the case of the bend. Video 9 contains a typical bend, as shown in Fig. 14. We equally divide the regions of interest into 5×2 subregions. With more than 120 min accumulation of moving objects, most of subregions are available to estimate a couple of orthogonal vanishing points (Subregion 6 cannot accumulate enough vehicles to obtain stable vanishing points, Subregion 5 is occluded by branches so that the line equations cannot be accurately estimated from vehicles). The recovered intrinsic parameters by the respective estimated horizontal vanishing points and the global vertical vanishing points are listed in Table VI. As we can see, the estimated parameters vary in a less than 15% range, which demonstrates that our approach has some stability to different subregion selections. However, it is not as stable as straight segment selection in the case of crossroad. Especially, in Subregions 7 and 8, the variance of different vehicle orientations makes it difficult to estimate accurate peaks in the voting space. By comparing the backprojection errors in pixels listed in Table VII, we can find that the average backprojection errors in pixels are larger than the case of straight segments, especially in Subregions 7 and 8. Another interesting phenomenon is that the selected horizontal vanishing pair closest to the estimated vanishing line does not indicate the least backprojection error, which is another disadvantage of our approach to deal with bends. Overall, we should admit that our approach cannot deal with bends as well as straight segments, but can achieve a coarse camera calibration of acceptable accuracy. The performance of our approach can also be improved if more abundant moving objects are captured in the case of bends.

4) *Stability Testing to Different Camera Heights:* The fourth experiment is conducted to test the stability of our approach to different camera heights, which are the only one user input of our camera calibration system. Four videos of the same scene are captured by the same camera with the camera heights descending by 4200 mm one by one, as shown in Fig. 15. We fix the division of regions of interest, respectively, for each camera, which are further processed by selecting the couple of horizontal vanishing points closest to the vanishing line. The estimated intrinsic parameters are shown in Table VIII. Even though the stability of our approach to deal with bends is not as good as straight segments, the recovered intrinsic parameters vary in about 10% range and are close to the results of the SIFT-based method, which also illustrates the effectiveness of our approach in dealing with bends in reality.

TABLE VI
RECOVERED INTRINSIC PARAMETERS OF DIFFERENT STRAIGHT SEGMENTS (PIXELS)

Parameter	f	u_0	v_0
Subregion 1	1966	989	600
Subregion 2	1830	897	578
Subregion 3	1849	1039	626
Subregion 4	1757	977	621
Subregion 7	2213	1068	493
Subregion 8	2130	997	678
Subregion 9	1762	1017	565
Subregion 10	1779	1019	584
SIFT [34]	1719	941	552

TABLE VII
COMPARISON OF BACKPROJECTION ERRORS IN IMAGES FOR THE BEND (PIXELS)

Subregion	1	2	3	4	7	8	9	10	DLT
Error	24	19	17	14	42	44	14	13	13

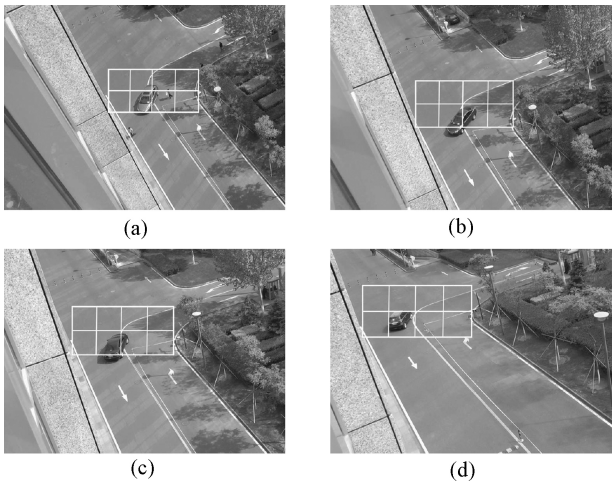


Fig. 15. Illustration of videos of different camera heights. (a) Video 12. (b) Video 13. (c) Video 14. (d) Video 15.

D. Accuracy According to an Accumulated Number of Moving Objects

As we know, our approach can estimate three orthogonal vanishing points on-the-fly by search extreme points in voting spaces. The more number of moving objects accumulated, the better estimation accuracy of vanishing points should be achieved. An experiment is conducted in the simplest case of Video 1 to test the effect of different numbers of accumulated moving objects. We measure the error by the difference between the extreme point and the final estimation. The trends of distance error according to the increasing number of moving objects for all three vanishing points are shown in Fig. 16. As we can see, the distance error becomes smaller and finally converges to the estimation value. Based on our experiments

TABLE VIII
RECOVERED INTRINSIC PARAMETERS OF DIFFERENT CAMERA HEIGHT (PIXELS)

Parameters	f	u_0	v_0
Video 12	2516	996	614
Video 13	2351	867	569
Video 14	2594	1016	621
Video 15	2456	998	589
SIFT [34]	2527	972	633

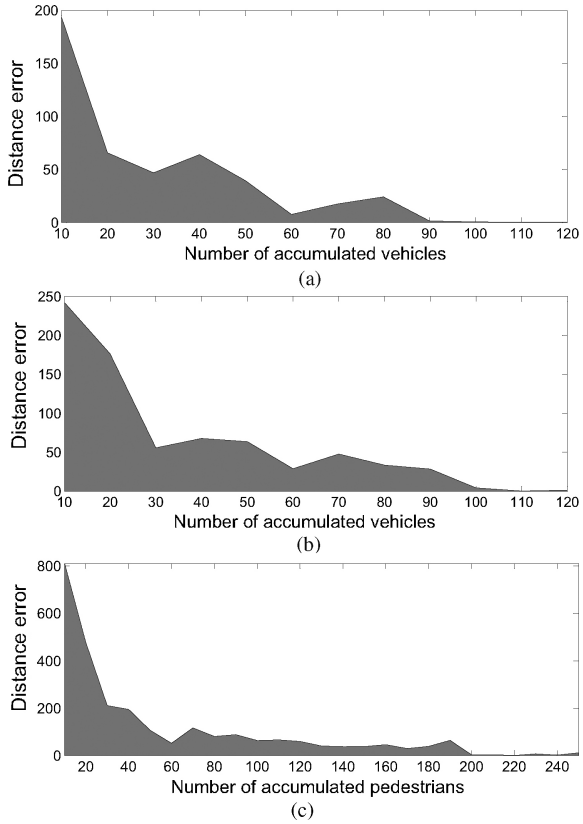


Fig. 16. Error trends with the increasing number of accumulated moving objects at the (a) first horizontal vanishing point, (b) second horizontal vanishing point, and (c) vertical vanishing point.

on all kinds of traffic scenes, accumulating more than 500 vehicles in a straight segment, more than 700 vehicles in a subregion of bends, and more than 1000 pedestrians in the whole scene can ensure the accuracy of vanishing points estimation. Since we can accumulate the same moving object every frame ignoring temporal relationship, this requirement is not difficult to be achieved in cases of normal traffic density.

E. Applications

Our approach makes use of motion and appearance information of moving objects in videos to achieve camera calibration, which has diverse applications in traffic scene surveillance. In the following, we list several applications of our practical camera calibration, and the successful applications can further illustrate the performance of our camera calibration method.

1) *Applications to 3-D Measurement in Images:* Camera calibration makes it possible to recover 3-D measurement

Fig. 17. Scene measurement from images.

TABLE IX
MEASUREMENT FROM IMAGES OF THE DIGITAL CAMERA

Label	l_1	l_2	l_3	l_4	l_5	l_6	l_7
Test	0.94	1.01	1.03	1.05	0.96	1.02	0.40
Real	1.00	1.00	1.00	1.00	1.00	1.00	0.46
Label	l_8	l_9	l_{10}	l_{11}	l_{12}	l_{13}	l_{14}
Test	0.40	0.48	0.46	0.52	0.48	0.50	1.94
Real	0.46	0.46	0.46	0.46	0.46	0.46	2.11
Label	l_{15}	l_{16}	l_{17}	l_{18}	l_{19}	l_{20}	l_{21}
Test	1.92	2.04	0.91	0.99	2.23	2.25	2.07
Real	2.11	2.11	1.13	1.13	2.11	2.11	2.11
Label	l_{22}	l_{23}	l_{24}	l_{25}	l_{26}	l_{27}	l_{28}
Test	1.10	1.12	0.49	1.39	0.51	0.48	1.42
Real	1.13	1.13	0.52	1.50	0.52	0.52	1.50

from images. At the same time, the accuracy of image-based 3-D measurement can also test the performance of camera calibration. As shown in Fig. 17, we take one length as unit length for reference, and 27 length ratios are measured from images. The measured value and the ground truth of every value are listed in Table IX.

As we can see, the average error of measurement is less than 10%, which demonstrates the accuracy of 3-D measurement in images. From the experimental results, two phenomena are worth mentioning here. The first one is that lines near the camera are measured more accurately than those far away. This is related to the measured pixel error on the image plane. The second one is that lines that are parallel to the ground plane or lie on the ground plane are measured more accurately than those which are perpendicular to the ground plane. This is because the horizontal vanishing points estimated from vehicles are more accurate than the vertical one from pedestrians. There are mainly three reasons. First, as artificial objects, vehicles are rigid and have more regular shapes than pedestrians. As a result, the extracted lines from vehicles should be more accurate than those of pedestrians. Second, the sizes of vehicles in images are often larger than the sizes of pedestrians. It also leads to less estimation accuracy from pedestrians. Third, the assumption that the trunk of a pedestrian is perpendicular to the ground plane is not as strict as the assumptions of vehicles. If we can make use of additional scene structures to improve the accuracy of

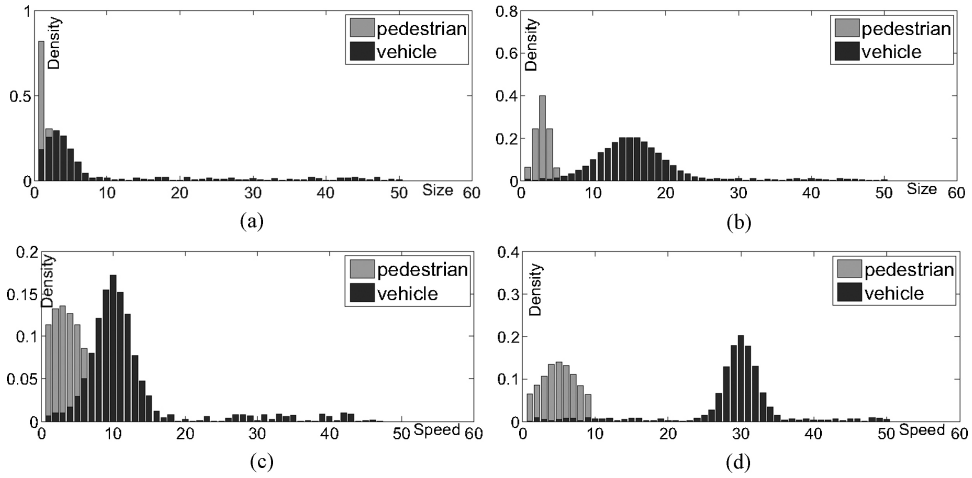


Fig. 18. Density of projected size (a) before and (b) after rectification. Density of projected speed (c) before and (d) after rectification.

TABLE X
CLASSIFICATION ACCURACY BEFORE AND AFTER RECTIFICATION

Classification accuracy	Before rectification	After rectification
Pedestrians	81.2%	97.3%
Vehicles	76.4%	95.2%

horizontal vanishing point estimation, the performance of 3-D measurement would be boosted.

2) *Applications to Automatic Object Classification:* In our framework, coarse object classification supplies abundant information for camera calibration. In fact, camera information can also feedback to help camera calibration. A pioneering work has been proposed by [25] to achieve view-independent object classification based on appearance normalization. Here, we also try to explore the camera model information to enhance the performance of object classification.

As we know, classification of objects in surveillance videos is difficult due to the perspective distortion of objects. The most common phenomenon is that close objects seem to be larger and move faster than those far away. As a result, 2-D image features such as size and velocity are not robust for object classification. With camera calibrated, we can rectify image features of objects in videos and make them robust for object classification. Two features, speed and size in image plane, are chosen to test the effect of camera calibration to object classification. With moving objects extracted from videos and labeled manually, we analyze the class-conditional densities of projected size and speed for pedestrians and vehicles before and after rectification, respectively, as shown in Fig. 18. It is evident that objects are much easier to be classified after feature rectification.

Here, we assume that these two features satisfy Gaussian distributions that are found, as shown in Fig. 18, and classification is realized based on Bayesian rules. The classification accuracy of moving objects before and after feature rectification is listed in Table X.

As we can see, classification accuracy has been greatly boosted by feature rectification. In our framework, we make use of motion and shape feature for coarse object classification, which can help to estimate three orthogonal vanishing points for camera calibration. Calibration results can then feedback to rectify 2-D image features for robust classification. Feature rectification for automatic object classification is also a good application of our camera calibration method.

3) *Applications to Model-Based Recognition:* As we know, 3-D object models supply global information to deal with pose variance and occlusions for object recognition, which has been researched over many years [4], [35], [36]. However, a calibrated camera is often required as a connection between the 3-D object and 2-D image data, which limits the applicability of 3-D model-based object recognition in real surveillance systems. Our approach makes use of motion and appearance information of objects in videos to achieve camera calibration with the camera height H as the only one user input, which greatly simplifies the calibration procedure and improves the applicability of model-based object recognition in traffic scene surveillance. We have applied our calibration method to model-based vehicle localization and recognition [3], which is illustrated in Fig. 19. As we can see, the 3-D model fits with image data very well in all kinds of pose and occlusions, which also demonstrates the accuracy of our approach for camera calibration.

F. Discussion

In the above, we have proposed a practical camera calibration approach based on motion and appearance information of moving objects in traffic scenes. Experimental results show that our approach can achieve acceptable camera calibration in all kinds of traffic scenes. However, some degenerate or near degenerate cases may also lead to inaccuracy or even failure of our approach for camera calibration. As we know, if the camera plane is parallel or perpendicular to the ground plane, we cannot recover the three orthogonal vanishing points from videos. In these cases, we need additional information such as more vertical or horizontal lines to achieve complete camera calibration. In surveillance applications, cameras always prefer to be mounted with a titled angle to the ground plane to cover a wider field of view. As a result, these extreme cases are not

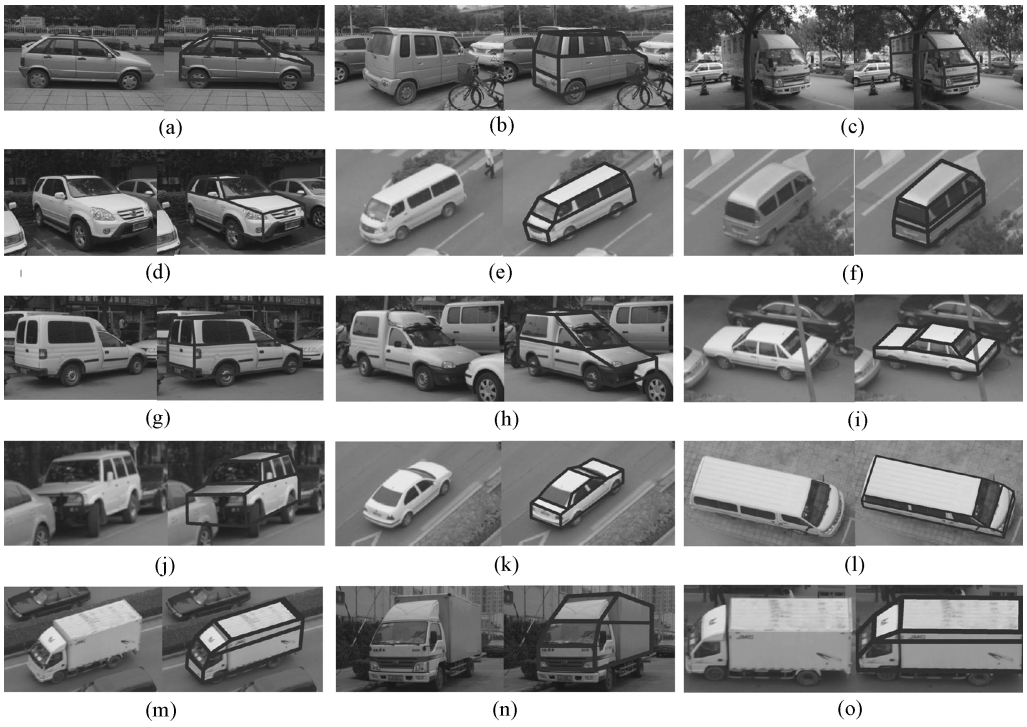


Fig. 19. Illustration of model based vehicle localization and recognition with camera calibrated with our approach [3]. (a) Hatchback (sideview). (b) Minivan (occluded by bicycles). (c) Truck (occluded by trees). (d) SUV (occluded by vehicles). (e) Van. (f) Minivan. (g) Pickup (occluded by vehicles). (h) Pickup (occluded by vehicles). (i) Sedan (occluded by pillars). (j) SUV (occluded by vehicles). (k) Sedan. (l) Van (top-down view). (m) Truck (view 1). (n) Truck (view 2). (o) Truck (view 3).

common in surveillance applications. In addition, some near degenerated cases where the estimated vanishing point is near infinity or much far away from the visible image plane would also cause inaccuracy of camera calibration (such as Video 5). Overall, our approach can achieve accurate calibration in almost all kinds of traffic surveillance scenes, but may fail in some extreme cases.

Another issue of our approach is that we need pedestrians in videos to estimate the third vanishing point. However, some traffic scenes like highways in reality have rare pedestrians, which makes it impossible to estimate the vertical vanishing point. In this case, we can make use of scene structures, such as buildings, trees, lamps, to accumulate the set of lines vertical to the ground plane. The learning-based method can be adopted to achieve tree or lamp detection automatically [37]. Furthermore, as we know, vehicles also have abundant lines along the vertical direction, how to extract information of vehicles to estimate the third vanishing point would also be our future work.

Our approach can achieve camera calibration almost automatically, but we still need to measure the camera height H as the only one user input. In all our experiments, camera heights were measured manually. In practice, camera heights may not be easy to obtain. As shown in Fig. 20(a), we can make two rays of different orientations (α_1 and α_2) from the position of the camera and obtain their projection on the ground plane. If we can ensure that the plane determined by these two rays is vertical to the ground plane, we can conveniently estimate the camera height H by measuring the distance d between the two projections on the ground plane as follows:

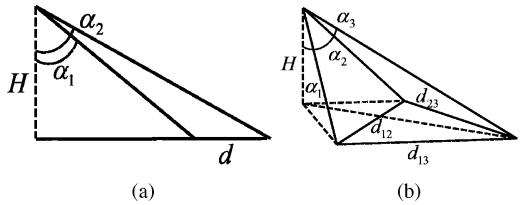


Fig. 20. Illustrations of camera height estimation. (a) Estimation from two rays. (b) Estimation from three rays.

$$H = \frac{d}{\tan(\alpha_1) - \tan(\alpha_2)}. \tag{17}$$

If we cannot ensure the plane is vertical to the ground plane, two rays are not enough and we should make three rays of different orientations (α_1, α_2, and α_3), as shown in Fig. 20(b). The camera height can be estimated by measuring the distances (d_{11}, d_{12}, and d_{13}) among the three projections on the ground plane, but the formula is much more complicated.

In addition, due to the unknown intrinsic structure of cameras, the height H of the camera optical center cannot be measured accurately. In our work, we use the distance between the camera and the ground plane to approximate this value. As we know, cameras are usually mounted very high in surveillance applications so that the measurement error of H is less than 2%. Even more, the error of H only affects the translation matrix $\mathbf{T}$. It can be shown that the estimation error of $\mathbf{T}$ is less than 2% with a 2% measurement error of H.

VI. Conclusion

In this paper, we proposed a novel method for camera calibration from traffic scene surveillance videos. With camera height H measured as the only user input, we can completely recover both intrinsic and extrinsic parameters of the camera based on appearance and motion information of moving objects in videos. Abundant experiments were conducted, and it was shown that our approach can achieve accurate camera calibration in most traffic scenes in reality. We also showed that our approach can be widely used in diverse surveillance applications. We can conclude that our approach improved the practicability and applicability of camera calibration in traffic scene surveillance.

References

[1] Z. X. Zhang, Y. Cai, K. Huang, and T. Tan, "Real-time moving object classification with automatic scene division," in *Proc. Int. Conf. Image Process.*, Sep. 2007, pp. 149–152.

[2] Q. Cai and J. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1241–1247, Dec. 1999.

[3] Z. Zhang, T. Tan, K. Huang, and Y. Wang, "Three-dimensional deformable-model-based localization and recognition of road vehicles," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 1–13, Jan. 2012.

[4] T. N. Tan, G. D. Sullivan, and K. D. Baker, "Model-based localization and recognition of road vehicles," *Int. J. Comput. Vision*, vol. 27, no. 1, pp. 5–25, 1998.

[5] A. Toshev, A. Makadia, and K. Daniilidis, "Shape-based object recognition in videos using 3-D synthetic object models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Dec. 2009, pp. 288–295.

[6] O. Faugeras, *Three Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA: MIT Press, 1993.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, MA: Cambridge Univ. Press, 2004.

[8] R. Tsai, "An efficient and accurate camera calibration technique for 3-D machine vision," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 1986, pp. 364–374.

[9] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[10] Z. Zhang, "Camera calibration with 1-D object," *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 892–899, Jul. 2004.

[11] H. Li, H. Zhang, F. Wu, and Z. Hu, "A new linear camera self-calibration technique," in *Proc. 5th Asian Conf. Comput. Vision*, Jan. 2002, pp. 23–25.

[12] B. Triggs, "Auto-calibration from planar scenes," in *Proc. Eur. Conf. Comput. Vision*, Jun. 1998, pp. 89–105.

[13] B. Caprile and V. Grimson, "Using vanishing points for camera calibration," *Int. J. Comput. Vision*, vol. 4, no. 2, pp. 127–140, Mar. 1990.

[14] R. Cipolla, T. Drummond, and D. Robertson, "Camera calibration from vanishing points in images of architectural scenes," in *Proc. Brit. Mach. Vision Conf.*, Sep. 1999, pp. 382–391.

[15] D. Liebowitz, A. Criminisi, and A. Zisserman, "Creating architectural models from images," in *Proc. EuroGraphics*, Jun. 1999, pp. 39–50.

[16] J. Deutscher, M. Isard, and J. MacCormick, "Automatic camera calibration from a single Manhattan image," in *Proc. Eur. Conf. Comput. Vision*, May 2002, pp. 175–205.

[17] N. Krahnstoever and P. R. S. Mendonca, "Bayesian autocalibration for surveillance," in *Proc. Int. Conf. Comput. Vision*, Oct. 2005, pp. 1858–1865.

[18] I. Junejo and H. Foroosha, "Robust auto-calibration from pedestrians," in *Proc. IEEE Int. Conf. Video Signal Based Surveillance*, Nov. 2006, p. 92.

[19] F. Lv, T. Zhao, and R. Nevatia, "Camera calibration from video of a walking human," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1513–1518, Sep. 2006.

[20] B. Micusik and T. Pajdla, "Simultaneous surveillance camera calibration and foot-head homology estimation from human detections," in *Proc. Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 1–8.

[21] M. Evans and J. Ferryman, "Surveillance camera calibration from observations of a pedestrian," in *Proc. IEEE Int. Conf. Advanced Video Signal Based Surveillance*, Aug.–Sep. 2010, pp. 64–71.

[22] I. N. Junejo and H. Foroosh, "Trajectory rectification and path modeling for video surveillance," in *Proc. Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–7.

[23] I. N. Junejo and H. Foroosh, "Euclidean path modeling for video surveillance," *Image Vision Comput.*, vol. 26, no. 4, pp. 512–528, 2008.

[24] B. Bose and E. Grimson, "Ground plane rectification by tracking moving objects," in *Proc. Joint Int. Workshops Visual Surveillance Performance Evaluation Tracking Surveillance*, Oct. 2003, pp. 1–8.

[25] C. Stauffer, K. Tieu, and L. Lee, "Robust automated planar normalization of tracking data," in *Proc. Joint Int. Workshops Visual Surveillance Performance Evaluation Tracking Surveillance*, 2003.

[26] Z. X. Zhang, M. Li, K. Huang, and T. Tan, "Practical camera auto-calibration based on object appearance and motion for traffic scene visual surveillance," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2008, pp. 1–8.

[27] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Jun. 1999, pp. 246–252.

[28] Z. Liu, K. Huang, and T. Tan, "Cast shadow removal with GMM for surface reflectance component," in *Proc. 18th IEEE Int. Conf. Pattern Recognit.*, Aug. 2006, pp. 727–730.

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2005, pp. 886–893.

[30] T.-J. Cham, A. Ciptadi, W.-C. Tan, M.-T. Pham, and L.-T. Chia, "Estimating camera pose from a single urban ground-view omnidirectional image and a 2-D building outline map," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2010, pp. 366–373.

[31] S. Battiato, G. M. Farinella, E. Messina, and G. Puglisi, "Robust image alignment for tampering detection," *IEEE Trans. Inform. Forensics Security*, 2012.

[32] A. Criminisi, "Accurate visual metrology from single and multiple uncalibrated images," Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 1999.

[33] D. Liebowitz and A. Zisserman, "Combining scene and auto-calibration constraints," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep. 1999, pp. 293–300.

[34] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[35] J. Ferryman, A. Worrall, G. Sullivan, and K. Backer, "A generic deformable model for vehicle recognition," in *Proc. Brit. Conf. Mach. Vision*, Jun. 1995, pp. 127–136.

[36] Y. Li, L. Gu, and T. Kanade, "A robust shape model for multi-view car alignment," in *Proc. Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 2466–2473.

[37] *The Pascal Visual Object Classes Homepage*. (2011) [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC

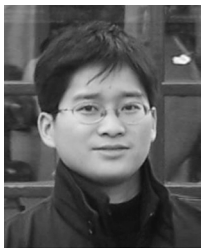
Zhaoxiang Zhang (M'08) received the B.S. degree in electronic science and technology from the University of Science and Technology of China, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligent systems from the Chinese Academy of Sciences, Beijing, China, in 2009.

In October 2009, he joined the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing, as a Faculty Member. His current research interests include computer vision, pattern recognition, image processing, and machine learning.

Tieniu Tan (M'92–SM'97–F'04) received the B.Sc. degree in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, London, U.K., in 1986 and 1989, respectively.

In October 1989, he joined the Computational Vision Group, Department of Computer Science, University of Reading, Reading, U.K., where he was a Research Fellow, a Senior Research Fellow, and a Lecturer. In January 1998, he joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China. From 2000 to 2007, he was the Director General of the Institute of Automation, CAS, and since 1998, has been a Professor and the Director of NLPR. He is also the Deputy Secretary-General (for cyberinfrastructure and international affairs) with CAS. He is the author of more than 350 research papers in refereed journals and conferences in the areas of image processing, computer vision, and pattern recognition. He is also the author or editor of nine books. He holds more than 50 patents. His current research interests include biometrics, image and video understanding, and information forensics and security.

Dr. Tan is a fellow of the International Association of Pattern Recognition (IAPR). He has served as the Chair or a Program Committee Member for many major national and international conferences. He has been the Founding Chair of the IAPR Technical Committee on Biometrics, and the Founding Chair of the IAPR/IEEE International Conference on Biometrics and the IEEE International Workshop on Visual Surveillance. He has been the Deputy President of the China Computer Federation and the Chinese Automation Association. He is currently the Vice President of the IAPR, the Executive Vice President of the Chinese Society of Image and Graphics, and the Deputy President of the Chinese Association for Artificial Intelligence. He has served as an Associate Editor or a member of the editorial boards of many leading international journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, *Pattern Recognition Letters*, and *Image and Vision Computing*. He is the Editor-in-Chief of the *International Journal of Automation and Computing*. He has given invited talks and keynotes at many universities and international conferences, and has received numerous national and international awards and recognitions.



Kaiqi Huang (M'07–SM'09) received the B.Sc. and M.Sc. degrees from the Nanjing University of Science Technology, Nanjing, China, and the Ph.D. degree from Southeast University, Nanjing.

Since 2004, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and in 2005, became an Associate Professor. He is the author of over 70 papers in international journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS, *Pattern Recognition*, ECCV, CVPR, ICIP, and ICPR. He is in charge of several fundings, including National Science Funding, and collaboration with some enterprises. He is also involved in several national research projects (e.g., 863, 973, and NSFC). His current research interests include visual-surveillance digital image processing, pattern recognition, and biological-based vision.

Yunhong Wang (M'98) received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively.

Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China, where she is also the Director of the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. From 1998 to 2004, she was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. Her current research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.