

Simultaneous surveillance camera calibration and foot-head homology estimation from human detections*

Branislav Micusik
Safety and Security Department
AIT Austrian Institute of Technology

Tomas Pajdla
Center for Machine Perception
Czech Technical University in Prague

Abstract

We propose a novel method for automatic camera calibration and foot-head homology estimation by observing persons standing at several positions in the camera field of view. We demonstrate that human body can be considered as a calibration target thus avoiding special calibration objects or manually established fiducial points. First, by assuming roughly parallel human poses we derive a new constraint which allows to formulate the calibration of internal and external camera parameters as a Quadratic Eigenvalue Problem. Secondly, we couple the calibration with an improved effective integral contour based human detector and use 3D projected models to capture a large variety of person and camera mutual positions. The resulting camera auto-calibration method is very robust and efficient, and thus well suited for surveillance applications where the camera calibration process cannot use special calibration targets and must be simple.

1. Introduction

Known camera calibration in visual surveillance systems is an important cue for better human detection and further tracking. The known camera internal calibration and its position w.r.t. a ground plane allows to compute a foot-head homology, a mapping transforming each image ground point to the expected human head location. That allows to evaluate at each location only a fraction of possible detector rotations and scales and decreases thus significantly false detections, shown in Fig. 1.

In this paper, we aim at automatic simultaneous camera calibration and the foot-head homology estimation just by observing a person standing at various locations. We demonstrate that human body can be used as a calibration target with appropriate human detection method. We avoid

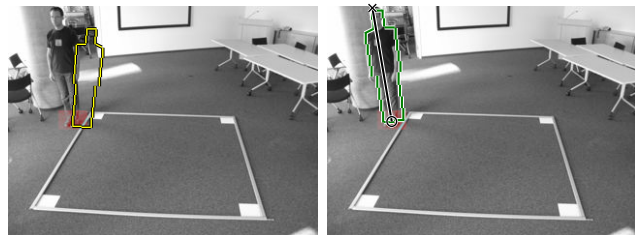


Figure 1. Camera calibration from human detections. Left: An initial (possibly wrong) human detection without knowing a camera-scene geometry. Right: A refined detection after utilizing an automatically estimated foot-head homology. The line represents the homology which maps the clicked foot point ‘o’ to the expected head location ‘x’ and allows to adjust human detector scale and rotation automatically.

the need of special calibration objects, checkerboards, or manually clicked points which require expert intervention and therefore not being of interest for real surveillance systems.

The contribution of this paper is two-fold. First, we bring a novel formulation of single camera calibration via the Quadratic Eigenvalue Problem (QEP). Second, we link a general contour based human detector with a camera calibration problem by incorporating synthetic human projections. We solve by QEP simultaneously for focal length, rotation and translation of a camera which occur in monomial relations. QEP is very suitable for such a difficult non-linear algebraic problem as an alternative to using Groebner basis, and therefore it is widely beneficial to bring any problem to a QEP formulation, *e.g.* as done for epipolar geometry estimation with unknown radial distortion [6, 17, 18] or focal length [4]. We recognize another problem yielding QEP and moreover, we formulate the problem as the rectangular QEP and show its significant improvement over the typically used squared QEP and standard vanishing point based techniques. The core idea is that we recognize that parallel shifted homographies with properly chosen coordinate system and parametrization can be written in such a form that it allows to factorize the homography and monomial

*This research received funding from Wiener Wissenschafts-, Forschungs- und Technologiefonds - WWTF, Project No ICT08-030, and from the projects FP6-IST-027787 DIRAC and MSM6840770038.

relations in an elegant way to solve for the unknowns in a closed form for minimal case and in the rectangular QEP for the overconstrained case.

Previous work Approaches to estimating the homology from human detections in single view [16, 12, 9, 10] often rely on estimating the vertical vanishing point and the horizon. The estimation of the vanishing points is usually the bottleneck of the approaches since it is extremely sensitive to noise [5, 11, 15]. Even very small inaccuracy in the vanishing point can cause huge inaccuracy in the focal length estimate, easily 100%. Therefore, these approaches are only of a limited use in practice.

In [9], in effort to avoid vanishing point computation, the epipolar geometry between two human detections is utilized. After deeper analysis, one can prove that estimating the proposed vertical fundamental matrix is algebraically equivalent to computing the intersection of many lines (in the least square sense) created by connecting the foot and head points. Therefore, this formulation brings no difference to computing the vertical vanishing point by the standard way [15] and therefore suffers from the same sensitivity to foot and head point detections.

We search for a method which could deliver the foot-head homology more robustly just from human observations and moreover, it would simultaneously deliver the camera focal length, translation and rotation w.r.t. the ground plane. Compared to [16], we estimate the parameters simultaneously and not step by step which obviously accumulates errors. Thanks to the contour based human detector, we utilize more points on the person than just feet and head usually used when bounding box detectors are employed. Human silhouettes have been considered in [13] for estimating the foot-head homology, however, the focal length and external camera parameters have not been explicitly estimated as by our approach.

2. Concept

In the following, we consider a single camera observing a scene with a common ground plane. The goal is to estimate a camera focal length, extrinsic camera parameters, and a foot-head homology by observing a person standing at several roughly parallel positions, see Fig. 2.

First, a person is detected in each frame by sliding a contour based detector at all possible scales and rotations as explained in Sec. 2.1. Then, the best N candidates are considered for each image and all the detections are used in a RANSAC-based estimation process to get the desired focal length, the extrinsics, and homology, explained in Sec. 2.2. The homology allows to compute the expected head location for each foot point and thus to select only feasible models to be evaluated at each point. It allows to re-run contour

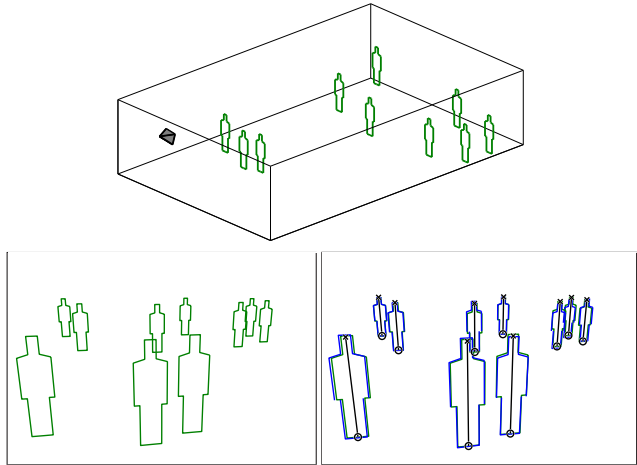


Figure 2. Concept. Top: A single camera observing a person at different locations. Bottom left: Person projections. Bottom right: Person detections with possible inaccuracy, and the resulting foot-head homology which maps each foot point marked by ‘o’ to a head position, marked by ‘x’.

detector again but with a finer resolution of detector scale and rotation. This two-way iteration strategy allows to decrease computational burden which would be necessary for evaluating all the detailed models from the beginning.

2.1. Human detection

A generative approach to recognizing objects has turned out to be a very promising direction [14, 19]. It has been shown [19] that the silhouettes alone without appearance features are very discriminative. The basic idea is to generate many silhouette images of a synthetic 3D model of objects of interest by changing a virtual camera viewpoint. Then, in the inference stage, a query image is traversed and each hypothesized image location is compared against all the stored object silhouette projections, or just the most representative ones.

In the context of surveillance, a contour based approach for human detection has been suggested in [2]. They have proposed an efficient generative approach, the contour templates, based on the integral image concept. Human body shape is represented by a piece-wise linear edge model. Its line orientations are quantized into eight predefined orientations to efficiently perform sums along the line segments in the gradient image to evaluate for cost of the model at a particular image location. Only up-right human positions are considered, therefore, the foot-head relation is a function of y coordinate only and is fixed manually.

In this paper, we work with a priori unknown foot-head relation and therefore at the beginning of the human detection, when there is no information about scene geometry, many hypothesized models capturing various person poses must be evaluated. To efficiently validate all the poses we

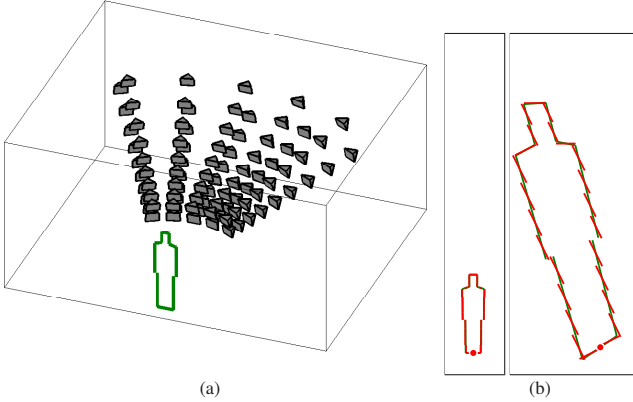


Figure 3. Generating the contour models. (a) A virtual camera is placed at different locations. Each camera is also rotated around its optical axis by a yaw angle which is not shown in the figure. (b) Two projections, the smallest and the largest, are shown. The vertical dimension of the frame corresponds to the image vertical dimension of 480pxls. The rightmost model shows splitting and snapping projected line segments into the discretized orientations.

adopt the method [2] but we suggest to generate synthetic projections of human contours from various camera viewpoints, see Fig. 3a, giving us an ability to model many more human poses than in [2].

The contour projections capture large scale and camera rotation variations as shown in Fig. 3b. We generate roughly a thousand model projections, which are related to each other by a projective transformation. Note that a simple isotropic scaling of the models or detector, as suggested by many methods to resolve scale ambiguity, does not correspond to correct transformations. Moreover, since we assume only one contour model, the dimensions of its parts do not necessarily have to fit to all humans. It means that, *e.g.* a thinner person with even different (head-hand-leg) proportions can still be well approximated by a model generated by a virtual camera at a different locus than the real camera observing that person.

To utilize integral images and to get sufficient accuracy for more general human orientations than in [2] while using only a limited number of eight discretized edge orientations, long edges of contour models are split into shorter ones as shown in Fig. 3b. At each image location, the model which gets the highest score is considered. Parabola-based non-maxima suppression is finally applied and the highest N maxima are taken as the best N detections per image.

2.2. Foot-head homology estimation

2.2.1 Shifted Homographies

Let us choose a world coordinate system at one human foot location and orientation of the coordinate axes as shown Fig. 4. The projection of a 3D point \mathbf{X} into a camera

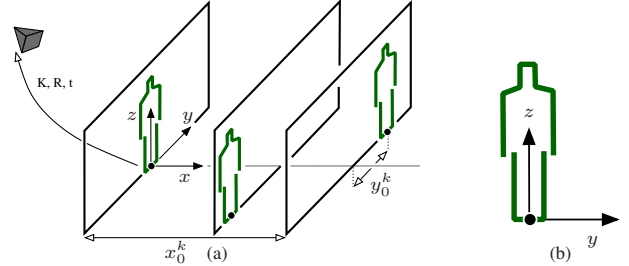


Figure 4. (a) A considered scene with a camera and a parallel standing person. (b) Coordinate system of the human edge model.

point \mathbf{u} reads as

$$\mathbf{u} = \alpha \mathbf{K} [\mathbf{R} \ \mathbf{t}] \mathbf{X}, \quad (1)$$

where α is a scale, and \mathbf{K} , \mathbf{R} , \mathbf{t} stand for camera calibration matrix, rotation and translation w.r.t. the world coordinate system [8]. We assume that all human positions are parallel and $x = 0$, since fixing the coordinate system into the first detection. For each human pose thus holds

$$\begin{aligned} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} &= \alpha \mathbf{K} [\mathbf{R} \ \mathbf{t}] \begin{pmatrix} x + x_0 \\ y + y_0 \\ z \\ 1 \end{pmatrix} = \mathbf{H} \begin{pmatrix} y \\ z \\ 1 \end{pmatrix} = \mathbf{H} \mathbf{x} \\ &= \alpha \mathbf{K} [\mathbf{r}_2 \ \mathbf{r}_3 \ x_0 \mathbf{r}_1 + y_0 \mathbf{r}_2 + \mathbf{t}] \mathbf{x} \end{aligned} \quad (2)$$

where \mathbf{r}_i is the i -th column of the rotation matrix \mathbf{R} and \mathbf{H} is a 3×3 homography matrix, u and v are detections in the image. It is evident that if we compute the homography between the human edge model and all their projections, all the homographies will be the same up to the third column. This allows us to estimate the homography simultaneously from more, let's say K , detections and design the following linear system

$$\mathbf{M} \mathbf{h} = \mathbf{M} (\mathbf{h}_1^\top \ \mathbf{h}_2^\top \ \mathbf{h}_3^{1^\top} \ \dots \ \mathbf{h}_3^{K^\top})^\top = \mathbf{0}, \quad (3)$$

where \mathbf{h}_i is the i -th column of the homography, *i.e.* $\mathbf{H}^k = [\mathbf{h}_1 \ \mathbf{h}_2 \ \mathbf{h}_3^k]$ between the k th detection and the human edge model in Fig. 4b. The matrix \mathbf{M} is composed of detected image model points and the corresponding human model points. There are $6 + 3K$ unknowns in Eq. (3). Each point correspondence adds 2 equations and for each detection, we must provide at least one point correspondence to give constraints on \mathbf{h}_3^k which is unique for each detection. Altogether there must hold $\text{rank}(\mathbf{M}) = (6 + 3K) - 1$. That means, for one detection 4 correspondences are required, for two detections 6, for three 7 and so on. Including more detections and more correspondences leads to an overconstrained problem solvable in the least square sense by SVD. We use all the corner points of the model and two human detections, *i.e.* two images. Hartley's point normalization [8] is used before filling the matrix \mathbf{M} to improve the numerical stability.

2.2.2 Focal length, Rotation, Translation

From Eq. (2) we know that the last column of the homography for each human pose factorizes to

$$\alpha K(x_0^k \mathbf{r}_1 + y_0^k \mathbf{r}_2 + \mathbf{t}) = \mathbf{h}_3^k, \quad (4)$$

with the unknown calibration matrix K , \mathbf{t} , x_0^k , y_0^k and scale α . Let assume square pixels and known principal point of the camera. Next, before solving Eq. (3), we express all image detections in the image coordinate system with its origin at the principal point. Thus, the calibration matrix is $K = \text{diag}(f, f, 1)$, where f is the focal length. We can write

$$\alpha x_0^k K(\mathbf{r}_2 \times \mathbf{r}_3) + \alpha y_0^k K \mathbf{r}_2 + \alpha K \mathbf{t} - \mathbf{h}_3^k = \mathbf{0}. \quad (5)$$

Recall that we have estimated \mathbf{h}_1 and \mathbf{h}_2 and we know from Eq. (2) and Eq. (3) that $\mathbf{h}_1 = \alpha K \mathbf{r}_2$ and $\mathbf{h}_2 = \alpha K \mathbf{r}_3$. Then the first summand in Eq. (5) becomes

$$\frac{x_0^k}{\alpha} \det(K^{-1}) K K^\top (\mathbf{h}_1 \times \mathbf{h}_2) = \frac{1}{f^2} \frac{x_0^k}{\alpha} K K^\top (\mathbf{h}_1 \times \mathbf{h}_2).$$

Substituting that back into Eq. (5) and multiplying by f^2 we get

$$\frac{x_0^k}{\alpha} K K^\top (\mathbf{h}_1 \times \mathbf{h}_2) + y_0^k \mathbf{h}_1 f^2 + \alpha K \mathbf{t} f^2 - \mathbf{h}_3^k f^2 = \mathbf{0} \quad (6)$$

Introducing $\tilde{x}_0^k = \frac{x_0^k}{\alpha}$, $\tilde{\mathbf{t}} = \alpha \mathbf{t} = (\tilde{t}_x \ \tilde{t}_y \ \tilde{t}_z)^\top$, and $\mathbf{e} = (\mathbf{h}_1 \times \mathbf{h}_2) = (e_x \ e_y \ e_z)^\top$ we get for each pose the following set of equations

$$\begin{aligned} \tilde{x}_0^k e_x + y_0^k h_{1x} + f \tilde{t}_x - h_{3x}^k &= 0 \\ \tilde{x}_0^k e_y + y_0^k h_{1y} + f \tilde{t}_y - h_{3y}^k &= 0 \\ \tilde{x}_0^k e_z + f^2 y_0^k h_{1z} + f^2 \tilde{t}_z - f^2 h_{3z}^k &= 0 \end{aligned} \quad (7)$$

with six unknowns f , $\tilde{\mathbf{t}}$, \tilde{x}_0^k , y_0^k . Each additional human pose adds two new unknowns, \tilde{x}_0^k and y_0^k , but provides three more equations.

Minimal Solution. To get a solution we need two detections, *i.e.* we get six equations with six unknowns. Recall that we set the world coordinate system to be in one of the detections, therefore $\tilde{x}_0^1 = y_0^1 = 0$, reducing the number of unknowns by two. To simplify the notation, let us denote symbols with index $k = 1$, *i.e.* denoting the first detection, as primed and symbols with $k = 2$, *i.e.* the second detection, as double primed.

From the first three equation we have then $f \tilde{t}_x = h'_{3x}$ and $f \tilde{t}_y = h'_{3y}$ and after elimination process of Eq. (7) we

get finally

$$\begin{aligned} \tilde{x}_0'' &= \frac{h_{1y} h'_{3x} - h_{1x} h'_{3y} + h_{1x} h''_{3y} - h_{1y} h''_{3x}}{e_y h_{1x} - e_x h_{1y}}, \\ y_0'' &= \frac{e_y h'_{3x} - e_x h'_{3y} + e_x h''_{3y} - e_y h''_{3x}}{e_x h_{1y} - e_y h_{1x}}, \\ f &= \sqrt{\frac{-\tilde{x}_0'' e_z}{y_0'' h_{1z} + h'_{3z} - h''_{3z}}}, \\ \alpha &= \|K^{-1} \mathbf{h}_1\|_2, \\ \mathbf{t} &= \tilde{\mathbf{t}}/\alpha = (h'_{3x}/f \ h'_{3y}/f \ h'_{3z})^\top / \alpha. \end{aligned} \quad (8)$$

The estimated scale α and the focal length f allow us to fully recover the rotation matrix R , see Eq. (2), thus $\mathbf{r}_2 = (\alpha K)^{-1} \mathbf{h}_1$ and $\mathbf{r}_3 = (\alpha K)^{-1} \mathbf{h}_2$.

Overconstrained Solution. We can rewrite Eq. (6), resp. Eq. (7), in the form of the Quadratic Eigenvalue Problem (QEP),

$$(D_1 + D_2 f + D_3 f^2) \mathbf{v} = \mathbf{0}, \quad (9)$$

with known design matrices D_i , the unknown focal length f and the vector \mathbf{v} with unknown parameters

$$\mathbf{v} = (\tilde{\mathbf{t}}^\top \ 1 \ \tilde{x}_0^2 \ y_0^2 \ \dots \ \tilde{x}_0^K \ y_0^K)^\top. \quad (10)$$

Eq. (9) is linear in elements of \mathbf{v} and quadratic in the focal length. The design matrices D_1 , D_2 , D_3 are sparse of size $3K \times (2K + 2)$ such that each detection adds 3 rows. If $K = 2$, they are square and Eq. (9) is equivalent to the aforementioned Minimal Solution. For $K > 2$, the matrices become rectangular as each new pose adds three equations but only two unknowns.

The QEP is mathematically well understood problem [1] if the matrices are square. However, it is not our case since our design matrices are non-square leading to a *rectangular* QEP. To profit from the overconstrained system, we propose a method to solve the rectangular QEP in the next section and show in Sec. 4 a significant improvement of the overconstrained system to the final estimate.

2.2.3 Rectangular QEP

In order to use standard solvers, the rectangular Quadratic Eigenvalue Problem (QEP) in Eq. (9) could be squared by left multiplication by D_1^\top , as suggested by [6] in connection with simultaneous fundamental matrix estimation and lens distortion. This trick has the virtue of preserving the true solution in the noiseless case. It has been shown in [18] that such a solution suffers from bias and significant variance in the presence of noise when the approximation does not hold well. Moreover, in our case, the sparsity of the matrices does not allow to use this trick as it leads to singularities

and to the ill-conditioned system. We therefore propose a different strategy.

The rectangular QEP can be converted to a linear rectangular generalized eigenvalue problem through a linearization by introducing a new variable $\mathbf{w} = f\mathbf{v}$,

$$D_1 \mathbf{v} + f(D_2 \mathbf{v} + D_3 \mathbf{w}) = \mathbf{0}, \quad \mathbf{w} - f\mathbf{v} = \mathbf{0}, \quad (11)$$

which can be rewritten as

$$\left(\begin{bmatrix} D_1 & 0 \\ 0 & I \end{bmatrix} - f \begin{bmatrix} -D_2 & -D_3 \\ I & 0 \end{bmatrix} \right) \begin{bmatrix} \mathbf{v} \\ \mathbf{w} \end{bmatrix} = \mathbf{0}, \quad (12)$$

with I being an identity matrix, and simplified to a *rectangular* generalized eigenvalue problem

$$(A - fB)\tilde{\mathbf{v}} = \mathbf{0}. \quad (13)$$

Solving this equation for rectangular matrices A, B has been studied in [3]. They propose an iterative approach based on perturbation of the rectangular matrices A, B . It has been shown in [18] that this iterative technique significantly helps in finding a better, more noise resistant, solution for radial distortion parameter and fundamental matrix estimation.

Going back to our problem, we observed that the technique does not converge, resp. converges extremely slow, by which we believe is due to the sparse structure of the problem. We therefore propose a modification of the algorithm [3] by using a different updating of the focal length f in step 3, giving us much better performance and desired convergence. The solution for the focal length f and parameter vector $\tilde{\mathbf{v}}$, resp. \mathbf{v} , is obtained by the following algorithm.

1. Construct rectangular matrices A, B . Initialize f by solving a square QEP, details are given below.
2. Update $\tilde{\mathbf{v}}$ by the right singular vector corresponding to the smallest singular value of the matrix $(A - fB)$.
3. Update f by $\tilde{v}(4 + L)/\tilde{v}(4)$, where L is the length of the vector \mathbf{v} , $\mathbf{v}(4)$ stands for the 4th element of that vector. Note, that the 4th element of \mathbf{v} , Eq. (10), and so also of $\tilde{\mathbf{v}}$ should be one and the $(4+L)$ element of $\tilde{\mathbf{v}}$ is equal to f in noiseless situations.
4. Iterate steps 2-3 until convergence of f to a minimum.

The square QEP for initialization is obtained as follows. For each third and higher detection add only two rows into the matrices D_1, D_2, D_3 . Solve the square QEP, e.g. by `polyeig(D1,D2,D3)` in MATLAB or by converting it to the generalized eigenvalue problem and get an initial focal length f as one of the real eigenvalues. In general, there are $2(2K + 2)$ solutions of Eq. (9) for f and corresponding eigenvectors. In practice, there is typically only one real

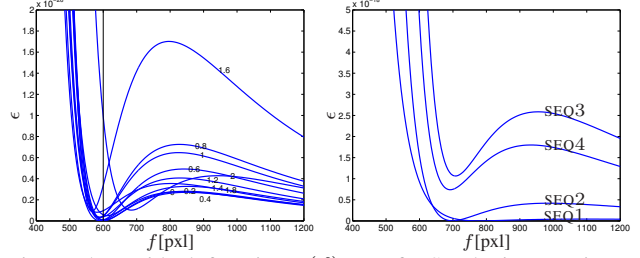


Figure 5. Residual function $\epsilon(f)$. Left: Synthetic experiment from Sec. 4 with the ground truth focal length $f = 600\text{pxl}$. The numbers above the curves correspond to noise level σ . Right: Real data. Four sequences with the same camera but at different orientations and persons are considered.

positive solution or none if the data is too noisy. The rest is zero, infinite or complex. The corresponding eigenvector is a solution for the vector \mathbf{v} in Eq. (10). However, if there are more solutions, they all can be evaluated by checking the validity of the solutions and only the best candidate further considered.

In principle, to solve Eq. (13) in the overconstrained case, one searches for

$$\epsilon^* = \min_{f, \tilde{\mathbf{v}}} \epsilon = \min_{f, \tilde{\mathbf{v}}} \frac{\|(A - fB)\tilde{\mathbf{v}}\|_2^2}{1 + f^2}, \quad \text{s.t. } \|\tilde{\mathbf{v}}\|_2^2 = 1, \quad (14)$$

see the proof in [3]. To investigate the behavior of the residual function we visualize the residual ϵ over a range of f . For each f , according to the step 2, we find the vector \mathbf{v} and evaluate ϵ . We evaluate that for noisy synthetic data as well as for real measurements, see Fig. 5. For real data, we run the overconstrained system on inliers only found by the Algorithm in Sec. 3. One can observe a strong local minimum around the expected value of f even for noisy data. We also see that the shift from ground truth position is in reasonable bounds what confirms the usability of the proposed overconstrained solution. The function is monotonic and decreasing between 0 and the optimal f . Thus, the minimum can be always found, e.g. by gradient descent starting with some small f .

2.2.4 Foot-head Homology

From the estimated calibration matrix K , rotation $R = [r_1 \ r_2 \ r_3]$ and translation \mathbf{t} , the foot-head homology can be constructed as

$$H_{FH} = K[r_1 \ r_2 \ lr_3 + \mathbf{t}][r_1 \ r_2 \ \mathbf{t}]^{-1}K^{-1}, \quad (15)$$

such that $\mathbf{u}_H \simeq H_{FH}\mathbf{u}_F$, where \mathbf{u}_F are homogeneous image points at foot locations and \mathbf{u}_H their corresponding head positions. The constant l is the height of a person in pixels, set to the height of the human model used for estimating the homographies in Eq. (2).

3. Algorithm

The method consists of the following steps

1. Detection of human bodies via the contour based method in Sec. 2.1, returning N best detections per image.
2. RANSAC: Random sampling of image pairs and in each image one of the best N detections. For each pair of the detections, the homography is estimated, Eq. (3), then the Minimal Solution solved, Eq. (8), and finally the foot-head homology estimated, Eq. (15). The Sampson distance [8] is used to indicate inliers and so to measure the quality of the homology estimate. The procedure is iterated until no further inliers can be found.
3. The inliers are used to solve the overconstrained system via the rectangular QEP to improve the estimate of the unknown variables.
4. Steps 1-3 are repeated, utilizing the current homology estimate, with more densely sampled contour template models. At each image location only models with head points in some pre-defined radius from the expected ones given by the homology are evaluated.

4. Experiments

Synthetic data We placed a 600×400 pxl camera with the focal length $f = 600$ pxl as shown in Fig. 2, and at each iteration, we randomly positioned ten planar mutually parallel human models in the scene and projected them into the image giving us the ground truth. To simulate inaccuracy in the contour based model detection stage, before projecting the models, we disturbed the models in the following manner. Each model was rotated in 3D by three Euler angles corresponding to zero mean Gaussian noise with variance σ^2 and scaled by the factor $1 + k$, where k is drawn by the same distribution but with $(\sigma/50)^2$. Such distorted models were further used for comparison and evaluation. It better simulates the real inaccuracy, the whole model was rotated and scaled instead of distorting just the individual corner points of the contour model in the image. It captures the situations when all persons do not stand parallel, are not all standing upright, and are not always of the same height. Two examples of distorted models at maximal investigated $\sigma = 2$ are shown in Fig. 6. At each noise level 100 trials were performed.

We evaluate three methods for computing the foot-head homology. First, the standard 8-point algorithm for homography estimation between the detected foot and head points [8]. Second, a method for homology computation given the vertical vanishing point and the horizon line, see

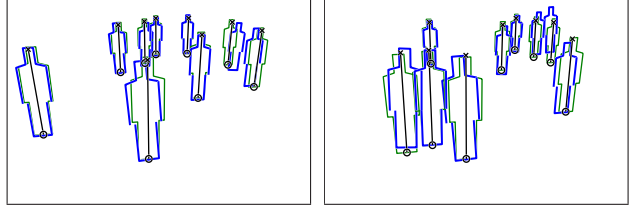


Figure 6. Synthetic data. Detections, shown by the blue thick models, at the noise level with $\sigma = 2$. Ground truth projections are shown by the thin green models. The estimated foot-head homology is visualized by black lines mapping each ground truth foot point marked by ‘o’ to a head position, marked by ‘x’.

p.83 [5]. We estimate the vanishing point in the least square sense by intersecting all lines connecting foot and head points re-normalized by the technique [11] considering the covariance matrices of the points. The horizon is estimated by fitting a line to the points computed as the intersections of lines composed of head-head and foot-foot points of all pose pairs. The vanishing point and the horizon are related as $\mathbf{l}_{inf} = (\mathbf{K}\mathbf{K}^T)^{-1}\mathbf{v}_p$. Assuming square pixels and known principal point, a linear algorithm for the focal length estimation can be designed [15]. Third, we compute the homology by our proposed technique with and without the overconstrained solution formulated in Sec. 2.2.2 as the QEP.

To measure residuals, we compute for each detection the distance of the ground truth head point and the point which is obtained by mapping the ground truth foot point by the estimated homology. It has no sense to explicitly investigate accuracy of factorized rotation and translation as we are particularly interested in correct homology within the image. We compute RMS error from all the detections in the image; its mean and standard deviation, and focal length estimates are plotted in Fig. 7. The result shows that our proposed method performs best as the noise level grows and provides very robust and stable estimate of the focal length. The plot confirms the known fact that estimating vanishing points and horizon line is very sensitive to noise and only small displacement can significantly affect the homology estimation and the focal length. The standard homography estimation allows neither to estimate the focal length nor to factorize the matrix into rotation and translation and it is presented to demonstrate that the factorization brings the stability into the estimation. The plot shows how the proposed overconstrained solution with the rectangular QEP contributes to the noise resistance and significantly outperforms the square QEP.

We repeated the same experiment but we shifted a principal point by 10 pixels in both directions. The result in Fig. 7c,d. shows that our method has already a bias at the beginning, however, the noise influence is prevalent and our method becomes soon more robust than the other methods.

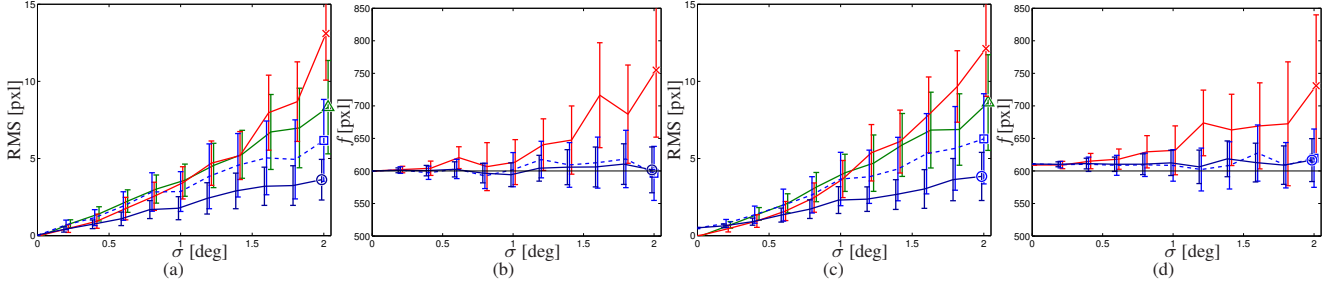


Figure 7. Synthetic data validation. (a): RMS residual error for: ‘o’ - the proposed method with the rectangular QEP; ‘□’ - the proposed method with the square QEP; ‘×’ - homology via a vanishing point and horizon; ‘△’ - the standard 8-point homography algorithm. (b): Focal length estimation and its dependency on noise level. (c),(d): The same experiment with a shifted principal point by 10 pixels in both directions.

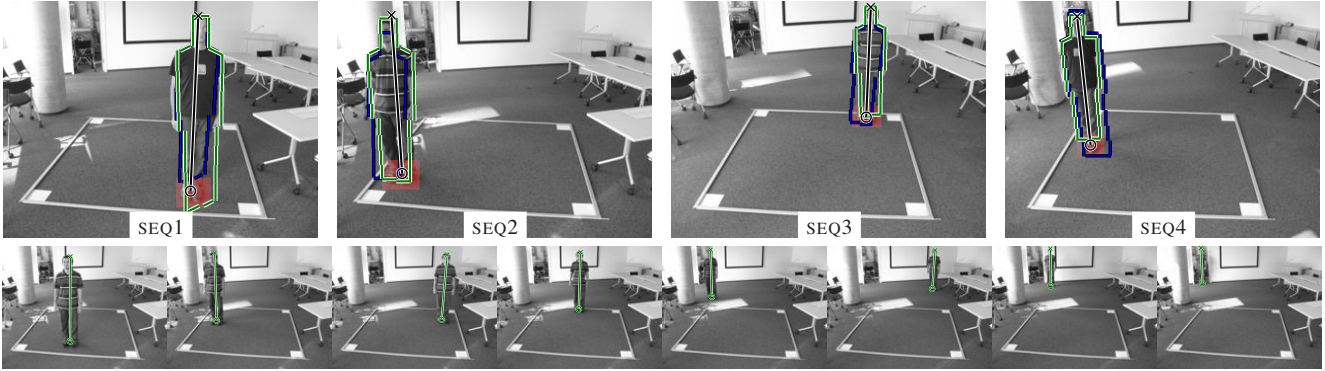


Figure 8. Real data. Top row: Each column corresponds to one of four sequences. Each image shows a detection before knowing a geometry and the detection after utilizing the estimated homology. Vertical lines represent the homology mapping manually established foot points marked by ‘o’ to head points ‘x’. Bottom row: Eight images from the SEQ2 with the estimated homology.

Real data We acquired four sequences with a Canon PowerShot G7 digital camera with a pre-set resolution of 640×480 pxl per image. Two different persons standing at random but roughly parallel positions were shot by the camera placed at two different heights and orientations resulting in four sequences SEQ1-4 with 20, 17, 15, and 12 images, respectively. The two persons differ in body proportions, they wear different clothes (notice that one person wears a T-shirt with a horizontal line pattern sometimes confusing the human detector and much looser trousers than the other one), see Fig. 8. The images possess quite strong radial distortion which can be observed by a deformed square lying on the ground plane. The radial distortion has not been modeled, despite that, the proposed method still delivers feasible results.

As for the synthetic data, we carried out experiments with the three methods. For the first method, the 8-point homography algorithm bundled in RANSAC is performed. The second method based on homology from a vertical vanishing point and horizon, both estimated in RANSAC framework with Liebowitz’s error [15] for measuring consistency of each line, given by foot-head points, to the estimated vanishing point. Third, we evaluate our proposed method summarized in Sec. 3.

For comparison and validation of the methods, we manually extracted foot and head points, and in the same spirit as for the synthetic data, we evaluated the RMS error and focal length estimate, reported in Tab. 1. To eliminate effect of the RANSAC procedure, we ran each method 50 times for all methods and report mean RMS error and the focal length estimate. One must take into account that the manual extraction of the foot and head points is error prone. Especially for the foot points, due to shoe occlusions, it is difficult to find the true points of the intersections of the ground plane and planes approximating the person. However, the relative differences between RMS errors tells us the qualitative performance difference of all the methods.

Tab. 1 shows that our method delivers the most stable result. The focal length is unknown to us, however, the EXIF tags in acquired images indicate a focal length of 665 pxl, which is close to our estimate. Having intuition from the synthetic experiment, we speculate that the focal length captures the inaccuracies coming from the off-the-center principal point and very likely the radial distortion. The vanishing point based homology method suffers, not surprisingly, from errors in the vanishing point estimation. The first two sequences are acquired by the camera having optical axis more parallel to the ground plane and therefore

		SEQ1	SEQ2	SEQ3	SEQ4
<i>Our method</i>	RMS	10.6	7.5	9.0	12.3
	foc.length	757	707	698	684
<i>VP-homology</i>	RMS	16.3	-	17.7	15.0
	foc.length	6748	imag	568	644
<i>8pt-homography</i>	RMS	14.5	19.7	16.3	20.4

Table 1. Results on real data. Both, RMS and the focal length are in pixels. Ground truth focal length is 665 pxl. All quantities in the table are in pixels.

the vanishing point closer to infinity. Perspective distortion in the vertical direction is less evident and, as well known, the estimation of the vanishing point much more sensitive to noise. That makes the estimate of the focal length in this case very biased and in the SEQ2 imaginary. However, our method is resistant to that and delivers comparable results to other two sequences.

To judge the quality of the results one must realize that human detectors deliver detections in much lower accuracy compared to point correspondences with subpixel accuracy utilized by standard calibration methods. The subpixel accuracy is necessary for Structure from Motion or 3D reconstruction but not in surveillance. The human detections can easily be shifted ± 5 pixels from the true positions in case of VGA resolution image, the persons do not stay always straight, and therefore the range of 10 pixels is a reasonable expectation for positioning the head location. The purpose of the calibration here is, first, to adjust the rotation and scale via the homology of a chosen human detector to let it running real time and suppress the missdetection rate. Second, to project the detections via the camera extrinsic calibration, *e.g.* centers of gravity of the detected bounding boxes, onto the ground plane *e.g.* as done manually in [7]. Tracking in the ground plane instead of the image plane contributes to more robust result and better handling of human mutual occlusions.

5. Conclusion

We proposed a method which significantly outperforms standard approaches utilized in surveillance systems for camera calibration and foot-head homology estimation. We found interesting relations when the assumption of parallel poses was adopted and we provided theoretical insight into that problem yielding a better numerical solution. Let us emphasize that the aim of our method is not to replace standard calibration techniques with sophisticated calibration targets when subpixel accuracy is required. What we show is that the accuracy required by surveillance applications can be efficiently obtained by our method without using any special calibration targets.

References

- [1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, 2000.
- [2] C. Beleznai and H. Bischof. Fast human detection in crowded scenes by contour integration and local shape estimation. In *CVPR*, 2009.
- [3] G. Boutry, M. Elad, G. H. Golub, P. Milanfar, and G. H. G. P. Milanfar. The generalized eigenvalue problem for non-square pencils using a minimal perturbation approach. *SIAM J. Matrix Anal. Appl.*, 27:582–601, 2005.
- [4] M. Bujnak, Z. Kukelova, and T. Pajdla. 3D reconstruction from image collections with a single known focal length. In *ICCV*, 2009.
- [5] A. Criminisi. *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*. Springer-Verlag, 2001.
- [6] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *CVPR*, pages (I):125–132, 2001.
- [7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *PAMI*, 30(2):267–282, 2008.
- [8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.
- [9] I. N. Junejo and H. Foroosh. Trajectory rectification and path modeling for video surveillance. In *ICCV*, 2007.
- [10] I. N. Junejo and H. Foroosh. Euclidean path modeling for video surveillance. *Image and Vision Computing*, 26(4):512–528, 2008.
- [11] K. Kanatani and Y. Sugaya. Statistical optimization for 3-D reconstruction from a single view. *IEICE Transactions on Information and Systems*, E88-D(10):2260–2268, 2005.
- [12] N. Krahnstoever and P. R. S. Mendonca. Bayesian autocalibration for surveillance. In *ICCV*, pages II:1858–1865, 2005.
- [13] L. Li and M. Leung. Unsupervised learning of human perspective context using ME-DT for efficient human detection in surveillance. In *CVPR*, 2008.
- [14] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *CVPR*, 2008.
- [15] D. Liebowitz. *Camera Calibration and Reconstruction of Geometry from Images*. PhD thesis, University of Oxford, 2001.
- [16] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *PAMI*, 28(9), 2006.
- [17] B. Micusik and T. Pajdla. Structure from motion with wide circular field of view cameras. *PAMI*, 28(7), 2006.
- [18] R. Steele and C. Jaynes. Overconstrained linear estimation of radial distortion and multi-view geometry. In *ECCV*, pages (I):253–264, 2006.
- [19] A. Toshev, A. Makadia, and K. Daniilidis. Shape-based object recognition in videos using 3D synthetic object models. In *CVPR*, 2009.