Surveillance Camera Autocalibration based on Pedestrian Height Distributions

Jingchen Liu
jingchen@cse.psu.edu

Robert T. Collins
rcollins@cse.psu.edu

Yanxi Liu
yanxi@cse.psu.edu

Laboratory for Perception, Action and Cognition
Pennsylvania State University
State College, PA, USA
http://vision.cse.psu.edu

Abstract

We propose a new framework for automatic surveillance camera calibration by observing videos of pedestrians walking through the scene. Unlike existing methods that require accurate pedestrian detection and tracking, our method takes noisy foreground masks as input and automatically estimates the necessary intrinsic and extrinsic camera parameters using prior knowledge about the distribution of relative human heights. Our algorithm is computationally efficient enough for online parameter estimation. Experimental results on both synthetic and real data show the robustness of our method to camera pose and noisy foreground detections.

1 Introduction

The main goal of surveillance camera calibration is to compute a mapping between objects in the 3D scene and their projections in the 2D image plane [10]. This helps to infer object locations and allows for more accurate object detection and tracking. For example, sampling-based pedestrian detection [6] yields better performance when hypotheses are generated in 3D and then projected into one or more image views. Detection approaches based on sliding windows can also benefit from calibration, since the search over orientation and scale can be constrained to a small range, reducing false positives [14].

In this paper, we present an automated calibration method that enables smart sampling of object size and orientation given either a 2D location in the image or 3D location in the scene. The method works directly on noisy observations collected by the surveillance system, which is appealing because traditional photogrammetric calibration from surveyed points is time-consuming and expensive.

Under the assumption of a single, flat ground plane, most existing work for automatic surveillance calibration requires relatively clean pedestrian detection results to estimate a foot-head homology: [14] uses an expensive human detector and locates multiple control points on human contours to establish correspondence constraints; [15] assumes that humans are vertical in the image and uses a rectangular bounding box to localize the pedestrians; [13] requires accurate detection of leg-crossings for calibrating the camera. These methods have
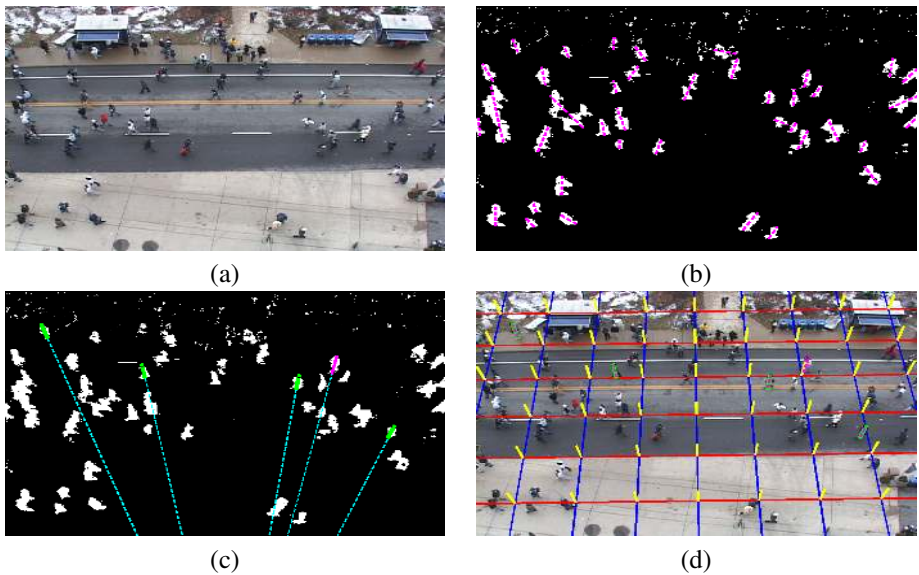
(a)

(b)

(c)

(d)

Figure 1: An example of surveillance camera calibration on Seq.1. (a) one video frame; (b) foreground masks after adaptive background subtraction, overlaid with major axis of each 2D blob; (c) after RANSAC vanishing point estimation and height distribution analysis, the RANSAC inliers from this frame are plotted in magenta, and inliers that also indicate a consistent 3D height are plotted in green; (d) final calibration results where each yellow line indicates predicted pedestrian orientation and height at a particular ground plane location.

limited practical use in more challenging real-world environments (Fig. 1) where field of view is large, image resolution is relatively low, and accurate detection of pedestrian shapes is almost impossible prior to calibration. Furthermore, to our best knowledge, all existing methods that use pedestrians to calibrate the camera require explicit correspondence of the same people over time. [7, 13, 14] assume that there is only a single person walking across the scene during the training stage, which rarely happens in crowded scenes. Others like [9, 15] require either tracking or manual input to specify the same person across different frames. Given the challenge in detecting people in scenes like Fig. 1, it is even harder to achieve robust tracking before calibration.

Our work addresses existing limitations in the literature with the following main novelties: (1) our algorithm is robust to noisy foreground detections in moderately crowded scenes; (2) we do not require correspondence of the same people across frames; (3) we look for a low-variance distribution of relative pedestrian heights, motivated by strong prior knowledge that 90% of human heights fall within a very small range of $\pm 7.6\%$ from the mean [16][1]. Our camera calibration method is based on the idea that objects (pedestrians) in the scene are all roughly the same height. Given a set of 2D observations, different hypothesized camera parameters will recover different explanations of the 3D scene; however, a *correct* camera model will recover the most realistic explanation, measured here by a low variation of relative pedestrian heights about an estimated mean.

---

[1]The statistics in [16] were generated from people of European descent; we assume similar results hold for other populations.

Using our approach, a rough camera calibration can be obtained quickly given noisy foreground blobs observed from crowded scenes. Assuming on average 2 inlier foreground blobs can be detected in each frame, and that blobs are extracted at 3fps, a total of 1800 "good" blobs can be collected within 5 minutes, which is empirically more than sufficient for accurate prediction of pedestrian orientation and height across the image.

2 Camera Geometry

Camera parameters to be estimated during the calibration stage include both intrinsic and extrinsic parameters. Assuming zero skew, known aspect ratio a,[2] and principal point at the image center $(0,0)$, focal length f is the only parameter to be estimated in the intrinsic parameter matrix:

$$K = \begin{bmatrix} f & 0 & 0 \\ 0 & f \cdot a & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

To specify the extrinsic parameters, we place the origin of the World Coordinate System (WCS) on the ground plane. Without loss of generality, assuming zero pan angle and no translation within the ground plane, the Camera Coordinate System (CCS) is determined by a translation h_c along the Z-axis (height of the camera above the ground) followed by a tilt angle θ around the X-axis and a roll angle ρ around the Z-axis. Combined with the intrinsic matrix, the full camera projection matrix is given by

$$P = K \begin{bmatrix} cos(\rho) & -sin(\rho) & 0 \\ sin(\rho) & cos(\rho) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & -sin(\theta) \\ 0 & sin(\theta) & cos(\theta) \end{bmatrix} \left[\begin{array}{ccc|c} 1 & & & 0 \\ & 1 & & 0 \\ & & 1 & -h_c \end{array} \right] \tag{2}$$

For a downward-tilted surveillance camera, θ has a value between $\pi/2$ (optic axis parallel to the ground plane) and π (facing straight down).

Let the vertical vanishing point be $v_0 = (v_x, v_y, 1)^T$.[3] Any vanishing point $(x, y, 1)$ on the horizon line associated with the ground plane (and every other horizontal plane in the scene) is orthogonal with v_0, which gives the constraint [11]

$$v_0^T \omega \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0, \tag{3}$$

where ω is the Image of Absolute Conic (IAC), defined as $\omega = K^{-T}K^{-1}$ [11], which imposes a metric on the uncalibrated space. Substituting Eqn. 1 into Eqn. 3, we get

$$v_x x + \frac{v_y}{a^2} y + f^2 = 0. \tag{4}$$

Eqn. 4 shows that the horizon line is determined by the vertical vanishing point together with the focal length f. It is also straightforward to recover the camera rotation parameters from

---

[2]For standard video formats, we set aspect ratio $a = 0.91$ for NTSC (720×480) and $a = 1.09$ for PAL (720×576); for unknown cases we assume $a = 1$.

[3]We are assuming a finite vanishing point in the image to simplify the presentation. The equations can be modified to handle vanishing points at infinity.

v_x, v_y and f as follows

$$\rho = \text{atan}(-av_x/v_y) \tag{5}$$

$$\theta = \text{atan2}(\sqrt{a^2 v_x^2 + v_y^2}, -af). \tag{6}$$

3 Vanishing Point Estimation via Foreground Blobs

Although it is often preferable to jointly optimize correlated parameters in order to maximally use all the available information, we argue that for this task it is more efficient to first estimate the vertical vanishing point independently from the other camera parameters. This is because the vanishing point can be computed using only foreground blob candidates in the image. We can easily collect many blobs in real-time from a short video sequence, especially in crowded scenes (Fig. 2). It also can be shown empirically that the estimation accuracy keeps improving as more blobs are incorporated into the robust random sample consensus (RANSAC) framework [5].
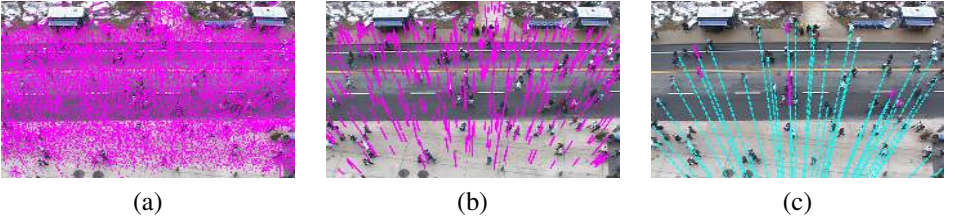

| (a) | (b) | (c) |

Figure 2: Vertical vanishing point estimation. (a) Major axis orientation of blobs extracted from a short video sequence. (b) Inlier axes found by RANSAC. (c) Lines connecting blob centroids to the computed vertical vanishing point overlaid on a sample frame.

The vanishing point estimation process is illustrated in Fig. 2. Adaptive background subtraction and connected component analysis is first performed to detect foreground blobs. The centroid, major and minor axis of each binary blob is determined from its first and second moments, which is equivalent to approximating each blob shape by an ellipse. The major axes of some ellipses represent vertical orientation of individuals who are found as a single blob; however, many others are outliers representing the orientation of multiperson blobs, fragmented blobs, or blobs whose second moments are corrupted away from vertical by arms and legs extending out from the person. To robustly find the set of ellipse axes that converge to the vertical vanishing point, we use RANSAC to find the intersection point voted for by the most axes. The line intersection computations are carried out in homogeneous coordinates so that the procedure works well even vanishing points at infinity.

Following RANSAC selection, we use the method of [2, 3] to fit a least squares vanishing point estimate to the inlier blobs. The major axes of all K inlier ellipses are organized into a matrix with normalized homogeneous coordinate representation:

$$A = \begin{bmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_K & y_K & z_K \end{bmatrix}. \tag{7}$$

The vanishing point direction should be a unit vector perpendicular to all of these vectors. Therefore, we minimize the sum of squared cosine angles from the vanishing point $v_0 = [v_x \quad v_y \quad v_z]^T$ to all inlier lines:

$$\arg\min_{v_0} \|A v_0\|^2 = \arg\min_{v_0} v_0^T (A^T A) v_0. \tag{8}$$

The *Mean Squared Error (MSE)* solution of v_0 is given by the eigenvector of $A^T A$ that corresponds to the smallest eigenvalue.

Given the vertical vanishing point, the major axis of each inlier blob $i = 1, \ldots N$ is reprojected onto the line connecting the vanishing point and the blob center. These projected line segments are represented as a pair of feet and head pixel locations $(p_f, \ p_h)$ for subsequent height analysis.

The above method for vertical vanishing point extraction is more efficient with large numbers of observations compared to other robust statistics [12, 13]. However, there may still exist outliers among the "RANSAC inliers" – although the orientations of inlier axes are now consistent, some blobs may still have inconsistent heights. This will be dealt with in the following stage.

4 Relative Height Computation

A key part of our approach is to analyze the distribution of relative pedestrian heights from inlier foot-head pairs found by the RANSAC procedure. We define the *relative height* of a pedestrian h_i as their actual 3D height h_i^{3D} divided by the camera height h_c. This relative height can be computed as an invariant using the cross ratio [4, 13] of distances between four points

$$h_i = \frac{h_i^{3D}}{h_c} = 1 - \frac{d(p_h, v_l) \cdot d(p_f, v_0)}{d(p_f, v_l) \cdot d(p_h, v_0)}. \tag{9}$$

where p_f and p_h are feet and head position of the pedestrian, v_0 is the vertical vanishing point, and v_l is the intersection of a line passing through these points and the horizon line, as illustrated in Fig. 3. It is worth mentioning that the distance measurement $d(a,b)$ in Eqn. 9 is directional. The feet position p_f should always be below the horizon in the image, however the head position p_h could be above the horizon, which would result in a negative distance ratio $d(p_h, v_l)/d(p_f, v_l)$, and hence $h_i > 0$, indicating that the height of the pedestrian is greater than the camera height.

Furthermore, Eqn. 9 is robust for vertical vanishing points near infinity. A camera with tilt angle θ close to $\pi/2$ projects vertical lines in 3D into parallel image lines, so that the vertical vanishing point is far away from the image center with large estimation variance. However, in such cases the ratio of $d(p_f, v_0)/d(p_h, v_0)$ is approximately 1, so the large estimation variance in v_0 does not corrupt the relative height estimate.

Given the estimated vertical vanishing point v_0 and pedestrian head and feet positions that were computed in Sec. 3, all we need is the horizon line for completing the computation of relative height h_i. Eqn. 4 showed that we can compute the horizon line from v_0 if we know the camera focal length f. In the next section, we compute the distribution of relative heights implied by different candidate values of f, and use prior knowledge about likely pedestrian height distributions to evaluate each hypothesized focal length.
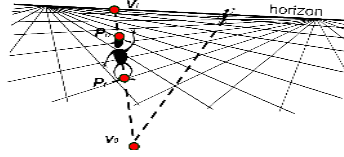
Figure 3: Estimating the relative height of a pedestrian using cross ratio invariance among the four points marked in red.

5 Calibration by Hypothesis Testing

Given the analysis above, the camera calibration problem now becomes looking for an appropriate focal length f that, together with the previously estimated v_0, will generate a relative height distribution that best matches pedestrian heights in the real-world. Therefore, we need to define a likelihood function to evaluate the similarity of a computed 1D relative height distribution with respect to the known human height distribution in the real world.

If we knew the absolute camera height h_c, we could solve for the 3D heights h_i^{3D} of people and compare the distribution of those to a normal distribution of 3D pedestrian heights, measured for European populations as having a mean of roughly 172 cm and a standard deviation of 7cm [16]. Even if we knew the absolute height of just one reference point in the scene, we could estimate the camera height, and from there infer 3D pedestrian heights. However, in this work we base an analysis on relative heights, bypassing the need for scene measurements, and allowing generalization to non-European populations.

A striking feature of the human height distribution is how closely it clusters about the mean, with 90% of human heights having a relative difference of less than 7.6% from the population average [16], or

$$|h_i^{3D} - E[h_i^{3D}]| / E[h_i^{3D}] \le .076 \ . \tag{10}$$

Note that Eqn. 10 also holds true for the relative heights $h_i = h_i^{3D}/h_c$, since the unknown but constant camera height h_c cancels from the equation.

We therefore define a distance metric

$$r(h_i, \mu) = \max\{\tau - \frac{|h_i - \mu|}{\mu}, 0\} \tag{11}$$

as a thresholded relative distance from a given h_i to the boundary $(1 \pm \tau)\mu$, where τ is the relative distance threshold (we set $\tau = 0.1$ for all experiments) and μ is an unknown population mean of relative heights. We then form a joint log-likelihood function over μ and f as

$$\mathcal{L}(O|f, \mu, v_0) \sim \sum_i \frac{1}{\mu} r(h_i, \mu)^2 \tag{12}$$

to evaluate the quality of a hypothesized focal length with respect to how well the relative heights it calculates cluster about population mean μ. For a given focal length, the relative pedestrian heights $\{h_i\} = \text{func}(O, f, v_0)$ are calculated based on image observation O (extracted foreground blobs) and the camera calibration parameters, using Eqn. 4 and 9 as described in previous sections.

To find the optimal parameters f^* and μ^* that yield the highest model likelihood, we enumerate values of f with a *field of view (FoV)* resolution of $0.5°$, and for each of those, we search for the optimal population average μ within a range of $median(h_i) \pm std(h_i)$.

The $\max\{\cdot, 0\}$ operation in Eqn. 11 indicates that only height candidates falling within a region of interest will contribute to the likelihood evaluation. This is analogous to the robust statistics of a truncated quadratic [1] in that it gives constant punishment to samples outside the region of interest, and therefore is robust to noisy sample data with a large proportion of outliers. More importantly, our prior knowledge about human height distribution indicates a meaningful threshold ($\tau = 10\%$), which makes this metric more favorable in our framework in comparison with other popular robust statistics such as least trimmed squares [8] that only consider a fixed number of nearest samples. The term $1/\mu$ in equation 12 gives further punishment to larger μ that result in larger regions of interest.

In addition to being robust to obvious outliers, we also want to deal with noise on the inliers due to inaccurate blob fitting or blob distortion. We assume zero-mean Gaussian noise on the feet-head locations with fixed variance σ^2, and by bootstrap sampling with these Gaussian distributions we can obtain multiple relative heights for each blob. The log-likelihood function corresponding to Eqn. 12 then can be rewritten in a more generalized form incorporating height uncertainty:

$$\mathcal{L}(O|f,\mu,v_0) = \sum_i \int \frac{1}{\mu} r(h_i,\mu)^2 p(h_i) \mathrm{d}h_i \tag{13}$$

$$\sim \frac{1}{\mu M} \sum_i \sum_m r(h_i^{(m)},\mu)^2, \tag{14}$$

where Monte Carlo integration is applied to obtain Eqn. 14 and M is the number of samples for each h_i.

6 Experimental Results

We first performed a test on two synthesized datasets with different viewing angles for validation, as shown in Fig. 4 (a1) and (b1), where white line segments are synthesized noisy data and the yellow lines are the estimated average height predicted for different ground plane locations. The performance is quantitatively evaluated by the mean squared relative error of feet-head projection

$$err = \int_{p_f} \left(\frac{d(p_h, p_h^*)}{d(p_f, p_h^*)} \right)^2 \tag{15}$$

where p_h and p_h^* are estimated and groundtruth head positions, respectively, given a selected foot position p_f in the image. All distances are measured in image coordinates, and the squared error is accumulated on p_f through the entire region of interest. This evaluation design reflects the main interest of many surveillance applications, which require correct estimation of pedestrian size and orientation in 2D for any specified pixel location.

We compared four different algorithms. **Alg.1** is our proposed approach using only one pedestrian height h_i calculated from Eqn. 9. **Alg.2** is the bootstrapped version that incorporates the height uncertainty of each h_i according to Eqn. 14 using $M = 30$ samples. **Alg.3** is our implementation of the approach in [13], which estimates the horizon using pairs of pedestrians incorporated into a robust estimation framework. **Alg.4** implements the method in [9],
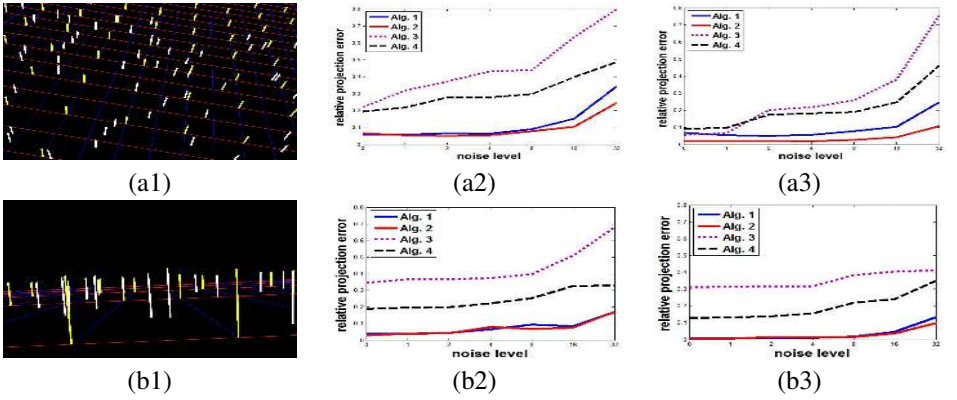
Figure 4: Synthesized test data: (a1) a highly elevated camera with narrow viewing angle; (b1) a low viewpoint camera with wide viewing angle; (a2),(b2) performance evaluation on both datasets with 20 inlier blobs; (a3),(b3) performance evaluation on both datasets with 640 inlier blobs.

which computes the MAP estimate of a feet-head homology in a Bayesian framework, where we use the independently estimated vertical vanishing point to reduce the sample space for the feet-head homology, and sum the uncertainty on the locations of pedestrians over the 3D ground plane.

We synthesize a uniform height distribution within $(90 \sim 110\%)\mu$ and also generate the same number of outliers as inliers, with random positions, orientations and heights. As we increase the variance of Gaussian noise on inlier feet-head positions, it can be seen from Fig.4 that our proposed approaches (Alg.1 and Alg.2) consistently outperform the others. The performance gain is more prominent for higher noise levels. The estimated vertical vanishing point (shared by all testing algorithms) at different noise levels is listed in Table 1. The parameter estimation errors for *FoV* and relative camera height ($h_c = 1/\mu$) are averaged over all noise levels and shown in Table 2 together with the groundtruth values.

Table 1: Vanishing point estimation of synthesized datasets.

	noise level	0	2	8	32
set 1:	20 inliers	$(-167, 826)$	$(-169, 831)$	$(-182, 850)$	$(-329, 1278)$
$(-168, 828)$	640 inliers	$(-168, 828)$	$(-168, 828)$	$(-163, 804)$	$(-198, 1020)$
set 2:	20 inliers	$(357, 7135)$	$(349, 7292)$	$(236, 4277)$	$(193, 1250)$
$(356, 7106)$	640 inliers	$(357, 7122)$	$(354, 6979)$	$(387, 9771)$	$(298, 3039)$

Table 2: Quantitative comparison of 4 algorithms on synthesized datasets.

		GT	Alg.1	Alg.2	Alg.3	Alg.4
set 1:	FoV(°)	51	50.6 ± 4.5	50.5 ± 3.7	65.2 ± 14.5	46.6 ± 3.2
	h_c^{-1}	.06	$.061 \pm .010$	$.061 \pm .008$	$.071 \pm .020$	$.050 \pm .006$
set 2:	FoV(°)	120	118.2 ± 5.6	117.7 ± 5.6	124.8 ± 11.4	114.0 ± 5.0
	h_c^{-1}	1.1	1.08 ± 0.11	1.07 ± 0.10	1.16 ± 0.37	1.23 ± 0.33

(a) Seq. 1   (b) Seq. 2   (c) Seq. 3
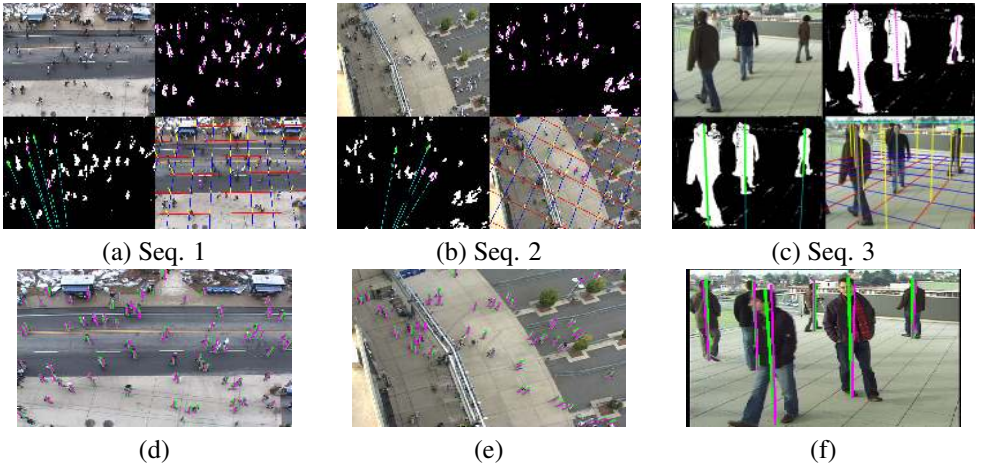
(d)   (e)   (f)

Figure 5: Calibration results on real sequences with inlier blobs on sample frames (a,b,c); Estimated feet-head projection (magenta) versus groundtruth labels (green) (d,e,f)

We also test our approach (Alg.2) on three real sequences. Two sequences were captured outside a stadium during a sporting event using a Sony DCR VX2000 digital video camcorder. The third sequence is from a publicly available dataset showing people walking on a terrace [4]. As shown in Fig. 5, our method correctly estimates pedestrian orientation and height over the ground plane for different camera settings. The image projection error for the foot to head mapping after calibration is evaluated from hand labeled data, as shown in Fig. 5 and Table 3.

Table 3: Root Mean Square Error of Feet-Head Projection on real Sequences

	Seq.1	Seq.2	Seq.3
absolute RMSE (pixel)	3.8	4.2	8.9
relative RMSE (Eqn. 15)	0.15	0.18	0.05

We also conduct 3D/2D distance estimation based on available groundtruth as illustrated in Fig.6. In (a), we estimate a total of 9 real-world distances (marked as black arrows) among 3 parallel lines, and compare these to on-site measurements. Our estimation of the 3D distance is a proportion of average human height, and the minimum mean squared relative error: $MSE = average\{(d^{(3D)}/d^{(GT)} - 1)^2\} = (3.5\%)^2$ is reached when we set average pedestrian height to be 6.58 feet. The standard deviation of the estimated 3D orientation of the 3 parallel lines is 0.22°. Given the uncertainty of actual human heights, height changes during walking and lens distortion, these errors are within a reasonable range. In Fig.6(b), the groundtruth horizon is located at 6% of the image height (17.3 pixels) from the top of the frame, while the average vertical distance on our estimated horizon (plotted in cyan) is 5.59%(16.10 pixels) from the top. The skew is 0.24°. Also, as a sanity check, the estimated horizon intersects the other camera of the same height (circled in red) mounted at the far end of the terrace.
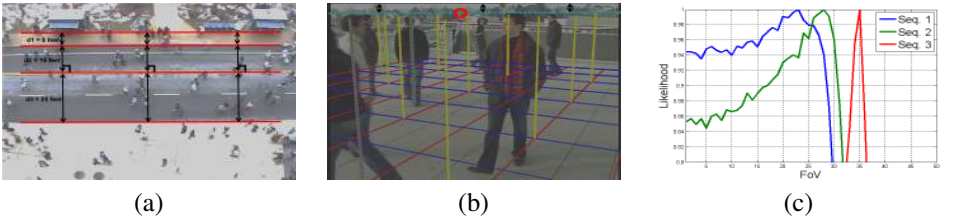
---

[4]http://cvlab.epfl.ch/data/pom/

Figure 6: Qualitative and quantitative evaluation on real sequence. (a) Groundtruth distances in scene 1; (b) Estimated horizon in scene 3; (c) Likelihood curves of FoV angle

The likelihood curve for different *FoV* angles (normalized to $[0, 1]$) is also plotted in Fig. 6(c) for each sequence, indicating the confidence interval for *FoV* estimation.

7   Conclusion

We propose a novel approach for unsupervised camera calibration based on noisy foreground detections in moderately crowded scenes. No measurements in the scene are required, nor is there a need to track or label correspondences of the same person between frames. Instead, we optimize the camera parameters so that the predicted relative human height distribution resembles our prior knowledge about how pedestrian heights are distributed in the real world. Our approach is robust against noise in foreground subtraction, and can generate reliable pedestrian orientation and height predictions after only a short observation period. Preliminary experiments show promising results under a variety of different camera views.

References

[1] M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1): 75–104, 1996.

[2] R.T. Collins. *Model Acquisition Using Stochastic Projective Geometry*. PhD thesis, University of Massachusetts, 1993.

[3] R.T. Collins and R.Weiss. Vanishing point calculation as a statistical inference on the unit sphere. In *Proc. ICCV*, pages 400–403, 1990.

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *Int. J. Computer Vision*, 40(2):123–148, 2000.

[5] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[6] Weina Ge and Robert T. Collins. Crowd detection with a multiview sampler. In *Proc. ECCV*, pages 1–14, 2010.

[7] Imran N. Junejo and Hassan Foroosh. Trajectory rectification and path modeling for video surveillance. In *Proc. ICCV*, pages 1–7, 2007.

[8] Kang-Mo Jung. Least trimmed squares estimator in the errors-in-variables model. *Applied Statistics*, 34(3):331–338, 2007.

[9] Nils Krahnstoever and Paulo R.S. Mendonca. Bayesian autocalibration for surveillance. In *Proc. ICCV*, pages 1858–1865, 2005.

[10] D. Liebowitz. *Camera Calibration and Reconstruction of Geometry from Images*. PhD thesis, University of Oxford, 2001.

[11] David Liebowitz and Andrew Zisserman. Combining scene and auto-calibration constraints. In *Proc. EuroGraphics*, pages 293–300, 1999.

[12] David Liebowitz, Antonio Criminisi, and Andrew Zisserman. Creating architectural models from images. In *Proc. EuroGraphics*, pages 39–50, 1999.

[13] Fengjun Lv, Tao Zhao, and Ramakant Nevatia. Camera calibration from video of a walking human. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(9):1513–1518, 2006.

[14] Branislav Micusik and Tomas Pajdla. Simultaneous surveillance camera calibration and foot-head homology estimation from human detections. In *Proc. CVPR*, pages 1–8, 2010.

[15] Diego Rother and Kedar A. Patwardhan. What can casual walkers tell us about the 3d scene. In *Proc. ICCV*, pages 1–8, 2007.

[16] Peter M. Visscher. Sizing up human height variation. *Nature Genetics*, 40(5):489–490, 2006.