SELF-CALIBRATION OF TRAFFIC SURVEILLANCE CAMERAS BASED ON MOVING VEHICLE APPEARANCE AND 3-D VEHICLE MODELING

Na Wang[1], Haiqing Du[1], Yong Liu[1], Zheng Tang[2], Jenq-Neng Hwang[2]

[1]Beijing Laboratory of Advanced Information Networks,
Department of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, China,{wangna, duhaiqing}@bupt.edu.cn
[2]Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA

ABSTRACT

This paper proposes an effective and practical method for self-calibration of traffic surveillance cameras. Based on analyzing multiple moving vehicles across multiple frames, the Canny edge detector and Hough transform are first adopted to obtain orthogonal horizontal vanishing points pairs, from which corresponding vertical vanishing points are derived. Next, mean shift clustering and Laplace linear regression are employed to deal with noise and outliers during estimation of vanishing points. To overcome the unreliable estimation issues of orthogonal vanishing points pairs, we further utilize the projective line segments obtained from 3-D vehicle model to create more reliable pairs and iteratively improve the calibration results. Finally, the estimation of distribution algorithm (EDA) is also applied to relax the assumptions made on camera parameters and the moving trajectories of vehicles during the iterations. Experimental results on different datasets prove the feasibility of our proposed scheme.

Index Terms — traffic camera self-calibration, vanishing points, horizon line, 3-D vehicle modeling, EDA optimization

1. INTRODUCTION

Camera calibration, which defines the relationship between 3-D real world and image plane, is an essential work for computer vision tasks, such as 3-D object detection, localization and tracking, and intelligent transportation system applications. Besides, a well-calibrated camera can help to address the problem of object appearance distortion. Due to the difficulties of obtaining useful calibration patterns in most of traffic surveillance, self-calibration has thus drawn more and more attention in recent years.

Many studies have shown that vanishing points are closely related to estimating camera parameters [1-3]. Many researches are devoted to extracting vanishing points from structures (buildings and landmarks) [2] [4-7], walking humans [8-12], and moving vehicles [13]. These methods face major challenges, since a lot of useful inherent information is not available in many traffic scenes.

In view of all the weakness of above methods, we propose a practical traffic camera self-calibration technique based only on moving vehicles without the existence of buildings, walking humans or zebra-crossings. The contributions of proposed approach include: 1) The Canny edge detector and Hough transform are employed, combined with mean shift clustering and Laplace linear regression for noise/outlier reduction, to estimate more precise initial orthogonal horizontal vanishing points pairs than using histogram of oriented gradient (HOG) [13], which can only obtain coarse direction corresponding to the symmetric axis direction and its perpendicular direction. 2) all the self-calibration process is based on vehicles only. No need to rely on the presence of walking pedestrians [3] [10] [13], buildings, traffic lines [5] [12], Manhattan world assumption [4], or assumption on vehicles running along straight roads/lines [13]. 3) line segments with high fitness evaluation scores (FES) obtained from fitting of the 3-D deformable vehicle model are exploited to estimate more accurate vanishing points and further optimize the camera parameters.

The rest of this paper is organized as follows. Section 2 presents our proposed traffic surveillance camera self-calibration processes. Experimental results on different datasets and analyses are shown in Section 3. Finally, we reach the conclusions in Section 4.

2. CAMERA SELF-CALIBRATION PROCESS

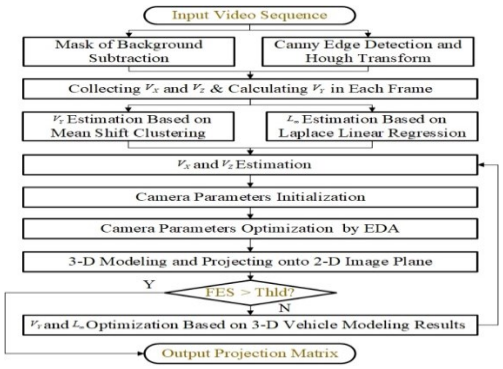
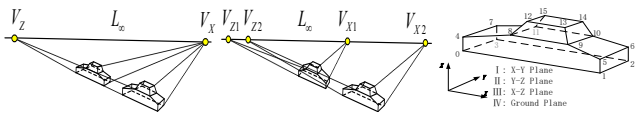


Fig. 1. Overview flow chart of proposed system.



(a) (b) (c)

Fig. 2. A vehicle appears in two consecutive frames. The horizon line (L_∞) and the orthogonal horizontal vanishing points pair V_X and V_Z. (a) The vehicle moves along a straight line. (b) The vehicle makes a slight turn. (c) A 3-D model used for creating parallel lines.

The projections of world parallel lines in an image intersect at a single point called vanishing point. All the camera parameters can be derived from a pair of two orthogonal vanishing points [1]. The overall architecture of our proposed camera self-calibration system is shown in Fig. 1, where the pair of orthogonal vanishing points V_X and V_Z can be initially estimated by extracting foreground information from multiple vehicles across multiple frames from a traffic video, combining with mean shift clustering and Laplace linear regression [14] to reduce the impact of outliers and noise. Further, 3-D vehicle modeling can be iteratively applied into the proposed system to generate more reliable pairs of orthogonal vanishing points and further optimize camera parameters. Meanwhile, EDA optimization is also utilized to find out the optimal camera parameters during the iterations and relax the assumptions by minimizing the reprojection error on the ground plane.

2.1. Orthogonal Vanishing Points Estimation

Figures 2(a)(b) show a vehicle in two consecutive frames. To obtain pairs of the horizontal vanishing points, V_X and V_Z, our first step is to extract parallel lines from the segmented region of a moving vehicle. Background subtraction with Otsu thresholding is adopted to extract the foreground vehicle mask, which can be refined by morphological operations and shadow removal [15] in YCbCr color space. We take advantage of the Canny operator to perform edge detection in the original image region surrounding the foreground mask. The 2-D projected lines in the image plane of two pairs of orthogonal parallel lines in the 3-D real world are detected by Hough transform along the direction of symmetry axis of the vehicle and its perpendicular direction. Examples of extracted lines from images in different videos are shown in Fig. 3.

The horizon line (L_∞), which connects the horizontal vanishing points V_X and V_Z, is considered as the extension of the ground plane at infinity. To ensure that candidates V_X and V_Z are distributed on L_∞, collected projected lines using Hough transform always include the two longest straight lines of the two strongest directions at the bottom of vehicles. Also, only 2 lines in each direction are considered from one vehicle of the same frame and redundant lines will be removed.

Theoretically, if a vehicle runs along a straight road, V_X and V_Z in every frame will lie on the same position (Fig. 2(a)). However, they intersect at different positions but on the same line L_∞ when it is making a turn from frame to frame (see Fig. 2(b)). Unfortunately, the presence of noise and outliers will result in more than one derived L_∞. To derive more accurate L_∞, Laplace linear regression [14] is thus employed.

Fig. 3. Image frames in different videos and corresponding parallel lines (red) found by the Canny operator and Hough transform.

As for estimating the corresponding vertical vanishing point V_Y, we assume the principal point $P(p_u, p_v)$ is located in the center of image preliminarily. This assumption can be relaxed by the EDA optimization described in Section 2.4. Since P is the center of mass of the triangle with vertices, V_X, V_Y and V_Z, we can locate a candidate of V_Y corresponding to each pair of V_X and V_Z candidates, as is demonstrated in Fig. 4(a). Similarly, many V_Y candidates lie on different locations due to the outliers and noise. Hence, mean shift clustering is applied to handle noise and outliers while estimating V_Y [14] since mean shift clustering retains the cluster which has the most candidate points while noise and outliers in small clusters can be suppressed effectively, i.e., can effectively avoid the issue that a number of outliers can overwhelm inliers using RANSAC approach [10]. Specifically, the vehicles that are occluded will be neglected.

To address the problem of estimating the most consistent pair of orthogonal V_X and V_Z based on the locations of P, V_Y and L_∞ [1][3], according to Fig. 4(b), first we choose a random point on L_∞ as V_X. Next, two auxiliary lines, $L_1$and L_2, can be obtained. At last, V_Z can be defined as the intersection of L_2 and L_∞.

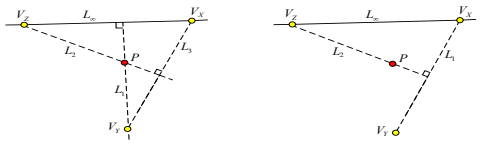
2.2. Computation of Camera Parameters

Vanishing points derived from 3-D parallel lines have been proved to be very helpful for estimating camera parameters [1][3]. Our ultimate target of pinhole self-calibration is to get a projection matrix P so that a 3-D point (X, Y, Z, 1) can be accurately projected onto a 2-D image plane at (u, v, 1):

$$[u, v, 1]^T \sim P \cdot [X, Y, Z, 1]^T = K \cdot [R|T] \cdot [X, Y, Z, 1]^T. \quad (1)$$

The matrix P associated with camera parameters can be decomposed into three matrices, denoted as the *intrinsic parameter matrix* K with five intrinsic parameters (focal length in x direction f_u, focal length in y direction f_v, coordinates of principal point p_u and p_v, and skew value s), the *rotation matrix* R determined by three extrinsic parameters ($roll$ angle around Z-axis γ, $pitch$ angle around X-axis β and yaw angle around Y-axis α), and the *translation matrix* T involving another three extrinsic parameters (t_X moving along X-axis, t_Y moving along Y-axis, and t_Z moving along Z-axis), respectively.

Following previous works [8-10], we first assume the camera with zero skew $s=0$ and equal scale, i.e., $f = f_u = f_v$. In addition, we assume the camera's principal point $P(p_u, p_v)$ to be at the image center. Also, the rough range of camera height is considered to be known. Accurate height value will be derived by EDA in Section 2.4. Except the skew,



(a) (b)

Fig. 4. (a) Vertical vanishing points V_Y derivation based on V_X, V_Z and P. Dotted lines L_1, $L_2$and L_3 are auxiliary lines. (b) V_X and V_Z Localization from V_Y, P, and L_∞. Dotted lines L_1 and L_2 are auxiliary lines.

all the other assumptions will result in reprojection errors. That is why the EDA optimization is further utilized to relax these assumptions later.

To derive translation matrix T, the origin (0, 0, 0) in 3-D world is set to the intersection of perpendicular lines through camera and the ground plane. Therefore, both t_X and t_Z equal to zero. What's more, t_Y is set to the negative of camera height. Accordingly, all camera parameters can be calculated from the orthogonal pair of the vanishing point coordinates $V_X(u_{V_X}, v_{V_X})$ and $V_Z(u_{V_Z}, v_{V_Z})$, which are derived from extracting foreground vehicle edge lines and the estimated location of $P(p_u, p_v)$ [3] [7-10].

2.3. Improvement by 3-D Deformable Vehicle Modeling

It is inevitable to introduce noise and outliers while extracting foreground mask and finding lines in region of vehicle edges. Since all vehicles are rigid body objects with many pairs of orthogonal parallel lines. Accordingly, a 3-D vehicle model is thus employed for identifying more reliable parallel lines. The 3-D vehicle model, which is made up of 16 vertices and 23 arcs as shown in Fig. 2(c), has been proposed to characterize one of 8 different types of vehicles [16]. There are 12 parameters (lengths, widths and so on) to define vehicle's shape and 3 parameters (position coordinates and orientation) for pose information. An effective optimization, called fitness evaluation score (FES), is proposed to find the model by measuring the fitness through comparing gradients along the directions perpendicular to the projected line segments in the image [16].

Based on a calibrated camera, projective wireframe is obtained by projecting a 3-D vehicle model into the image plane. To calculate FES, with the pixel data within the rectangle around the projected line segments, we calculate the gradient directions and magnitude values. If the gradients of pixels have large magnitudes concentrated along the perpendicular direction of a projected line, it can be proved that the projected line segment fits the vehicle well. More specifically, all the 15 parameters in a deformable 3D vehicle model can be estimated by measuring the FES between the projection of model and image data with a method called estimation of multivariate normal algorithm-global ($\text{EMNA}_{\text{global}}$), which aims at solving problems with multiple variables not independent of each other. To reach a better fitting model of a vehicle, for each parameter in the deformable model in its corresponding initial range, the local optimal value can be obtained after a few iterative computations determined by FES. Even if there is a small deviation in camera parameters, 3-D vehicle modeling still performs robust and fits the vehicle in the image well by $\text{EMNA}_{\text{global}}$ [16].

(a) (b) (c) (d)

Fig. 5. Results of 3-D vehicle models using different camera parameters. (a)(b) Results of pairs of parallel lines with high FES from 3-D vehicle models. (c)(d) Camera is well-calibrated.

Corresponding 2-D geometric primitives are obtained by projecting the 3-D model onto image through projection matrix P derived from the previous calibration result and FES can be calculated for segments in the wireframe. If some segments fit the target in the image poorly i.e., FES is lower than an empirical threshold, our system starts to optimize camera parameters by collecting vanishing candidate points using the segments with high FES in the projection wireframe until it reaches a more precise result. In other words, it is controlled by the feed-back based on 3-D vehicle modeling results. Estimating vanishing points using projective lines from 3-D vehicle model achieves better performance than using Hough transform, which is validated by experiments on real datasets. Figure 5 shows 3-D modeling results with different camera parameters. In the projection wireframe, those vertical line segments with high FES (green line segments as shown in Fig. 5(a)(b)) are collected for estimating vertical vanishing points. Similarly, the intersections of two pairs of horizontal parallel lines with high FES (orange line segments as shown in Fig. 5(a)(b)), which are in parallel to the ground plane and corresponding to the symmetrical axis direction and its perpendicular direction, are used for estimating horizontal vanishing points. These horizontal vanishing points should distribute on the L_∞. In spite of projective lines from the 3-D vehicle model fit the vehicle better than lines found by Hough transform, we still inevitably obtain many candidate points of V_Y due to some noise and outliers. At the same time, there are also more than one L_∞, that is why we are calling for mean shift clustering for better estimating V_Y and Laplace linear regression for better estimating accurate L_∞, as is described in Section 2.1.

2.4. Camera Parameters Optimization by EDA

A set of uniform square grid points are created in the 3-D world and they are projected onto the image plane through projection matrix. In theory, the reprojection points are supposed to locate at the intersections of lines connecting all the vanishing points with each edge point of square grid. We define reprojection error as the sum of distance from projection points to the lines and apply EDA to relax the assumptions on intrinsic parameters by searching for camera parameters to minimize reprojection error [14]. Therefore, for

each camera parameter in its corresponding initial range, the local optimal value can be obtained after a few iterative computations. So far, all the assumptions we rely on are that all the vehicles drive on a visible ground plane and the camera approximate height range is known. As a result, the proposed system can be widely used in traffic surveillance camera calibration scenarios since little human intervention or unreasonable assumptions are required.

3. EXPERIMENTAL RESULTS

Several experiments and corresponding discussions are presented to demonstrate the performance of the proposed camera self-calibration system. One of the experiments is conducted on PETS 2000 database [17], which includes vehicles and walking human with camera calibration ground truth (GT). We also use a 5-minute traffic surveillance video of 1920×1080 resolution containing vehicles and human. The ground truth for this dataset is manually computed based on two orthogonal vanishing points [1]. The third video BrnoCompSpeed [18], with available ground truth, is recorded from a highway scene containing many vehicles but few pedestrians and buildings. This video is used to prove the practical necessity of the proposed method.

The experimental settings are described as follows. In 3-D vehicle modeling, the parameters empirically adopted in $\text{EMNA}_{\text{global}}$ are the maximum iteration number 100, and the stopping threshold for the gradient magnitude 2, sample size of the initial population *R=2000*, selected population *N=100*. For the parameters in EDA for optimization of camera parameters, we empirically set *R=2000* and *N=20*. The maximum iteration number is 100, and the stopping threshold of reprojection error ratio between two generations is 0.1. From testing on a number of different simulation parameters, we come to the conclusion that different choices of simulation parameters do not cause too much difference on the performance of the proposed algorithm. The deviation ranges of camera parameters are: $0.1 \times f_u$ for f_u, $0.1 \times f_v$ for f_v, 10 pixels for p_u and p_v, and 20 degrees for γ, β, and α, respectively. Here we choose several algorithms for fair comparisons. First compared method is proposed in [3], which uses located head/foot points based on walking humans, and no additional techniques are employed to handle noise or outliers. The 2nd compared method [10] is based on walking human head/foot localization and RANSAC is used to reduce noise. Another algorithm [13] for comparison relies on both vehicles and human. Moreover, our proposed method without 3-D vehicle modeling optimization nor EDA optimization is also chosen for comparison. Last compared one is our proposed method without only EDA optimization. To demonstrate our method is feasible in the scenes with only moving vehicles, the third dataset with few pedestrians and buildings is specifically tested here, all the other methods cannot work in this scenario. The comparison results of estimated camera parameters and reprojection errors (μ_e) among different schemes are present in Table 1.

As can be seen from the results of first two datasets in Table 1, effective handling of noise and outliers during estimating vanishing points can greatly enhance the accuracy of results. Our proposed method achieves the best overall performance in every dataset. This fully demonstrates that noise and outliers can be dealt with mean shift clustering and Laplace linear regression more efficiently than RANSAC. Besides, we effectively avoid the trouble of fine-tuning the threshold during L_∞ estimation. What's more, the incorporation of 3-D vehicle models fitting to optimize the camera calibration makes results more reliable. In addition, EDA optimization helps to lift the assumptions on intrinsic parameters, which greatly reduce the error. In Seq. 1, method in [13] performs worst because vehicles make a turn in the test video while our scheme can still achieve more accurate results. It is evident that we do not have to restrict vehicles from moving along the straight line/road. Last but not the least, experiments on the third dataset have verified that another advantage of our proposed method requires only moving vehicles without depending on the existence of buildings, pedestrians or zebra crossing lines in the field of view. This will greatly enhance the applicability of our scheme, such as to calibrate the traffic surveillance cameras on the highway.

Estimated Parameters	f_u	f_v	p_u	p_v	γ	β	α	μ_e
1.GT	1360	1360	384	288	0.05	-17.7	-51.2	N/A
1. [3]	-120	-120	0	0	+1.72	+9.8	-4.3	12.4
1. [10]	-108	-108	0	0	+0.81	+7.9	-7.9	8.8
1. [13]	-233	-233	0	0	+2.08	-7.4	+4.0	25.9
1. no FES	-26	-26	0	0	+0.42	+1.2	+3.6	6.0
1. no EDA	-22	-22	0	0	+0.21	+0.7	+1.4	3.3
1. Proposed	-14	-17	+2	0	-0.24	+0.7	+0.8	5.2E-3
2. GT	1853	1851	954	540	0.16	-19.1	-57.2	N/A
2. [3]	-285	-283	+6	0	+2.11	-0.7	+13.7	39.5
2. [10]	-235	-233	+6	0	+1.74	-5.4	+11.6	35.8
2. [13]	-141	-139	+6	0	+2.19	-6.1	+7.0	28.6
2. no FES	-31	-29	+6	0	+0.85	-2.1	+3.6	12.0
2. no EDA	-20	-18	+6	0	+1.09	-1.2	+3.0	10.1
2. Proposed	-11	-14	+6	-1	-0.54	-1.1	+2.2	1.9E-3
3. GT	1368	1369	962	539	-3.51	-21.3	-49.2	N/A
3. no FES	-36	-37	-2	+1	+1.11	-1.7	+1.4	14.3
3. no EDA	-25	-26	-2	+1	+0.66	-1.6	+0.6	5.4
3. Proposed	-17	-21	+2	+1	+0.49	-1.1	-0.9	4.3E-3

Table 1. Results of differences between estimated camera parameters and ground truth using different methods on different test datasets (f_u, f_v, p_u, p_v, μ_e, unit=pixel; γ, β, α unit=degree). GT denotes ground truth.

4. CONCLUSION

This paper presents an effective and practical self-calibration method for traffic surveillance cameras solely based on moving vehicles without relying on the availabilities of humans, buildings, or traffic lines. With only two orthogonal vanishing points and an approximate range of camera height, we can systematically and reliably estimate all the camera parameters. We take advantage of iterative 3-D deformable vehicle model fitting for further camera parameters optimization to enhance the system performance. Moreover, EDA is also adopted in the optimization loop to relax the assumption on camera intrinsic parameters.

5. REFERENCES

[1] B. Caprile and V. Torre, "Using vanishing points for camera calibration," *Int. J. Computer Vision* (IJCV), vol. 4, no. 2, pp. 127-139, 1990.

[2] R. Cipolla, T. Drummond, and D. P. Robertson, "Camera calibration from vanishing points in image of architectural scenes," in *Proc. British Machine Vision Conference* (BMVC), Nottingham, 1999.

[3] F.-J. Lv, T. Zhao and R. Nevatia, "Self-calibration of a camera from video of a walking human," in *Proc. IEEE Int. Conf. Pattern Recognition* (ICPR), vol. 1, pp. 562-567, 2002.

[4] J. Deutscher, M. Isard, and J. Maccormick, "Automatic Camera Calibration from a Single Manhattan Image," in *European Conf. on Computer Vision* Springer-Verlag, pp.175-205, 2002

[5] X. Lu, Y. Wang, Z. Ling, K. Wang, and G. Wang, "A method for vehicle-mounted camera calibration under urban traffic scenes," in *Chinese Automation Congress* IEEE, pp.556-560, 2014

[6] T. N. Schoepflin and D. J. Dailey, "Dynamic camera calibration of roadside traffic management cameras for vehicle speed estimation," *IEEE Trans. on Intelligent Transportation Systems*, vol. 4, no. 2, pp. 90-98, June 2003.

[7] J. P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *Proc. IEEE Int. Conf. on Computer Vision*, Kyoto, pp. 1250-1257, 2009.

[8] F. Lv, Tao Zhao and R. Nevatia, "Camera calibration from video of a walking human," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1513-1518, Sept. 2006.

[9] N. Krahnstoever and P. R. S. Mendonça. "Autocalibration from tracks of walking people," in *Proc. British Machine Vision Conference 2006*, Edinburgh, pp.107-116, 2013

[10] Q. Wu, T.-C. Shao and T. Chen. "Robust self-calibration from single image using RANSAC," in *Int. Conf. on Advances in Visual Computing,* Springer Berlin Heidelberg, pp. 230-237, 2007.

[11] W. Kusakunniran, H. Li and J. Zhang, "A direct method to self-calibrate a surveillance camera by observing a walking pedestrian," in *Proc. Int. Conf. Digital Image Computing: Techniques and Applications* (DICTA), pp. 250-255, 2009.

[12] M. Hodlmoser, B. Micusik, and M. Kampel, "Camera auto-calibration using pedestrians and zebra-crossings," in *Proc. IEEE Int. Conf. on Computer Vision Workshops.* pp. 1697-1704, 2011.

[13] Z. Zhang, M. Li, Huang K, et al. "Practical camera auto-calibration based on object appearance and motion for traffic scene visual surveillance," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 1-8, 2008.

[14] Z. Tang, Y. S. Lin, K. H. Lee, J. N. Hwang, J. H. Chuang and Z. Fang, "Camera self-calibration from tracking of moving persons," in *Proc. Int. Conf. on Pattern Recognition* (ICPR), pp. 265-270, 2017.

[15] Z. Tang, J. N. Hwang, Y. S. Lin and J. H. Chuang, "Multiple-kernel adaptive segmentation and tracking (MAST) for robust object tracking," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), pp. 1115-1119, 2016.

[16] 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.

[17] PETS Data Base [Online]. Available FTP:.uk/pub/PETS2000/, accessed 2000.

[18] J. Sochor, R. Juránek, J. Špaňhel, et al, "BrnoCompSpeed: Review of Traffic Camera Calibration and Comprehensive Dataset for Monocular Speed Measurement," 2017.