Practical Camera Auto-Calibration Based on Object Appearance and Motion for Traffic Scene Visual Surveillance

Zhaoxiang Zhang, Min Li, Kaiqi Huang and Tieniu Tan
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{zxzhang, mli, kqhuang, tnt}@nlpr.ia.ac.cn

Abstract

Camera calibration, as a fundamental issue in computer vision, is indispensable in many visual surveillance applications. Firstly, calibrated camera can help to deal with perspective distortion of object appearance on image plane. Secondly, calibrated camera makes it possible to recover metrics from images which are robust to scene or view angle changes. In addition, with calibrated cameras, we can make use of prior information of 3D models to estimate 3D pose of objects and make object detection or tracking more robust to noise and occlusions.

In this paper, we propose an automatic method to recover camera models from traffic scene surveillance videos. With only the camera height H measured, we can completely recover both intrinsic and extrinsic parameters of cameras based on appearance and motion of objects in videos. Experiments are conducted in different scenes and experimental results demonstrate the effectiveness and practicability of our approach, which can be adopted in many traffic scene surveillance applications.

1. Introduction

Camera calibration, as a fundamental topic in computer vision, is not only essential for many computer vision problems like stereo, metrology and reconstruction, but also benefits many application tasks like intelligent visual surveillance. Firstly, camera calibration can help to deal with perspective distortion of object appearance on 2D image plane which is a very difficult problem to solve for most 2D image feature based methods. Secondly, calibrated cameras make it possible to recover discriminant metrics robust to scene or view angle changes, which is greatly helpful for some applications like classification or tracking among multi-cameras. Thirdly, with cameras calibrated, we can make use of prior information of 3D models to estimate real 3D pose of objects in videos and make object detection or tracking more robust to noise and occlusions.

Due to its importance, much work has been done in the field of camera calibration with all kinds of approaches proposed. The common practice for camera calibration is to collect a set of correspondences between 3D points and their projections on image plane [4, 5]. However, a time-consuming wide site survey is required and it is difficult to measure 3D points which are not laid on the ground plane in wide surveillance scenes. Alternative strategies are proposed by Tsai [9] with a 3D known metric structure and Zhang [10] with a known planar template of unknown motion. However, requirement of calibrated templates limits the practicability of surveillance algorithms to different scenes. In addition, calibrated templates are not available in wide-field surveillance scenes because their projections are of very small size on image plane to supply poor accuracy for calibration.

Auto-calibration methods seem to be a more suitable way to recover camera parameters for surveillance applications. Since most surveillance applications make use of only one static camera, auto-calibartion cannot be achieved from camera motion but from inherent structure of monocular scenes. Caprile and Torre [2] described methods to use vanishing points to recover intrinsic parameters from a single camera but extrinsic parameters from multi-cameras. Liebowitz and *etc*. [6] developed a method to estimate intrinsic parameters by Cholesky decomposition and applied it to a scene reconstruction problem. Deutscher and *etc*. [3] made use of vanishing points in a Manhattan world to recover camera parameters for visual tracking. These methods are based on extraction of vanishing points from static scene structures such as buildings and landmarks.

In the absence of inherent scene structures, methods described above are not available. Researchers make use of object motion in videos to take the place of scene structure. Lv and *etc*. [8] obtained 3 orthogonal vanishing points by extracting head and feet positions of humans in videos on the assumptions of constant human height and planar human motion. As we know, precise pedestrian detection is

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

very difficult in surveillance videos due to noise and shadows. Further more, the approach requires accurate localization of head and feet positions of humans, which is more challenging in low resolution surveillance videos. Bose and *etc*. [1] tracked vehicles and detected constant velocity linear paths to realize ground plane rectification instead of recovering intrinsic and extrinsic camera parameters.

In this paper, we propose a novel automatic camera calibration method from traffic scene surveillance videos. With moving objects extracted from videos using motion information, three vanishing points corresponding to three orthogonal directions in real 3D world are estimated based on motion and appearance of moving objects. With only the camera height H measured, we can recover both intrinsic and extrinsic camera parameters, which is of great help for all kinds of surveillance applications. Experiments are conducted to evaluate the performance of this calibration algorithm in different traffic scenes. Experimental results demonstrate the accuracy and practicability of our approach.

The remainder of the paper is organized as follows. In Section 2, we introduce our method to extract accurate foreground areas with shadows removed. The strategy to estimate 3 orthogonal vanishing points from motion and appearance of video objects is described in Section 3. In Section 4, we introduce our method to realize calibration only from 3 orthogonal vanishing points and camera height H. Experimental results and analysis are given in Section 5. Finally, we draw our conclusions in Section 6.

2. Motion Detection

Motion and appearance of moving objects in surveillance videos supply plentiful information for calibration. In this section, we introduce our method for extraction of accurate foreground areas with shadows removed. As we know, Gaussian Mixture Model (GMM) is a popular method in the field of motion detection due to its outstanding ability to deal with slow lighting changes, periodical motions in clutter background, slow moving objects and long term scene changes. However, this method still has disadvantages that 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 [11] to deal with disadvantages mentioned above and the method can be summarized as follows:
(1) The intensity of each pixel is modeled as the product of irradiance component and reflectance component.
(2) The reflectance value of each pixel is modeled as a mixture of Gaussian.
(3) Every new pixel is matched against each of the existing Gaussian distributions. A match is defined as a pixel value within 2.5 standard deviations of a distribution.
(4) Sort the Gaussians and determine whether it is background.
(5) Adjust the Gaussians and their prior weights.
(6) If there is no match, replace the least probable Gaussian and set mask pixel to background.

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

(a) One frame of videos (b) Background recovered (c) Detected moving objects

Figure 1. Motion detection results with shadows removed (cited from [11])

3. Vanishing Points Estimation

A vanishing point is defined as the intersection of a series of projected parallel lines, which is very useful for autocalibration. In this section, we propose our method to estimate three orthogonal vanishing points from appearance and motion of moving objects in videos.

In fact, conventional traffic surveillance scenes have a series of helpful general properties for vanishing points estimation which are summarized as follows:

- Almost all moving objects including vehicles and pedestrians are moving on the ground plane.
- Vehicles always run along the roadway which can be seen to be straight or contain one or more approximately straight segments in the field of camera view.
- Image projection of vehicles are rich in line segments along two orientations which correspond to the symmetrical axis direction and its perpendicular direction in most view angles.
- In most cases, pedestrians are walking with their trunks perpendicular to the ground plane.

These four properties are found in most traffic surveillance scenes and they can be used to estimate three orthogonal vanishing points, which are described in detail as follows.

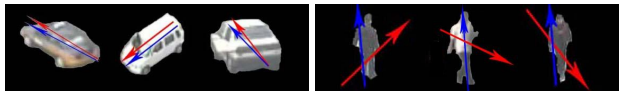
3.1. Coarse moving object classification

We extract two kinds of directions for every moving objects detected from videos. The first one is the velocity direction in images, which can be calculated due to its position change of unit time. The second one is the main axis

direction θ, which can be estimated from moment analysis of silhouette as:

$$\theta = arctan(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}) \quad (1)$$

Here, μ_{pq} is the central moment of order (p, q). The difference between these two directions supplies coarse category information. As we know, the direction difference is quite significant for pedestrians moving in videos while the two directions are very close for vehicles in most cases of camera view as shown in Figure 2. As a result, we take the difference of these two directions as discriminant feature for coarse classification. K-Mean clustering seems to be a good method for classification. However, due to large view angle variance in the camera view field, we should adopt more reliable strategy to avoid serious misclassification. In practice, we set two threshold values $\theta_1 = 5°$ and $\theta_2 = 20°$. The object is labeled as a vehicle if its direction difference is less than θ_1 and as a pedestrian if its direction difference is larger than θ_2. Those objects whose direction difference is between θ_1 and θ_2 are discarded. The latter estimation of vanishing points benefits from this strict classification strategy.



(a) Illustration of vehicles (b) Illustration of pedestrians

Figure 2. Illustrations in different view angles (Red arrowhead stands for velocity direction; blue arrowhead stands for main axis direction)

It is evident that this classification is not very accurate but enough for us to extract three orthogonal vanishing points as described in the following.

3.2. Line Equations Estimation

The four general properties in traffic surveillance scenes we summarized before supply important information for recovery of camera models.

Here, we assume that the roadway is straight in the field of view. Special cases of non-straight roadways will be discussed in Section 5. In this case, most vehicles are running in the same or inverse direction of the 3D world so that the symmetrical axes of most vehicles should be parallel to each other, which are also parallel to the ground plane. This supplies important information for us to extract horizonal vanishing points. As we have described before, image projection of vehicles are rich in line segments along two orientations which correspond to the symmetrical axis direction and its perpendicular direction. With these two orientations extracted for each vehicle detected from videos, we can estimate their intersections corresponding to the two horizonal vanishing points, respectively.

Due to perspective distortion, projected orientation of 3D direction is not unique and related to its position in images. For accuracy, we try to extract two accurate line equations corresponding to the two perpendicular directions for every vehicle detected from videos. Instead of sensitive edge point detection and combination to edge lines, these two orientations are extracted by Histogram of Orientated Gradient (HOG) in two stages. For every moving region labeled as vehicle 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. Those two bins with the largest values are chosen as coarse line orientations and a N bin HOG is calculated in each bin again to extract accurate line orientation, respectively. For every orientation estimated accurately, the line with this orientation slides from top to 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 Figure 3.

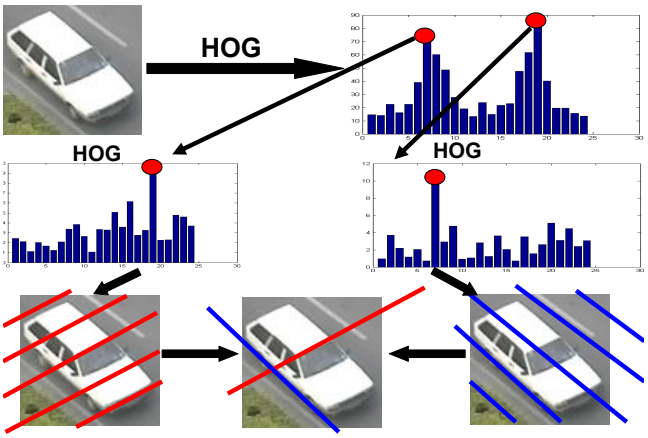


Figure 3. Flowchart for estimation of line equations for vehicles

For every vehicle, we can extract two line equations corresponding to two 3D directions. Motion direction is applied to distinguish these two directions. The line with its orientation close to motion direction corresponds to the symmetric axis direction of the vehicle while the other one corresponds to the perpendicular direction.

Pedestrians do not have so significant gradient orientations. However, as we know, most pedestrians are walking with their trunk perpendicular to the ground plane in most situations. Instead of localizing head and feet position in a small region, we take the line with main axis orientation passing by its centroid to describe trunk pose, which is more robust to be used for estimation of vertical vanishing points.

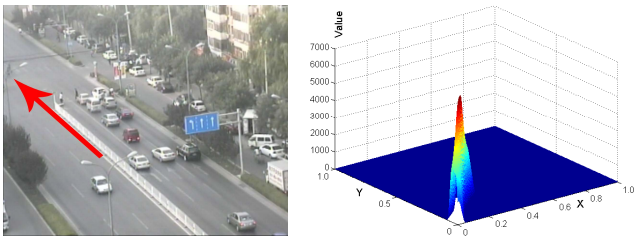
3.3. Intersection Estimation

There are three kinds of lines estimated from detected objects in videos. The first kind corresponds to the symmetric axis direction in reality. The second kind corresponds to perpendicular direction of symmetric axis. The third kind corresponds to the perpendicular direction of the ground plane.

With abundant objects detected from videos, we can collect large sets of lines for each kind and make use of them to estimate vanishing points. Due to large portion of outliers and noise in videos, lines are in fact not intersected at the same point. Various approaches can be adopted for robust estimation of the intersection point from redundant line equations. The simplest way is to solve simultaneous line equations based on least square strategy. Also, the problem can be transformed to estimate a point the sum of whose distance to all lines is minimal. This is an optimization problem which can be solved by Levenberg-Marquardt method. In addition, RANSAC is another strategy to solve this problem which has been used in [8].

In spite of the accuracy and robustness supplied by the above methods, they are not suitable to our case. In traffic scene surveillance, with videos processed and the frame number increases, more and more moving objects are detected from videos and the set of extracted lines become larger and larger. The intersections should be estimated from large sets of lines every moment without repeated calculation.

An improved voting strategy is adopted here for incremental estimation of intersections. It is based on the thought that every point on the line is the possible candidate as the intersection. The possibility satisfies 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 lines. Compared to other estimation method, this strategy can estimate the positions of vanishing points every moment without repeated calculation. Compared to traditional voting method, this strategy supplies more spiculate 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 Figure 4. Line equations from vehicles are taken to estimate 2 horizonal vanishing points while those from pedestrians are taken to estimate 1 vertical vanishing points. In this way, we can extract 3 orthogonal vanishing points (u_1, v_1), (u_2, v_2), (u_3, v_3) from appearance and motion information of moving objects in traffic scene surveillance videos.



(a) Illustration of a traffic scene (b) Voting surface

Figure 4. Illustration of estimating vanishing points from traffic scenes

4. 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 3D world to an image can be conveniently represented in homogeneous coordinates by the projection matrix **P**:

$$\lambda_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{T} \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{2}$$

As we know, the **P** can be further decomposed into the 3×3 rotation matrix **R**, the 3×1 translation vector **T** and the intrinsic parameter matrix **K** which has the form as

$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

With the assumption of zero skew ($s = 0$) and unit aspect ratio ($\alpha_u = \alpha_v = f$) for surveillance cameras, the **K** is simplified to have only 3 degrees of freedom.

4.1. Recovery of K and R

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

$$\begin{bmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \lambda_3 u_3 \\ \lambda_1 v1 & \lambda_2 v_2 & \lambda_3 v_3 \\ \lambda_1 & \lambda_2 & \lambda_3 \end{bmatrix} = \mathbf{P} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{KR} \tag{4}$$

Since the rotation matrix **R** satisfies $\mathbf{R} \cdot \mathbf{R}^T = \mathbf{I}$, (4) can be rearranged to derive constraints on **K** as:

$$\begin{bmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ 0 & \lambda_2^2 & 0 \\ 0 & 0 & \lambda_3^2 \end{bmatrix} \begin{bmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{bmatrix}^T = \mathbf{K}\mathbf{K}^T \tag{5}$$

Under the assumption of unit aspect ratio and zero skew, (5) can be solved to recover 3 intrinsic camera parameters and the 3 unknown factors, λ_i^2. A more robust strategy is to assume the main point (u_0, v_0) lying on the middle of image plane so that we only need to solve f from (5).

With $\mathbf{K}$ and λ_i solved, they can be substituted into (4) to solve the rotation matrix $\mathbf{R}$.

4.2. Recovery of T

Traditionally, the translation matrix $\mathbf{T}$ is recovered from correspondence between two or more views. However, only one static camera is usually used in surveillance applications. In this case, we can choose one arbitrary reference point (u_4, v_4) from image plane to correspond to the origin of the world coordinate system so that:

$$\lambda_4 \begin{bmatrix} u_4 \\ v_4 \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{T} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \mathbf{KT} \tag{6}$$

This supplies two constraints about $\mathbf{T}$, which leaves the scale factor λ_4 and is not sufficient to completely solve $\mathbf{T}$.

As we know, surveillance cameras are always mounted quite high from the ground plane so that the Z coordinate of the optical center can be simply estimated as the distance H between the camera and the ground plane. We will derive two other constraints from this metric.

The first property we can use is that the image projected point (u, v) of every point in $z = H$ plane are on the line across the two horizonal vanishing points (u_1, v_1) and (u_2, v_2). This lead to a linear equation about $\mathbf{T}$ as:

$$(u - u_1)(v_1 - v_2) - (v - v_1)(u_1 - u_2) = 0 \tag{7}$$

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

$$-\mathbf{R}^{-1}\mathbf{T} = \begin{bmatrix} x_c \\ y_c \\ H \end{bmatrix} \tag{8}$$

where (x_c, y_c) is the coordinate of optical center on the world coordinate system. So another linear equation about $\mathbf{T}$ can be derived from (8). The above derived simultaneous equations are sufficient to recover the translation matrix $\mathbf{T}$.

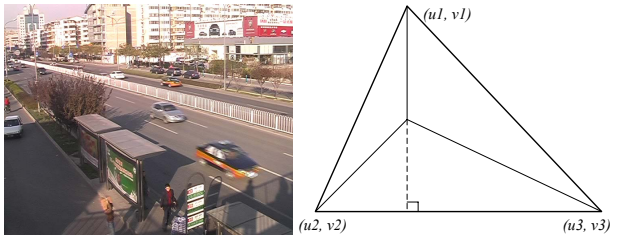
In this section, we propose our method of complete calibration of surveillance scenes with three estimated orthogonal vanishing points and the measured camera height H. In the next section, experiments are conducted to evaluate the performance of our auto-calibration method.

5. Experimental Results and Analysis

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

5.1. Illustration of the Procedure

One frame of a 720×576 traffic scene video captured by a Panasonic NV-MX500 digital video is shown in Figure 5(a). The three orthogonal vanishing points are estimated as $(u1, v1) = (-217, 70)$, $(u2, v2) = (1806, 31)$ and $(u3, v3) = (427, 4906)$ as shown in Figure 5(b). Using



(a) Illustration of a traffic scene (b) Triangle of vanishing points

Figure 5. Illustration of estimating camera parameters

the methods described in Section 4, we can recover the intrinsic camera parameters: $\alpha_u = \alpha_v = 884$, $(u_0, v_0) = (336, 226)$. With the camera height measured as 7420mm and the center of the image taken as the reference point, the rotation matrix $\mathbf{R}$ and the translation matrix $\mathbf{T}$ can be recovered as:

$$\mathbf{R} = \begin{bmatrix} -0.5244 & 0.8512 & 0.0190 \\ -0.1484 & -0.1134 & 0.9824 \\ 0.8384 & 0.5124 & 0.1858 \end{bmatrix} \tag{9}$$

$$\mathbf{T} = \begin{bmatrix} 781\text{mm} \\ 2020\text{mm} \\ 29180\text{mm} \end{bmatrix} \tag{10}$$

To test the effectiveness of our approach, we capture two other videos in different view angles without changing the intrinsic parameters of the camera. The frames of the two videos are illustrated in Figure 6.

(a) Frame of video1 (b) Frame of video2

Figure 6. Illustration of two videos from different view angles

In addition, we take the digital camera to capture two images from different view angles with overlap. The intrinsic parameters are recovered from interest point correspondence by SIFT [7]. The recovered camera intrinsic parameters including the original one are shown in Table 2.

As we can see, the intrinsic parameters recovered respectively from three videos vary in a small range less than 2%.

Table 1. Recovered intrinsic parameters of the digital camera

parameter	f	u_0	v_0
original video	884	336	226
video1	872	325	234
video2	893	342	238
SIFT	880	332	231

Further more, it is comparable to the method based on interest point correspondence. It is shown that our calibration method is accurate to be adaptive to different view angles. Many vanishing points based auto-calibration methods are not applicable because they cannot estimate the position of vanishing points accurately. In our approach, we make use of motion and appearance information of moving objects, which is very redundant for recovery of vanishing points. In addition, voting strategy is applied to get rid of outliers. In this case, our approach can estimate vanishing points quite accurately so that we can recover accurate camera parameters.

5.2. Comparison to Point Set Correspondence Based Method

For traffic scene surveillance, the most conventional method for camera calibration is based on point correspondence between 3D real scenes and 2D images [4]. In order to estimate accurate projection matrix, we need to survey the whole scene and label points distributed averagely within the whole scene. To compare with our approach, we manually labeled more than 60 points and mark the corresponding point in the image plane as shown in Figure 7. 20

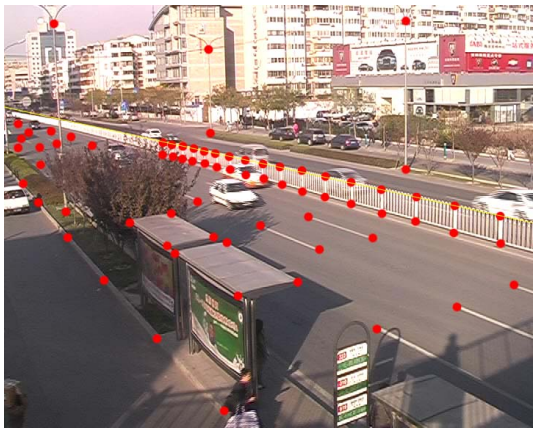


Figure 7. Correspondence of labeled points on image plane

points of them are selected and Direct linear transformation (DLT) method is applied to recover the projection matrix **P** based on the least square strategy as:

$$\mathbf{P} = \begin{bmatrix} -195.67 & 912.46 & 83.77 & 9429158 \\ 50.74 & 17.79 & 897.79 & 9270693 \\ 0.68 & 0.70 & 0.15 & 32739 \end{bmatrix} \quad (11)$$

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

$$\mathbf{P} = \begin{bmatrix} -181.94 & 925.3 & 79.3 & 10504946 \\ 58.87 & 15.8835 & 911.18 & 8403957 \\ 0.84 & 0.51 & 0.19 & 29180 \end{bmatrix} \quad (12)$$

Experiments are conducted to compare the projection matrix calculated by DLT and our method using the other 40 corresponding pairs. We find that our approach gives more than 3% higher accuracy. The possible reason is that labeling of points in surveillance scenes are focus on the ground plane. It is difficult to collect abundant points which are not lied on the ground plane. In addition, manually labeled points cannot cover the whole scene averagely. In contrast, our approach makes use of motion and appearance information of moving objects, which supplies very redundant direction information to achieve accurate calibration.

5.3. Testing with Real Scene Measurement from Images

The performance of camera calibration can be evaluated by measurement of real length ratio from images. As shown in Figure 8, we take one length as unit length and 27 length ratio are measured from images. The measured value and the ground truth of every value are listed in Table 2.

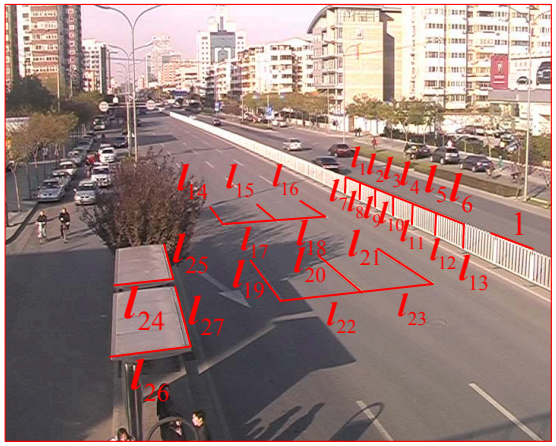


Figure 8. Scene measurement from images

As we can see, the average error of measurement is less than 10% which demonstrate the effectiveness of our approach. Two phenomena from the experimental results show some disadvantages of our approach. The first one is that those 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

Table 2. 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

is that those lines which are parallel to the ground plane are measured more accurately than those perpendicular to the ground plane. That is because the horizonal vanishing points estimated from vehicles are more accurate than the vertical one from pedestrians. More accurate estimation of vertical vanishing points can boost performance of our approach.

5.4. Degenerate Cases

Some degenerate cases may lead to invalidation of our approach. As we know, if the camera plane is parallel or perpendicular to the ground plane, we cannot recover the whole 3 orthogonal vanishing points from videos. In these cases, we should have more information like more vertical or horizonal lines to realize complete camera calibration. Fortunately, surveillance applications always like to mount the camera with a titled angle to the ground plane to cover a wider view field. As a result, these extreme cases are not common at all in surveillance applications.

5.5. Discussion

In the above, we assume that the roadway is straight in the field of view, which is not always true in real applications. Even though the roadway is not straight, there must be one or more approximately straight segments in the field of view. The longest straight segment will generate the global conspicuous peak in voting space to estimate the two orthogonal horizonal vanishing points. As a result, our approach still works in this case.

Another special case is that there are more than one roadway in the field of view. For example, the surveillance scene contains a crossroad as shown in Figure 9. This will lead to two evident peaks in the voting surface for horizonal vanishing points estimation. In most cases, the two roads have not the same traffic flow in a period of time. As a result, the two peaks are of different height so that they can be distinguished from each other. Two groups of three orthogonal vanishing points can be estimated and the camera parameters can be recovered from these two groups with a least square strategy.

Figure 9. Illustration of scene containing crossroad

In our framework, we design a very simple strategy for object classification in videos. Two thresholds are adopted and a part of samples are discarded. There are two reasons for us to use this strategy. The first is that voting based estimation need not very accurate classification. The second is that our calibration result is useful for classification so that it can even be feed back to the classification step to output more accurate result.

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

5.6. Applications

Accurate automatic calibration from videos has great potential to be applied to all kinds of traffic scene surveillance applications. In recent years, appearance based object recognition is more popular than 3D model based method. The reason is that the 3D model based method needs prior camera calibration step which limits its applications. With our approach applied, the 3D model based method can be automatically applied to all kinds of traffic surveillance scenes without manual calibration. Also, classification of objects in surveillance video is difficult due to the perspective distortion of objects. The most common phenomenon is that close objects seems to be larger and move faster than those far away. With our approach applied to object classification, 2D motion and shape features like speed and size

can be normalized to be invariant to view angle changes. In this case, motion and shape features can greatly contribute to the classification accuracy.

6. Conclusions

In this paper, we have proposed a practical camera auto-calibration method for traffic scene surveillance. With only the camera height H measured, we can completely recover both intrinsic and extrinsic parameters of cameras based on appearance and motion of moving objects in videos. Experimental results have demonstrated accuracy and practicability of our approach, which can be used in all kinds of surveillance applications like model based object recognition, coarse object classification and metric measurement from images.

Acknowledgement

This work is funded by research grants from the National Basic Research Program of China (2004CB318110), the National Science Foundation (60605014, 60332010, 60335010 and 2004DFA06900), and the CASIA Innovation Fund for Young Scientists. The authors also thank the anonymous reviewers for their valuable comments.

References

[1] B. Bose and E. Grimson. Ground plane rectification by tracking moving objects. In *Proceedings of the Joint International Workshops on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[2] B. Caprile and V. Grimson. Using vanishing points for camera calibration. *International Journal of Computer Vision*, 4:127–140, 1990.

[3] J. Deutscher, M. Isard, and J. MacCormick. Automatic camera calibration from a single manhttan image. In *Proceedings of European Conference on Computer Vision*, 2002.

[4] O. Faugeras. Three dimensional computer vision: A geometric viewpoint. *MIT Press*, 1993.

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[6] D. Liebowitz, A. Criminisi, and A. Zisserman. Creating architecural models from images. In *Proceedings of EuroGraphics*, 1999.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[8] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9), 2006.

[9] R. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1986.

[10] Z. Zhang. A flexible new technique for camera calibration. In *Proceedings of 7th International Conference on Computer Vision*, 1999.

[11] Z. X. Zhang, Y. Cai, K. Huang, and T. Tan. Real-time moving object classification with automatic scene division. In *In Proc. of International Conference on Image Processing*, 2007.