Camera Calibration from Periodic Motion of a Pedestrian

Shiyao Huang, Xianghua Ying*, Jiangpeng Rong, Zeyu Shang and Hongbin Zha
Key Laboratory of Machine Perception (Ministry of Education)
School of Electronic Engineering and Computer Science, Center for Information Science
Peking University, Beijing 100871, P.R. China
{h41, xhying, rjp, shangzeyu, zha}@cis.pku.edu.cn

Abstract

Camera calibration directly from image sequences of a pedestrian without using any calibration object is a really challenging task and should be well solved in computer vision, especially in visual surveillance. In this paper, we propose a novel camera calibration method based on recovering the three orthogonal vanishing points (TOVPs), just using an image sequence of a pedestrian walking in a straight line, without any assumption of scenes or motions, e.g., control points with known 3D coordinates, parallel or perpendicular lines, non-natural or pre-designed special human motions, as often necessary in previous methods. The traces of shoes of a pedestrian carry more rich and easily detectable metric information than all other body parts in the periodic motion of a pedestrian, but such information is usually overlooked by previous work. In this paper, we employ the images of the toes of the shoes on the ground plane to determine the vanishing point corresponding to the walking direction, and then utilize harmonic conjugate properties in projective geometry to recover the vanishing point corresponding to the perpendicular direction of the walking direction in the horizontal plane and the vanishing point corresponding to the vertical direction. After recovering all of the TOVPs, the intrinsic and extrinsic parameters of the camera can be determined. Experiments on various scenes and viewing angles prove the feasibility and accuracy of the proposed method.

1. Introduction

In the field of computer vision, camera calibration is one of the most fundamental issues for many applicants including 3D reconstruction, object recognition, metrology, and surveillance. Considerable efforts have been made to compute the intrinsic and extrinsic parameters of the camera with consideration of speed, accuracy and robustness [8, 10, 17, 26]. Vanishing point based calibration methods have been proved to be suitable for the situation of structured scenes. Caprile and Torre [3] proposed a method to calibrate a camera with known aspect ratio and skew from a single view of the TOVPs. They demonstrated that the principal point of the camera coincides with the orthocenter of the triangle with vertices being the TOVPs. Cipolla *et al.* [4] proposed a simple and geometrically intuitive method using the TOVPs and one reference point to determine both intrinsic and extrinsic parameters, and the method was realized with various viewpoints in indoor and outdoor architectural scenes when the TOVPs are available.

Since calibration objects are often absent in surveillance scenes and parameters of surveillance cameras may be changed over time, using vanishing points extracted from image sequences of a pedestrian seems to be an imperfect, but not-that-bad choice for camera calibration, though the extracted vanishing points are not always as accurate as those in structured scenes like buildings with sufficient and strong rigidity constraints of parallelism and orthogonality. Lv *et al.* [16] recovered the vertical vanishing point and the horizon line by detecting leg-crossings of a walking human. In order to determine the two orthogonal vanishing points in the horizontal plane, they need to point out two orthogonal lines on the ground which must be simultaneously taken with the pedestrian. Krahnstoever and Mendonca [13] presented a Bayesian method for calibration using the foot-to-head homology acquired from the visual surveillance of human activity, where the probabilistic spatial distribution of the tracks on the horizontal plane is required. Junejo and Foroosh [11] used the detected head and feet locations to compute two epipoles as two orthogonal vanishing points. They assumed that the camera's intrinsic parameters are almost fully known except for the focal length. The Total Least Squares method was applied to the observation points to estimate the focal length. Micusik and Pajdla [18] proposed a method for automatic simultaneous camera calibration and the foot-head homology estimation by observing a person standing at various locations in the scene with the same pose, e.g., standing to attention and facing the same direction in front of the camera. They formulated the calibration of intrinsic and extrinsic camera parameters as a Quadratic Eigenvalue Problem. Kusakunniran *et al.* [14] utilized the cross-ratio relationship in projective geometry to directly estimate a full projection matrix. However, they required observing three or more positions of person walking on a ground plane, where the three positions are not collinear. It means

* Corresponding Author

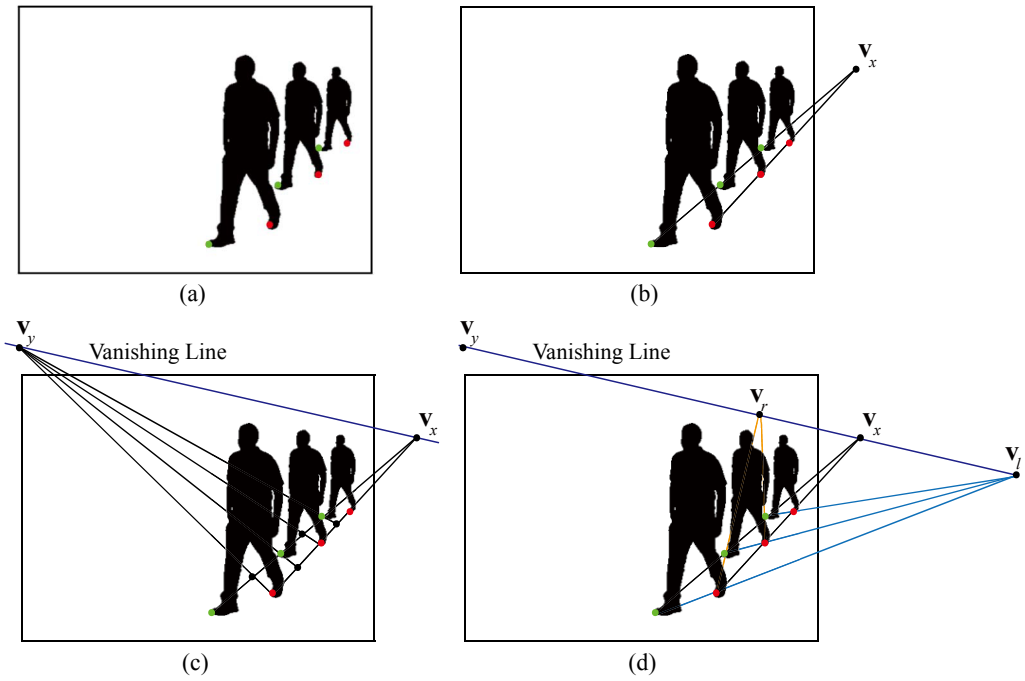


Figure 1. Recover the vanishing point $\mathbf{v}_x$ corresponding to the walking direction, and the vanishing point $\mathbf{v}_y$ corresponding to the perpendicular direction of the walking direction in the horizontal plane. (a) Extract the toes of the shoes from the pedestrian blobs in the lambda-shaped frames when the two legs are maximally separated. Left toes are marked as red, and right toes as green. (b) Use the left and right toes respectively to construct two image lines. The vanishing point $\mathbf{v}_x$ is their intersection. (c) Use harmonic conjugate properties in projective geometry to recover the images of the midpoints of adjacent toes, marked as black, and construct lines perpendicular to the walking direction on the ground, the vanishing point $\mathbf{v}_y$ is their intersection. (d) Another method to detect $\mathbf{v}_y$: Acquire two other vanishing points $\mathbf{v}_l$ and $\mathbf{v}_r$ on the horizon line, and construct harmonic conjugate system to determine $\mathbf{v}_y$ (see Section 4 for details).

that the person would walk around in the field of view of the camera. It is not difficult to find out that all these previous methods require assumption or prior knowledge about scenes [16], non-natural or pre-designed special motion [14, 18], or only two orthogonal vanishing points extracted but not all of the TOVPs [13, 11].

In this paper, we are motivated to develop a method which does not necessitate any assumption of specific scenes or motions but just exploits the periodic motion of human walking in a straight line. In the absence of favorable information in the scenes or pre-designed special motion, we recover the TOVPs just from the image sequence of a pedestrian walking in a straight line. Indeed, for real surveillance scenes, pedestrians often walk only one pass in the field of view of the camera, unusually walk around as [14, 18]. In many cases, the assumption of walking approximate in straight line is not very difficult to be satisfied, since the minimal data for our calibration method are just continuous three steps, namely, four continuous shoe prints on the ground. The main contribution of the paper is that we consider the shoe prints as the stable and easily detected features in the image sequence of pedestrians, and the TOVPs are recovered from these features more robustly, and then for camera calibration. To the best of our knowledge, this is the first work showing that it is possible to calibrate camera through the images of shoes of pedestrians. Since the techniques for detecting of periodic motion of human [5, 15, 20, 1, 21] and foot pose estimation [19, 12] have been well exploited, we are able to extract the toes of the shoes on the ground plane efficiently from the periodic motion. We are especially interested in special case when the two legs are maximally separated as shown in Figure 1, since it corresponds to a critical phase where the toe-to-toe distance reaches maximum value, and usually shoes contact the ground plane. We called such case as "lambda-shaped" one. Since the toes of the shoes on the ground plane will keep fixed on the ground for a relatively long time, they can be detected easily and robustly, with comparisons of the head and feet locations as used in [13, 16, 11, 18]. We divide the set of detected toes into two sets related to the left and right toes, respectively. The vanishing point corresponding to the walking direction can be detected by computing the intersection of the two parallel lines formed by the left and right toes. Furthermore, we use the harmonic conjugate properties in projective geometry to recover the vanishing point corresponding to the perpendicular direction of the walking direction in the horizontal plane and the vertical vanishing point. After recovering all of the TOVPs, the intrinsic and extrinsic parameters of the camera can be determined from the detected TOVPs.

2. Overview of the proposed method

When a pedestrian is walking in a straight line, it easily ensures the pace of a pedestrian as constant within a short period, thus the adjacent toes corresponding to the same shoe on the ground plane are equidistant. If we divide the detected toes into two sets related to the left and right toes, we can acquire two sets of 2D points whose correspondences in 3D are equidistant and respectively lying on two parallel lines (see Figure 2a). During the periodic motion of human walking, the two legs will separate to the max distance when the front shoe just touches the ground, which will remain almost stationary on the ground plane for a relatively long time until the back shoe moves forward and replaces it as the front shoe. We define this special case when the two legs are maximally separated as "lambda-shaped" one, and its corresponding frames in the image sequence as lambda-shaped frames.

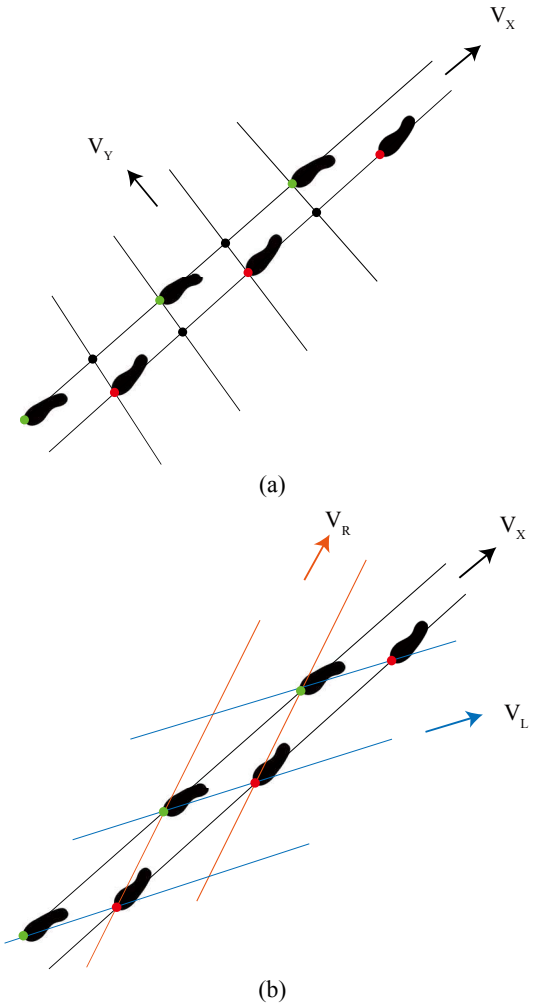


Figure 2. Top view of shoe prints, related to Figure 1. (a) The constructing procedures corresponding to Figure 1c. (b) The constructing procedures corresponding to Figure 1d.

Since the toes of the shoes on the ground plane will keep fixed on the ground for a relatively long time, they can be detected easily and robustly.

Given an image sequences of a pedestrian walking in a straight line, we extract the toes of the shoes from the pedestrian blobs in the lambda-shaped frames, as shown in Figure 1a (see Section 3 for details). We construct two image lines corresponding to the left and right toes on the ground plane, respectively, and then the vanishing point corresponding to the walking direction, $\mathbf{v}_x$ can be determined as shown in Figure 1b.

The images of the midpoints of adjacent toes can be recovered by constructing harmonic conjugate systems, with the help of $\mathbf{v}_x$, as shown in Figure 1c and Figure 2a. We connect the midpoints and its corresponding toes in the other side in order to construct lines perpendicular to the direction of the walking direction on the horizontal plane, as shown in Figure 2a. Then, the vanishing point $\mathbf{v}_y$ is determined by computing their common intersection, as shown in Figure 1c. We also propose another approach to detect $\mathbf{v}_y$: Firstly connect the left and right toes to get two groups of lines, as shown in Figure 2b. Secondly compute the common intersections of these lines to acquire two novel vanishing points $\mathbf{v}_l$ and $\mathbf{v}_r$ on the horizon line. Finally, construct harmonic conjugate system to determine $\mathbf{v}_y$, as shown in Figure 1d (see Section 4 for details).

The detection of the vertical vanishing point $\mathbf{v}_z$ is illustrated in Section 4.3 (see Figure 6). The detailed implementation of calibration algorithm using the recovered TOVPs is described in Section 5.

3. Detect image points of toes on ground plane

3.1. Extract pedestrians in lambda-shaped frames

Given an image sequence, a statistical background model [8] can help extract the moving foreground objects. Many efforts have been made to detect pedestrian [7, 6, 23, 25]. For each frame of the sequence, the blob of the pedestrian can be fast extracted with a blob tracker if no strong shadow exists [10]. We need to first pick out the lambda-shaped frames to enable the next step to detect the toe positions from the corresponding blobs in these frames.

We first apply PCA to blob in each of frames. Denote the first and second eigenvalue of the covariance matrix at i-th frame as $v_i^{(1)}$ and $v_i^{(2)}$, then we define $k_i = v_i^{(2)}/v_i^{(1)}$, where the superscripts (1) and (2) are used to distinguish the first and second eigenvalue. The curve related to k_i as shown in Figure 3a has the following properties: It reaches peaks in lambda-shaped frames and valleys in leg-crossing frames, thus pick out the peaks and we can determine the lambda-shaped frames, as shown in Figure 3a. Due to the tiny change of step frequency and noise, the curve may not keep to a strictly fixed period and appear unsmooth, we can apply quadratic curve fitting around the local peaks to deal with

it, as illustrated in Figure 3b. If the included angle between the walking direction and the projection of the viewing direction on the ground plane is too small, the two legs in the blobs will not separate, thus the above procedure cannot pick out the lambda-shaped frames exactly but just provides some initials for the lambda-shaped frames. We need to detect the fixed pixels of the blobs around the initial frames. Once the percentage of the fixed pixels has been higher than the preset threshold starting from some frames, it means the front shoe has been fixed on the ground, and the lambda-shaped frames are determined.

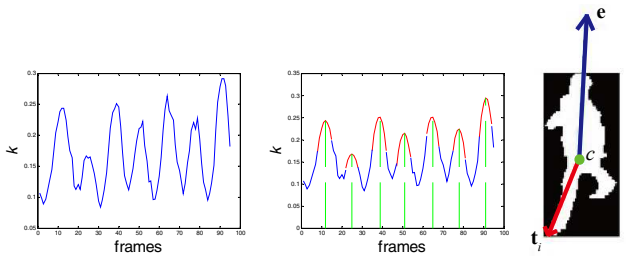
3.2. Detect toe positions on pedestrian blobs

If the colors of the pedestrian's shoes favorably contrast with those of the background, the shoes can be extracted perfectly [12]. Nevertheless, a pedestrian in a surveillance scene does not necessarily wear particular shoes. Therefore we provide a method to detect the front toe position from a pedestrian's blob in the lambda-shaped frame.

We denote the blob's center as c and its first eigenvector as $\mathbf{e}$, illustrated in Figure 3c. For each pixel in the blob, we define the vector from c to the i-th accessed pixel as $\mathbf{t}_i$, initial toe position $\mathbf{f}$ is the pixel that corresponds to the minimal dot product of $\mathbf{e}$ and $\mathbf{t}$:

$$\mathbf{f} = \arg\min_{\mathbf{t}}(\mathbf{e} \cdot \mathbf{t}) \tag{1}$$

Note that the minimal dot product is a negative value. In case that two or more pixels correspond to the minimal scalar product, we choose the one that is most apart from the principle axis as the optimal initial toe position. Making use of the property that the toes will keep fixed on the ground for a relatively long time, we can optimize the initial toe position $\mathbf{f}$ to acquire the refined toe position on the ground plane.



(a) (b) (c)

Figure 3. Extract the front toe position of a pedestrian in a lambda-shaped frame. (a) The plot of $k_i = v_i^{(2)}/v_i^{(1)}$. (b) Adopt quadratic curve fitting near the local peaks of (a). Pick out the curve peaks to determine the lambda-shaped frames. (c) A pedestrian's blob in a lambda-shaped frame. c is the center position. $\mathbf{e}$ is the first eigenvector. For each pixel of the blob, $\mathbf{t}_i$ is the vector from c to the i-th accessed pixel. The initial toe position is the pixel that corresponds to the minimal dot product of $\mathbf{e}$ and $\mathbf{t}$.

4. Recover the three orthogonal vanishing points

4.1. Recover the vanishing point corresponding to the walking direction

Denote the left and right toe positions as $\{\mathbf{f}_i^{(l)}\}_{i=1,\ldots,\mathrm{M}}$ and $\{\mathbf{f}_j^{(r)}\}_{j=1,\ldots,\mathrm{N}}$, where $|\mathrm{M}-\mathrm{N}| \le 1$, and the superscripts (l) and (r) are used to distinguish the left and right toes, and the subscripts i and j run over all left and right toes (see Figure 4). Let the lines best fitting these points be $\mathbf{S}_l = (\mathbf{w}_l, b_l)$ and $\mathbf{S}_r = (\mathbf{w}_r, b_r)$, where $\mathbf{w}^{\mathrm{T}}\mathbf{f} + b = 0$, and the subscript l and r are used to distinguish the lines related to the left and right toes. $\mathbf{S}_l$ and $\mathbf{S}_r$ are easily determined as:

$$(\mathbf{w}^*, b^*) = \arg\min_{(\mathbf{w}^*, b^*)} \sum_k \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{f}_k + b|}{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}_k\mathbf{w})^{1/2}} \tag{2}$$

where $\boldsymbol{\Sigma}_k$ is the covariance matrix of $\mathbf{f}_k$, and the subscript k runs over all left toes or right toes. The vanishing point corresponding to the walking direction on the horizontal plane $\mathbf{v}_x$ is then detected by computing the intersection of $\mathbf{S}_l$ and $\mathbf{S}_r$ (see Figure 4).

4.2. Recover the vanishing point corresponding to the perpendicular direction of the walking direction on the horizontal plane

We present two approaches to detect the vanishing point $\mathbf{v}_y$ corresponding to the perpendicular direction of the walking direction on the horizontal plane in Step 1 and Step 2, and combine them into a unique solution in Step 3.

Step 1: A natural idea to recover $\mathbf{v}_y$ is to find two or more image lines whose corresponding 3D lines are mutually parallel and perpendicular to the walking direction, then $\mathbf{v}_y$ can be determined by computing common intersection of these image lines. We propose a construction procedure of these desired image lines as follows: Denote $\mathbf{m}_k^{(l)}$ as the image point of the midpoint related to $\mathbf{f}_k^{(l)}$ and $\mathbf{f}_{k+1}^{(l)}$ in 3D, and similarly denote $\mathbf{m}_k^{(r)}$ related to $\mathbf{f}_k^{(r)}$ and $\mathbf{f}_{k+1}^{(r)}$ (see Figure 4).

We use the harmonic conjugate properties in the projective geometry to compute $\mathbf{m}_k^{(l)}$ and $\mathbf{m}_k^{(r)}$. Assume that in 3D space, $M^{(l)}$ is the midpoint of $F_1^{(l)}$ and $F_2^{(l)}$, X is the point at infinity in the line determined by $F_1^{(l)}F_2^{(l)}$. The cross ratio [22] of these four points is determined as:

$$\left(F_1^{(l)}F_2^{(l)}, M^{(l)}X\right) = \frac{(F_1^{(l)}F_2^{(l)}M^{(l)})}{(F_1^{(l)}F_2^{(l)}X)} = \frac{F_1^{(l)}M^{(l)}}{F_2^{(l)}M^{(l)}} \cdot \frac{F_2^{(l)}X}{F_1^{(l)}X} = -1 \tag{3}$$

As the cross ratio of these four points equals -1, we say that $F_1^{(l)}$, $F_2^{(l)}$, $M^{(l)}$ and X make up a harmonic system of points,

or $F_1^{(l)}$ and $F_2^{(l)}$ are harmonic conjugate points relative to $M^{(l)}$ and X. According to the property of projection transformation that cross ratio is a projective invariant [22], the projections of these four points $F_1^{(l)}$, $F_2^{(l)}$, $M^{(l)}$ and X in the image plane, namely $\mathbf{f}_k^{(l)}$, $\mathbf{f}_{k+1}^{(l)}$, $\mathbf{m}_k^{(l)}$, and $\mathbf{v}_x$, also satisfy the harmonic conjugate relationship:

$$\left(\mathbf{f}_k^{(l)}\mathbf{f}_{k+1}^{(l)}, \mathbf{m}_k^{(l)}\mathbf{v}_x\right) = \left(F_1^{(l)}F_2^{(l)}, M^{(l)}X\right) = -1 \tag{4}$$

$\mathbf{m}_k^{(l)}$ can be determined by solve (4). Use the exactly same properties we can also compute $\mathbf{m}_k^{(r)}$ from the harmonic conjugate relationship:

$$\left(\mathbf{f}_k^{(r)}\mathbf{f}_{k+1}^{(r)}, \mathbf{m}_k^{(r)}\mathbf{v}_x\right) = -1 \tag{5}$$

Construct parallel lines using image points pairs: $\mathbf{m}_k^{(l)}$ and $\mathbf{f}_k^{(r)}$, $\mathbf{m}_k^{(r)}$ and $\mathbf{f}_{k+1}^{(l)}$ in sequence, thus we can construct totally T such lines ($\mathrm{T} = \mathrm{M} + \mathrm{N} - 2$). After computing all the midpoints, $\mathbf{v}_y$ can be robustly determined: Re-denote the parallel lines we construct as $\{(\mathbf{p}_t^{(l)}, \mathbf{p}_t^{(r)})\}_{t=1,\ldots,\mathrm{T}}$, and $\{(\boldsymbol{\Sigma}_{\mathbf{p}_t^{(l)}}, \boldsymbol{\Sigma}_{\mathbf{p}_t^{(r)}})\}_{t=1,\ldots,\mathrm{T}}$ as the associated covariance matrices, where the subscript t runs over all parallel lines. Denote $\mathbf{q}_t$ as the midpoint of $\mathbf{p}_t^{(l)}$ and $\mathbf{p}_t^{(r)}$, $\mathbf{v}_y$ is then detected as:

$$\mathbf{v}_y = \arg\min_{\mathbf{v}} \sum_{t=1}^{T} \left(\frac{\left|\mathbf{w}_t^{\mathrm{T}}\mathbf{p}_t^{(l)} + b_t\right|}{\left(\mathbf{w}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{p}_t^{(l)}}\mathbf{w}_t\right)^{\frac{1}{2}}} + \frac{\left|\mathbf{w}_t^{\mathrm{T}}\mathbf{p}_t^{(r)} + b_t\right|}{\left(\mathbf{w}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{p}_t^{(r)}}\mathbf{w}_t\right)^{\frac{1}{2}}} \right) \tag{6}$$

where $(\mathbf{w}_t, b_t)$ is the line determined by $\mathbf{q}_t$ and $\mathbf{v}$.

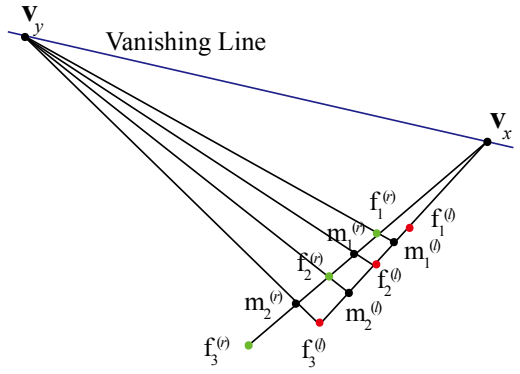


Figure 4. The first approach to recover the vanishing point $\mathbf{v}_y$ corresponding to the perpendicular direction of the walking direction on the horizontal plane. Use the harmonic conjugate property to compute the images of the midpoints $\mathbf{m}_k^{(l)}$ and $\mathbf{m}_k^{(r)}$ of adjacent toes, then construct lines connecting the midpoints and the corresponding toes of the other shoe, $\mathbf{v}_y$ is detected by computing their common intersection.

Step 2: $\mathbf{v}_y$ can be recovered in another way which constructs harmonic conjugate system by vanishing points directly. As illustrated in Figure 5a, we can acquire a group of parallel lines by connecting $\mathbf{f}_k^{(l)}$ with $\mathbf{f}_{k+1}^{(r)}$ in sequence, and another group of parallel lines by connecting $\mathbf{f}_k^{(r)}$ and $\mathbf{f}_{k+1}^{(l)}$. Denote the common intersection of the first group of lines as $\mathbf{v}_l$, and the common intersection of the second group of lines as $\mathbf{v}_r$. These two intersections are both vanishing points collinear with $\mathbf{v}_x$ and $\mathbf{v}_y$ on the horizon line. We can also construct harmonic conjugate system to detect $\mathbf{v}_y$ as:

$$\left(\mathbf{v}_l, \mathbf{v}_r, \mathbf{v}_x, \mathbf{v}_y\right) = -1 \tag{7}$$

The reason is as follows: As shown in Figure 5b, X is the point at infinity in the direction of walking, Y is the point at infinity in the direction perpendicular to the direction of walking, L and R are the points at infinity respectively corresponding to the direction of the lines connecting the left and right toes in the specific way we described above (see Figure 2b), O can be any point not at infinity. Denote four lines $l = OA, r = OB, x = OX, y = OY$, then the cross ratio of these four lines equals the cross ratio of the four points [22], i.e.:

$$(LR, XY) = (lr, xy) \tag{8}$$

Due to the symmetry of left and right toes during the periodic motion of straight walking: x is the internal angular bisector of $\angle\mathrm{LOR}$, y is the external angular bisector of $\angle\mathrm{LOR}$, namely, $\alpha = \beta, \theta = \varphi$ (see Figure 5b). Thus we can compute the cross ratio of the four lines as well as that of the four points:

$$\begin{aligned}(LR, XY) = (lr, xy) &= \frac{\sin(l,x)}{\sin(r,x)} \cdot \frac{\sin(r,y)}{\sin(l,y)} \\ &= \frac{\sin(-\alpha)}{\sin(\beta)} \cdot \frac{\sin(-\theta)}{\sin(-\varphi)} = -1\end{aligned} \tag{9}$$

According to the property of projection transformation that cross ratio is a projective invariant, the correspondences of the four points L, R, X, Y in the image plane, namely $\mathbf{v}_l$, $\mathbf{v}_r$, $\mathbf{v}_x$, $\mathbf{v}_y$, also satisfy the harmonic conjugate relationship:

$$\left(\mathbf{v}_l, \mathbf{v}_r, \mathbf{v}_x, \mathbf{v}_y\right) = (LR, XY) = -1 \tag{10}$$

Since $\mathbf{v}_l, \mathbf{v}_r, \mathbf{v}_x$ have been determined, by solving equation (10), we can obtain $\mathbf{v}_y$.

Step 3: Denote $\mathbf{v}_y$ acquired in step 1 and step 2 as $\mathbf{v}_y^{(1)}$ and $\mathbf{v}_y^{(2)}$, respectively. We combine both results to uniquely determine $\mathbf{v}_y$ in the balance of different situations. If the viewpoint's height is similar to the pedestrin's height, the two lines fitting the left toes and right toes are close to overlap, we tend to prefer $\mathbf{v}_y^{(2)}$ instead of $\mathbf{v}_y^{(1)}$ because in the first approach, the extremely short distance between a

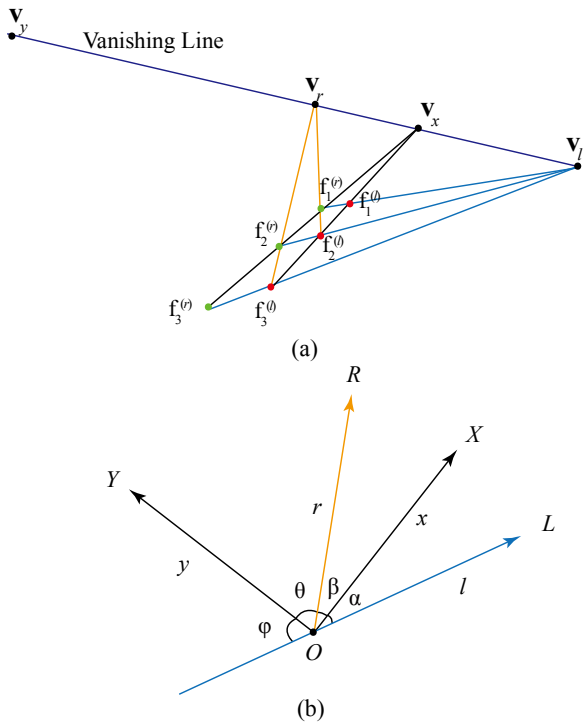


Figure 5. The second approach to recover the vanishing point $\mathbf{v}_y$ corresponding to the perpendicular direction of the walking direction on the horizontal plane. (a) Connect the adjacent left and right toes to acquire two groups of lines, then compute their common intersections to locate $\mathbf{v}_l$ and $\mathbf{v}_r$, finally construct harmonic conjugate system to detect $\mathbf{v}_y$. (b) Top view of the constructed harmonic conjugate system.

midpoint on one side and the toe on the other side makes the construction of parallel lines difficult to implement. If the viewing direction is nearly parallel with the walking direction, $\mathbf{v}_y^{(1)}$ is more reliable than $\mathbf{v}_y^{(2)}$, as in the second approach, the lines connecting the left and right toes do not always intersect at a point in the horizon line due to noise and outliers. Except for these two extreme conditions, we utilize both $\mathbf{v}_y^{(1)}$ and $\mathbf{v}_y^{(2)}$ to determine $\mathbf{v}_y$, and employ a parameter to reasonably balance their weights.

4.3. Recover the vertical vanishing point

We recover the vanishing point $\mathbf{v}_z$ corresponding to the vertical direction by constructing W vertical poles and compute their common intersection ($\mathrm{W} = \mathrm{M} + \mathrm{N} - 1$). As shown in Figure 6a, the top of a pole is the head position in a lambda-shaped frame, which is the tangent point on the common tangent line to all the lambda-shaped blobs, and the tangent line should pass through $\mathbf{v}_x$. The corresponding bottom is the midpoint of the two toes in a lambda-shaped frame, which can also be acquired by using the harmonic conjugate property. Denote the image point of midpoint related to $\mathbf{f}_k^{(l)}$ and $\mathbf{f}_k^{(r)}$ in 3D as $\mathbf{n}_k^{(l)}$, and the image point of midpoint related to $\mathbf{f}_k^{(r)}$ and $\mathbf{f}_{k+1}^{(l)}$ in 3D as $\mathbf{n}_k^{(r)}$, as shown in Figure 6b. By utilizing the harmonic conjugate properties, and $\mathbf{v}_l$ and $\mathbf{v}_r$ being available, we can compute $\mathbf{n}_k^{(l)}$ and $\mathbf{n}_k^{(r)}$ as:

$$\left(\mathbf{f}_k^{(l)}\mathbf{f}_k^{(r)}, \mathbf{n}_k^{(l)}\mathbf{v}_l\right) = \left(\mathbf{f}_k^{(r)}\mathbf{f}_{k+1}^{(l)}, \mathbf{n}_k^{(r)}\mathbf{v}_r\right) = -1 \qquad (11)$$

By computing the common intersection of these vertical poles, the vanishing point $\mathbf{v}_z$ is detected with the similar approach to detect $\mathbf{v}_y$ as illustrated in (6).

5. Calibration algorithm

5.1. Calibration with the recovered TOVPs

If a camera with zero skew and unit aspect ratio, i.e., a three-parameter camera model, the camera parameters left for us to determine are the focal length f, the principal point $(u_0, v_0)^{\mathrm{T}}$, the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{T}$. Now we have obtained the TOVPs as shown in previous sections. The calibration algorithm goes as the following steps:

Determine the intrinsic matrix: For triangle $\mathbf{v}_x\mathbf{v}_y\mathbf{v}_z$, its orthocenter coincides with the principal point $\mathbf{p} =$

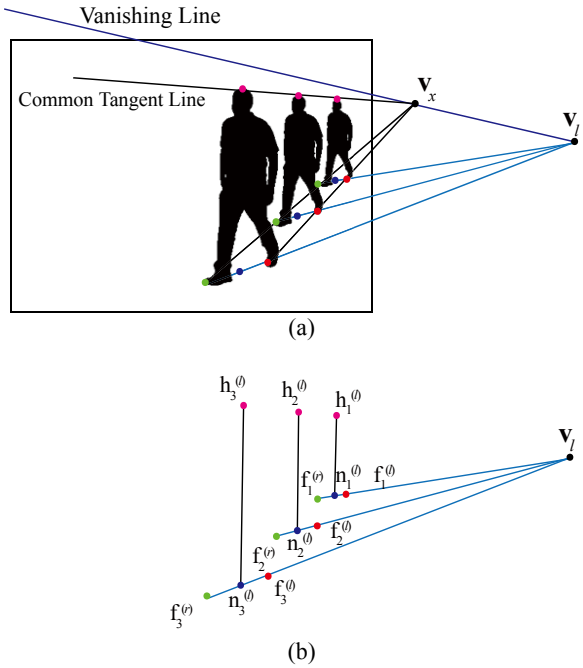


Figure 6. Construct vertical poles to detect the vertical vanishing point $\mathbf{v}_z$. (a) The head positions are the tangent points on the common tangent line to all of the lambda-shaped blobs, and marked as pink. The tangent line should pass through $\mathbf{v}_x$. The midpoints of two toes in lambda-shaped frames are computed using harmonic conjugate property, and marked as blue. (b) Connect the head positions and corresponding midpoints to construct vertical poles. $\mathbf{v}_z$ is detected by computing their common intersection.

$(u_0, v_0)^{\mathrm{T}}$of the image plane. If denote the coordinates of $\mathbf{v}_x$, $\mathbf{v}_y$ and $\mathbf{v}_z$ as $(x_1, y_1)^{\mathrm{T}}$, $(x_2, y_2)^{\mathrm{T}}$ and $(x_3, y_3)^{\mathrm{T}}$, when the origin of the coordinate system is located at the principal point, then focal length f can be determined by the following equations:

$$\begin{cases} x_1x_2 + y_1y_2 + f^2 = 0 \\ x_2x_3 + y_2y_3 + f^2 = 0 \\ x_3x_1 + y_3y_1 + f^2 = 0 \end{cases} \tag{12}$$

Detailed proof and explanations for these properties can be found in [3].

Determine the rotation matrix: From obtained $\mathbf{X} = (x_1, y_1, f)^{\mathrm{T}}$, $\mathbf{Y} = (x_2, y_2, f)^{\mathrm{T}}$ and $\mathbf{Z} = (x_3, y_3, f)^{\mathrm{T}}$, rotation matrix can be easily acquired as:

$$\mathbf{R} = \begin{bmatrix} \mathbf{X}^T/\mathrm{norm}(\mathbf{X}) \\ \mathbf{Y}^T/\mathrm{norm}(\mathbf{Y}) \\ \mathbf{Z}^T/\mathrm{norm}(\mathbf{Z}) \end{bmatrix} \tag{13}$$

Determine the translation vector: Given a 3D point, denote its camera coordinates as $\mathbf{M}_{\mathrm{C}}$, and its world coordinates as $\mathbf{M}_{\mathrm{W}}$, they satisfy:

$$\mathbf{M}_{\mathrm{C}} = \mathbf{R}\mathbf{M}_{\mathrm{W}} + \mathbf{T} \tag{14}$$

For the translation vector $\mathbf{T} = (T_{\mathrm{X}}, T_{\mathrm{Y}}, T_{\mathrm{Z}})^{\mathrm{T}}$, T_{Z} is the camera height where the origin of the world coordinate system is on the horizontal plane. If the height of walking human H is given, T_{Z} can be determined by the cross ratio of the four points A, B, C, D:

$$T_Z = \frac{H}{1 - (BC, AD)} \tag{15}$$

where D is the vanishing point $\mathbf{v}_z$, B and C are the head and bottom positions in lambda-shaped frames, A is the intersection of the horizon line l_{Inf} (determined by the two ground vanishing points $\mathbf{v}_x$ and $\mathbf{v}_y$) and the line passing through B, C and D. If we assign a point in the image plane as the correspondence of the world coordinate system's origin, T_{X} and T_{Y} can be computed by solving the equation of perspective projection.

5.2. Minimal data and robust calibration

We need at least adjacent two left toes and two right toes to construct enough lines and detect all of the TOVPs, thus the minimal data for our calibration method are continuous three steps, which generate four continuous toe positions of the shoe prints on the ground. It is obvious that a minimal data with such short path lengths easily satisfies the assumption of walking approximate in a straight line, namely, a little change of walking direction is not so serious. Additionally, if provided more data, we can complete even more accurate and robust calibration against noise and outliers, by dividing the extracted toes into groups of four, namely, groups of minimal data. From each group and its corresponding head position, we can detect a set of TOVPs and then compute camera parameters. For the intrinsic parameter sets as $\{f_i, \mathbf{p}_i\}_{i=1,\ldots,\mathrm{C}}$, where f_i and $\mathbf{p}_i = (u_{0,i}, v_{0,i})^{\mathrm{T}}$ are the focal length and principle point determined by the i-th group, we can adopt RANSAC to eliminate unreasonable sets first and then determine the ultimate parameters by majority voting or least square methods.

5.3. Degenerate case

The degenerate case is that two lines determined by the left and right toes become almost coincided. In this case, $\mathbf{v}_z$ and $\mathbf{v}_x$ are still available, with $\mathbf{v}_y$ undetermined. Under assumption that the principle point $\mathbf{p} = (u_0, v_0)^{\mathrm{T}}$coincides with the center of the image plane, we can solve the third equation in (12) and obtain focal length f. Substituting f into the first two equations in (12), we can estimate the vanishing point $\mathbf{v}_y$. $\mathbf{R}$ and $\mathbf{T}$ can be determined with the similar procedure as described in Section 5.1.

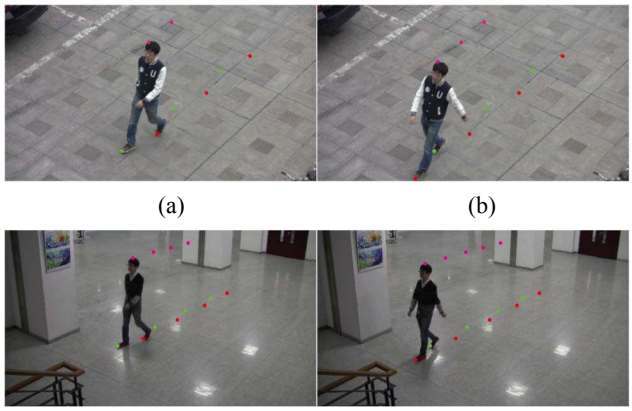
6. Experiments

In order to verify the proposed method, we use both sequences recorded by ourselves and downloaded from EPFL data set [2, 9]. We recorded sequences of pedestrians in various scenes from different heights and viewing angles. The sequences were recorded with Cannon LEGRIA HFS21 and have a resolution of 1920×1080. Some images taken in two different sequences Seq. #1 and Seq. #2 are shown in Figure 7. Seq. #1 was shot downwards from a balcony on the second floor of a building (see Figure 7ab), Seq. #2 was shot from the top corner of a corridor (see Figure 7cd), which are typical sequences of surveillance scenes shot at close or medium range. The detected left toes, right toes and head positions have been marked with different colors in these images.

For each sequence, we also record a sequence of a checkerboard in different positions and orientations. The calibration results obtained from the method of [26] are used as the ground truths. The comparisons between the ground truths and the intrinsic parameters estimated by the proposed method are presented in Table 1.

Seq. #		f	u_0	v_0
1	Ground truth [26]	5324.54	1011.57	521.29
	Ours	4433.11	1012.83	470.45
2	Ground truth [26]	2229.51	996.65	270.06
	Ours	2107.92	1165.38	329.04

Table 1. The comparisons of calibration results obtained from [26] and the proposed method



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

Figure 7. Some detected lambda-shaped frames in two of our sequences used in experiments. (a) and (b) are from Seq. #1, (c) and (d) are from Seq. #2. The detected left toes are marked as red, right toes are marked as green and head positions are marked as pink. (The personal information has been hidden to comply with the blind review policy.)

From Table 1, we may find that the calibration results of the proposed method are not so far from the ground truths. It is satisfying as expected and shows that camera calibration through the shoes of pedestrians is feasible. Most of the previous pedestrian based calibration methods are suitable for sequences took at medium or long ranges, but may not work or fail when dealing with sequences took at close ranges. In the two typical surveillance scenes as shown in Figure 7, where a pedestrian may well walk in a straight line for one pass in only a few steps and then goes out of sight, the proposed method may be the only solution so far.

We also use the multi-camera pedestrian sequences downloaded from EPFL data set [2, 9] to verify the proposed method. The available sequences have an original resolution of 720×576. As shown in Figure 8, Seq. #3 was shot outside a building on a terrace, which is also a common surveillance scene. The ground truths provided by EPFL data set [2, 9] are obtained from the Tsai model [24]. Comparisons between the ground truths obtained from [24] and the intrinsic parameters estimated by the proposed method are presented in Table 2.

Seq. #		f	u_0	v_0
3	Ground truth [24]	856.36	355.51	241.21
	Ours	740.98	407.43	223.48

Table 2. The comparisons of calibration results obtained from [24] and the proposed method

As presented in Table 2, the calibration results of the proposed method are basically consistent with the ground truths obtained from [24]. Experiments on sequences

(a) (b)

Figure 8. Two detected lambda-shaped frames in Seq. #3 from EPFL data set [2, 9]. The detected left toes are marked as red, right toes are marked as green and head positions are marked as pink. Note that the pedestrian's height and detected head positions in the lambda-shaped frames hardly change with the slight rotation of his head.

recorded by ourselves and downloaded from EPFL data set [2, 9] in different scenes and viewing angles prove the feasibility and accuracy of the proposed method, especially for sequences shot in common surveillance scenes at close or medium ranges.

7. Conclusions

In some typical surveillance scenes, such as subway stations, supermarkets, hospitals, hotels and etc., a majority of surveillance cameras are mounted in some top corners of the ceilings and at close or medium ranges from monitored persons. We find that in such situations, the shoes of pedestrians are very prominent in frames, especially the shoe prints, which are easily detectable, and may generate very regular pattern or, more precisely, repeatable pattern on the ground. As we known, repeatable pattern may be a very good choice for camera calibration, e.g., the commonly used checkerboard pattern. Therefore, this paper aims at employing such repeatable patterns for camera calibration. To the best of our knowledge, this is the first work showing that it is possible to calibrate camera through the images of shoes of pedestrians. By recognizing the "lambda-shaped" frames when two legs are maximally separated and left and right shoes both contact the ground, we can determine the image positions of the toes on the ground plane, as well as the corresponding head positions. Then we recover the TOVPs by utilizing the harmonic conjugate property to mine the metric information implicitly existing among the left and right shoe prints. After detecting all of the TOVPs, the intrinsic and extrinsic parameters of the camera can be determined. The minimal data for calibration in the proposed method are just continuous three steps, namely, four continuous shoe prints on the ground, thus easily ensuring the assumption of walking approximate in straight line. The degenerate case when left and right toes become almost collinear on the ground plane is also well discussed in this paper. Our ongoing work is to utilize traces of shoes of pedestrians to calibrate multiple cameras.

Acknowledgement

This work was supported in part by NNSFC Grant No. 61322309, NNSFC Grant No. 61273283, and NNSFC Grant No. 91120004.

References

[1] A. B. Albu, R. Bergevin and S. Quirion. Generic temporal segmentation of cyclic human motion. *Pattern Recognition*, 41(1):6-21, 2008. 2

[2] J. Berclaz, F. Fleuret, E. Turetken and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 33(9):1806-1819, 2011. 7, 8

[3] B. Caprile and V. Torre. Using vanishing points for camera calibration. *International Journal of Computer Vision,* 4(2): 127-139, 1990. 1, 6

[4] R. Cipolla, T. Drummond, and D. Robertson. Camera Calibration from Vanishing Points in Image of Architectural Scenes. In *British Machine Vision Conference*, 1999. 1

[5] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 781-796, 2000. 2

[6] P. Dollars, C. Wojek, B. Schiele and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743-761, 2012. 3

[7] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179-2195, 2009. 3

[8] O. Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993. 1, 3

[9] F. Fleuret, J. Berclaz, R. Lengagne and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267-282, 2008. 7, 8

[10] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 1,

[11] I. N. Junejo and H. Foroosh. Trajectory rectification and path modeling for video surveillance. In *International Conference on Computer Vision*, 2007. 1, 2

[12] D. Jung, Y. Yun and J. Choi. 3D pose estimation for foot motion tracking from image sequences. In *International Conference on Consumer Electronics*, 2011. 2, 4

[13] N. Krahnstoever and P. Mendonca. Bayesian autocalibration for surveillance. In *International Conference on Computer Vision*, 2005. 1, 2

[14] W. Kusakunniran, H. Li and J. Zhang. A direct method to self-calibrate a surveillance camera by observing a walking pedestrian. In *Digital Image Computing: Techniques and Applications*, 2009. 1, 2

[15] I. Laptev, S. J. Belongie, P. Perez and J. Wills. Periodic motion detection and segmentation via approximate sequence alignment. In *International Conference on Computer Vision*, 2005 2

[16] 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):1513–1518, 2006. 1, 2

[17] S. J. Maybank and O. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123-151, 1992. 1

[18] B. Micusik and T. Pajdla. Simultaneous surveillance camera calibration and foot-head homology estimation from human detections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2

[19] V. Paelke, C. Reimann and D. Stichling. Foot-based mobile interaction with games. In *International Conference on Advances in Computer Entertainment Technology*, 2004. 2

[20] Y. Ran, I. Weiss, Q. Zheng and L. S. Davis. Pedestrian detection via periodic motion analysis. *International Journal of Computer Vision*, 71(2):143-160, 2007. 2

[21] E. Ribnick and N. Papanikolopoulos. 3D reconstruction of periodic motion from a single view. *International Journal of Computer Vision*, 90(1):28-44, 2010. 2

[22] J. G. Semple and G. T. Kneebone. *Algebraic projective geometry*. Oxford University Press, 1998. 4, 5

[23] P. Sermanet, K. Kavukcuoglu, S. Chintala and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 3

[24] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323-344, 1987. 8

[25] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa and D. Geroimo. Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):797-809, 2014. 3

[26] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330-1334, 2000. 1, 7