Robust Auto-Calibration from Pedestrians[*]

Imran Junejo and Hassan Foroosh
University of Central Florida, Orlando, FL 32816, U.S.A.

Abstract

*The knowledge of camera intrinsic and extrinsic parameters is useful, as it allows us to make world measurements. Unfortunately, calibration information is rarely available in video surveillance systems and it is difficult to obtain once the system is installed. Auto-calibrating cameras using moving objects (humans) has recently attracted a lot of interest. Two methods are proposed by Lv-Nevatia(*2002*) and Krahnstoever-Mendonça(*2005*). The inherent difficulty of the problem lies in the noise that is generally present in the data. We propose a* robust *and a general linear solution to the problem by adopting a formulation different from the existing methods. The uniqueness of formulation lies in recognizing two harmonic homologies present in the geometry obtained by observing pedestrians, and then using properties of these homologies to obtain linear constraints on the unknown camera parameters. Experiments with synthetic as well as on real data are presented - indicating the practicality of the proposed system.*

1 Introduction

Observation of human activities from stationary cameras is of significant interest to many applications. This is mainly due to the fact that the computer vision research has advanced to systems that can accurately detect, recognize and track objects as they move through a scene. Based on this, it is possible for other systems to make higher level inferences. For example, consider the problem of monitoring an area of interest (e.g. a building entrance, parking lot, port facility, an embassy, or an airport lobby) using stationary cameras where the intent is to monitor as large areas as possible by generally deploying non-overlapping cameras. The goal for such a system can be to model the behavior of objects of interest in such scenarios (e.g. cars or pedestrians, depending on the situation). Typically, one can employ path modeling techniques or activity learning techniques for single or multiple cameras (e.g. [5, 9, 10]) and even establish relations between the camera system [8, 14, 19]. It is known that due to perspective projection the measurements made from the images do not represent metric data. Thus the obtained object trajectories and consequently the associated probabilities represent projectively distorted data, unless we have a calibrated camera. This is evident from a simple observation: the objects grow larger and move faster as they approach the camera center, or two objects moving in parallel direction seem to converge at a point in the image. The projective camera thus makes it difficult to characterize objects - in terms of their sizes, motion characteristics, length ratios and so on - unless more information is available about the camera being used. This is where the camera calibration steps in.

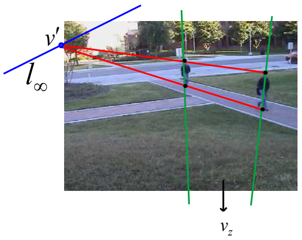


Figure 1. Auto-Calibration Geometry:A pedestrian, in two views, provides vertical vanishing points and an another vanishing points lying on the horizon line of the ground plane.

This paper proposes a robust auto-calibration method to estimate camera intrinsics and extrinsics by observing pedestrians in a scene. Original work on camera calibration using vanishing points started from the seminal paper by Caprile and Torre[2]. Liebowitz et al.[12] developed a method to compute the camera intrinsics by using the Cholesky Decomposition [16]. Similarly, Cipolla et al.[3] use three orthogonal vanishing points and one reference points to determine both intrinsic and extrinsic parameters.

For calibration using walking humans, Lv et al.[13] tracks humans to obtain the horizon line and the vanishing points to calibrate the camera. However, their formulation does not handle robustness issue. Also, they require a pedestrian to walk along two different directions for some

[*] The support of the National Science Foundation(NSF) through the Grant # IIS-0644280 is gratefully acknowledged.

duration or equivalently, two or more pedestrians along non-parallel directions for some duration. Recently Krahnstoever and Mendonça[11] proposed a Bayesian approach for auto-calibration by observing pedestrians. Foot-to-head homology is decomposed to extract the vanishing lines and the horizon line for calibration. They also incorporate measurement uncertainties and outlier models. However, their method requires prior knowledge about unknown calibration parameters and prior knowledge about the location of people; the proposed algorithm is also non-linear.

We propose a robust linear solution to estimate the camera intrinsic and extrinsic parameters by observing pedestrian. See Figure 1 for an example of the scenario. The detected head and feet locations of a person, over at least two instances, are used to estimate two harmonic homologies: head-to-foot and frame-to-frame. The former is referred to as the vertical homology, vertical vanishing points being the vertex. The later is referred to as the horizontal homology as the vertex lies on the horizon line. Linear constraints on the unknown camera parameters are obtained by using properties of these homologies. The noise in the data points is minimized by using total least squares method to solve an over-determined system of equations, where the outliers are removed by truncating the Rayleigh quotient [4].

A brief introduction to the concepts related to a pinhole camera are presented in Section 2. The unique geometry of the problem is explained in Section 3. The procedure to robustly determine the camera parameters is defined in Section 4. We present experimental results in Section 5 before concluding (Section 6).

2 Camera geometry

The projection of a **3D** scene point $\mathbf{X} \sim \begin{bmatrix} X & Y & Z & 1 \end{bmatrix}^T$ onto a point in the image plane $\mathbf{x} \sim \begin{bmatrix} x & y & 1 \end{bmatrix}^T$, for a perspective camera can be modeled by the central projection equation:

$$\mathbf{x} \sim \underbrace{\mathbf{K}\begin{bmatrix} \mathbf{R} & | - \mathbf{RC} \end{bmatrix}}_{\mathbf{P}} \mathbf{X}, \quad \mathbf{K} = \begin{bmatrix} \lambda f & \gamma & u_o \\ 0 & f & v_o \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where $\sim$ indicates equality up to a non-zero scale factor and $\mathbf{C} = \begin{bmatrix} C_x & C_y & C_z \end{bmatrix}^T$ represents camera center. Here $\mathbf{R} = \mathbf{R}_x\mathbf{R}_y\mathbf{R}_z = \begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{r}_3 \end{bmatrix}$ is the rotation matrix and $-\mathbf{RC}$ is the relative translation between the world origin and the camera center. The upper triangular 3×3 matrix $\mathbf{K}$ encodes the five intrinsic camera parameters: focal length f, aspect ratio λ, skew γ and the principal point at (u_o, v_o). As argued by [1, 15], it is safe to assume $\lambda = 1$ and $\gamma = 0$; moreover $(u_o = 0, v_o = 0)$ is assumed to lie in the center of the image.

Image of a family of parallel lines pass through a common point in the image. This point is referred to as the vanishing point. Since the proposed method uses only two vanishing points, without loss of generality, we refer to them as $\mathbf{v}_z$ for the vertical direction direction and $\mathbf{v}_x$ for the x-direction. Writing $\mathbf{P} \sim \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 \end{bmatrix}$, the vanishing points are given as: $\mathbf{v}_x \sim \mathbf{p}_1$ and $\mathbf{v}_z \sim \mathbf{p}_3$. Moreover, two families of parallel lines in mutually orthogonal direction intersect on a line called the horizon line. The horizon lines is given by $\mathbf{l}_\infty = \mathbf{p}_1 \times \mathbf{p}_2$ (See Figure 1).

The aim of camera calibration is to determine the calibration matrix $\mathbf{K}$. Instead of directly determining $\mathbf{K}$, it is common practice (for e.g. [6]) to compute the symmetric matrix $\boldsymbol{\omega} = \mathbf{K}^{-\mathbf{T}}\mathbf{K}^{-1}$ referred to as Image of the Absolute Conic(**IAC**). **IAC** is then decomposed uniquely using the Cholesky Decomposition [16] to obtain $\mathbf{K}$.

Knowledge of vanishing points of mutually orthogonal directions is used to put constraints on $\boldsymbol{\omega}$, which in our case is $\boldsymbol{\omega} = diag(w_{11}, w_{11}, 1)$.

Once the camera matrix $\mathbf{K}$ is determined, the camera extriniscs are extracted as:

$$\mathbf{r}_1 = \pm\frac{\mathbf{K}^{-1}\mathbf{v}_x}{\|\mathbf{K}^{-1}\mathbf{v}_x\|}, \mathbf{r}_3 = \pm\frac{\mathbf{K}^{-1}\mathbf{v}_z}{\|\mathbf{K}^{-1}\mathbf{v}_z\|}, \mathbf{r}_2 = \frac{\mathbf{r}_3 \times \mathbf{r}_1}{\|\mathbf{r}_3 \times \mathbf{r}_1\|}, \tag{2}$$

where $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ represent three columns of the **rotation matrix**. Due to special geometry of the problem, two of the three unknown angles are determined. The remaining angel is determined only up to a fixed rotation ambiguity. The sign ambiguity can be resolved by the cheirality constraint [6] or by known world information, like the maximum rotation possible for the camera.

3 Harmonic homologies from pedestrians

While observing pedestrians, one can notice the varied geometry associated with such a setup. As an object or a pedestrian of height h traverses the ground plane, each location on this plane corresponds to exactly one location on the head plane. Without loss of generality, for a simple case of two frames, as shown in Fig. 1, this correspondence can be mapped by a homography. In fact, this special type of homography is referred to as a homology [6]. We refer to this homology as the *vertical* homology:

$$\mathbf{H}_v = \mathbb{I} - \mu_v \frac{\mathbf{v}_z\mathbf{l}_1^{\mathbf{T}}}{\mathbf{v}_z^{\mathbf{T}}\mathbf{l}_1} \tag{3}$$

where $\mathbf{v}_z$ and $\mathbf{l}_1$ are, respectively, apex and axis of the homology. Another important geometric relation, so far ignored in existing literature on camera calibration from pedestrians, is the homology existing between different locations of a pedestrian. As shown in Fig. 1, since the height

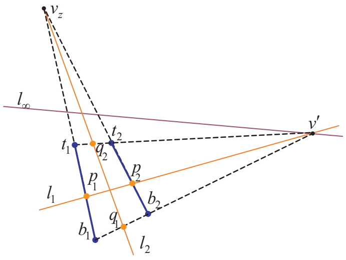
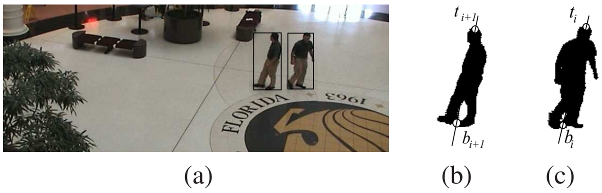


Figure 2. Harmonic Homologies: Tracking pedestrians over any two frames provides two harmonic homologies. See text for more details.



(a) (b) (c)

Figure 3. (a) shows an instance of a video sequences where a pedestrians is moving in the scene. (b) and (c) represent the pedestrian in two different frames. The head and foot location are denoted by t_i and b_i. See text for more details.

of a pedestrian is the same in all the frames, the line joining the head locations (t_1 and t_2) intersects the line joining the feet location (b_1 and b_2) at a point $\mathbf{v}'$ on the line at infinity ($\mathbf{l}_\infty$), forming another homology which we refer to as *horizontal* homology:

$$\mathbf{H}_h = \mathbb{I} - \mu_h \frac{\mathbf{v}'\mathbf{l}_2^{\mathbf{T}}}{\mathbf{v}'^{\mathbf{T}}\mathbf{l}_2} \tag{4}$$

where $\mathbf{l}_1, \mathbf{l}_2, \mathbf{v}'$ and $\mathbf{v}_z$ are as depicted in Fig. 2. A homology has five degrees of freedom. Therefore, three point correspondences are sufficient to uniquely determine a homology. An observation of the scenario at hand that reduces the required number of points to two is that the homologies observed so far can be replaced by *harmonic* homologies. Harmonic homology is a special type of a homology such that $\mu_h = \mu_v = -1$[17]. This reduces the degree of freedom to four i.e. only two point correspondences are needed to determine this harmonic homology. As shown in Fig. 2, t_1, b_1 corresponds to t_2, b_2, respectively to construct the harmonic homology ($\mathbf{H}_v$). Similarly, t_1, t_2 respectively corresponds to b_1, b_2 to determine $\mathbf{H}_h$. Also, a harmonic cross-ratio exists, given as:

$$Cross(\mathbf{v}_z, t_1, p_1, b_1) = \left(\frac{\overline{\mathbf{v}_z b_1}}{\overline{\mathbf{v}_z t_1}}\right) / \left(\frac{\overline{p_1 b_1}}{\overline{p_1 t_1}}\right) = -1 \tag{5}$$

Similarly, $Cross(\mathbf{v}', t_2, q_2, t_1) = -1$ for $\mathbf{H}_h$. [11] employs only the vertical homology (not *two harmonic homologies*) and therefore require more than two point correspondences to solve the problem.

Initial homology estimation: $\mathbf{H}_h$ and $\mathbf{H}_v$ are estimated from the detected head/foot location of an observed pedestrians. For instance, to estimate $\mathbf{H}_v$, $\mathbf{v}_z = \overline{t_1 b_1} \times \overline{t_2 b_2}$. The axis of the homology ($\mathbf{l}_1$) is obtained by using Eq. (5) to determine p_1 and p_2. $\mathbf{H}_v$ is obtained similarly. Using harmonic homology, instead of the standard 5 d.o.f homology, is more suited to the problem at hand. The main reason being that the object height does not change over various frames. Therefore, one can assume that two instances of object (i.e. a pedestrian detected at different time instance) correspond to the boundaries of a symmetric object. In our case, this symmetric object happens to be two parallel lines. Moreover, only two points correspondences are required to uniquely determine $\mathbf{H}_h$ and $\mathbf{H}_v$. Thus, in absence of noise, only two frames are required to calibrate the camera.

3.1 Determining head/foot locations

The proposed method requires point correspondences, which are head/feet positions of the pedestrians. Moving foreground objects (or region of interest), with shadows removed, can be extracted and tracked fairly accurately with statistical background models (for e.g. [8, 5, 18]). Lv et al.[13] perform eigendecomposition of the detected blob to extract head/feet location. An example of a detected pedestrian is shown in Fig. 3.

A simpler approach can be adopted to extract the head and foot location [11]. As shown in Fig. 3, these locations can easily be estimated by calculating the center of mass and the second order moment of the lower and the upper portion of the bounding box of the foreground region (Fig. 3(b)(c)).

4 Robust auto-calibration

Previous section illustrated the generality of the problem posed by pedestrians passing through a scene. The main issue with camera calibration by observing pedestrians is that head/feet detection is noisy. For example, a pedestrians may walk casually so that the posture might not be straight. Violations such as these result in measurements that can be viewed as *outliers*. Thus, some scheme needs to be adopted to minimize the influence of these outliers on *true* data points so that accurate results may be obtained.

By inspecting the geometry depicted in Fig. 2, further relations can be established between the two homologies and the vanishing points. The vertical vanishing point can thus be given as:

$$\widehat{\mathbf{v}_z} = \mathbf{l}_2 \times \boldsymbol{\omega} \mathbf{v}' \tag{6}$$

And also the $\mathbf{v}'$ can be given as:

$$\widehat{\mathbf{v}'} = \mathbf{l}_1 \times \boldsymbol{\omega} \mathbf{v}_z \tag{7}$$

This reformulated definition of the vanishing points can be used to constrain $\boldsymbol{\omega}$ using the harmonic cross-ratio (from $\mathbf{H}_v$) :

$$Cross(\widehat{\mathbf{v}_z}, t_1, p_1, b_1) + 1 = (\frac{\overline{\widehat{\mathbf{v}_z} b_1}}{\overline{\widehat{\mathbf{v}_z} t_1}}) / (\frac{\overline{p_1 b_1}}{\overline{p_1 t_1}}) + 1 = 0 \tag{8}$$

and similarly from $\mathbf{H}_h$:

$$Cross(\widehat{\mathbf{v}'}, t_2, q_2, t_1) + 1 = (\frac{\overline{\widehat{\mathbf{v}'} t_1}}{\overline{\widehat{\mathbf{v}'} t_2}}) / (\frac{\overline{q_2 t_1}}{\overline{q_2 t_2}}) + 1 = 0 \tag{9}$$

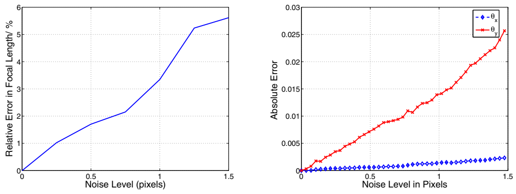
Unfortunately, Eq. (8) and Eq. (9) are not independent. Hence, we have only one constraint on $\boldsymbol{\omega}$. Unless we have more information, we can only solve for one unknown in $\boldsymbol{\omega}$. Fortunately, Eq. (8) and Eq. (9) are simplified into linear equations of the form: $a_i^j w_{11} + b_i^j = 0$, where the subscript i indicates the frame number and the superscript $j = \{1, 2\}$ indicates the two equations obtained per image pair using Eq. (8) and Eq. (9). Thus from each pair of images we obtain two equations with one unknown. Consequently, as each combination provides two equations, for n frames, $2 \times \begin{pmatrix} n \\ 2 \end{pmatrix}$ such combinations are possible. Equations obtained from a sequence are used to construct an over-determined system of equations:

$$\begin{bmatrix} a_1^1 & b_1^1 \\ a_1^2 & b_1^2 \\ \vdots & \vdots \\ a_n^1 & b_n^1 \\ a_n^2 & b_n^2 \end{bmatrix} \begin{bmatrix} w_{11} \\ 1 \end{bmatrix} = 0 \tag{10}$$

To increase the robustness, a suitable function should be selected that is less forgiving to outliers. One such example is the *truncated quadratic* function, commonly used in computer vision. The errors are weighted up to a fixed threshold, but beyond that, errors receive constant values. Thus the influence of outliers goes to zero beyond the threshold. We use the Rayleigh quotient to remove the outlier influence:

$$\rho(w_{11}) = \sum^{n} \frac{\mathbf{x}^\mathbf{T} \mathbf{A} \mathbf{x}}{\mathbf{x}^\mathbf{T} \mathbf{x}} < \xi \tag{11}$$

where $\mathbf{x} = \begin{bmatrix} w_{11} \\ 1 \end{bmatrix}$, $A = \begin{bmatrix} a_i^j & b_i^j \end{bmatrix}^\mathbf{T} \begin{bmatrix} a_i^j & b_i^j \end{bmatrix}$ and ξ is the threshold. The Rayleigh quotients are estimated from the observation points and ξ is set to the median of all the residual errors. Observation points having residual errors above ξ are removed as outliers.



(a) (b)

Figure 4. Performance of auto-calibration method VS. Noise level in pixels.

After outlier removal, the *outlier-free* observation points are used to construct the over-determined system of Eqs. (10). The system is then solved by TLS i.e. using the Singular Value Decomposition (SVD)[7]. The correct solution is the eigenvector corresponding to the smallest eigenvalue.

5 Experiments and results

In order to estimate the accuracy of the proposed method, we experiment with synthetic and the real data.

Synthetic data: We rigourously test the proposed method for estimating the camera parameter i.e. f. Twenty vertical lines of same height but random location are generated to represent pedestrian in our synthetic data. The ends of the lines indicate the head or the foot locations. We gradually add a Gaussian noise with $\mu = 0$ and $\sigma \leq 1.5$ pixels to the data-points making up the vertical lines. Taking two vertical lines at a time, the four points i.e. two head and two foot location are used to obtain $\mathbf{H}_h$ and $\mathbf{H}_v$. Vanishing points derived in Eqs. (6),(7) are substituted in to Eqs. (9), (8) to construct the over-determined system of equations, as described in Section 4. While varying the noise from 0.1 to 1.5 pixel level, we perform 1000 independent trials for each noise level, the results are shown in Figure 4. The relative error in f increases almost linearly with respect to the noise level. For a maximum noise of 1.5 pixels, we found that the error was under 7%. The absolute error in the rotation angles also increase linearly and is well under 1 degree.

Real Data: The proposed system has been tested on multiple sequences. The image sequences have a resolution of 320 × 240 pixels and captured at multiple locations. The tracker is able to accurately establish correspondences over a variety of environmental conditions. Different pedestrians from a single sequences are used to obtain the camera parameters. Then, as reported by [20], the mean of the estimated focal length is taken as the ground truth and the standard deviation as a measure of uncertainty in the results. This comparison of the results should be a good test of the stability and consistency of the proposed method. Due to space limitations, we only show results for the obtained focal lengths.

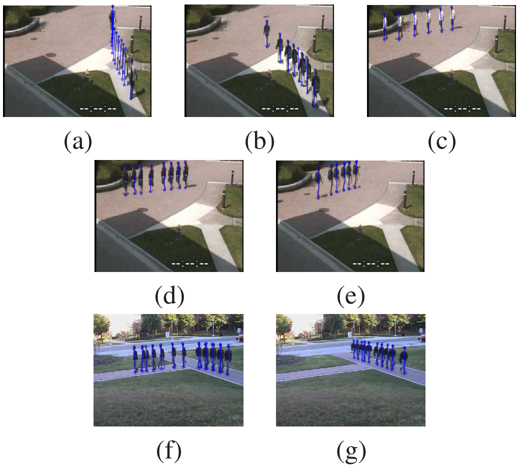

(a) (b) (c) (d) (e) (f) (g)

Figure 5. The figure depicts instances of the data set used for testing the proposed method. The estimated head and foot locations are marked with circle. Different frames are super-imposed on the background image to better visualize the test data.

Seq $\#1$	Recovered Focal Length (f)
Fig. 5a	$f = 2362.48$
Fig. 5b	$f = 2341.72$
Fig. 5c	$f = 2287.68$
Fig. 5d	$f = 2295.54$
Fig. 5e	$f = 2252.24$
Seq $\#2$	Recovered Focal Length (f)
Fig. 5f	$f = 2046.06$
Fig. 5g	$f = 1905.12$
Seq $\#3$	Recovered Focal Length (f)
Fig. 5f	$f = 840.68$
Fig. 5g	$f = 837.84$

Table 1. The recovered focal length for (*starting from top*) Seq $\#1$, Seq $\#2$ and Seq $\#3$. See text for more details.

Two video sequences are used for testing. **Seq** $\#1$ contains less than 5 minutes of data. As shown in Figure 5a-e, different pedestrians are chosen for auto-calibration. Using the method described above, the focal length is determined using the robust TLS method. The results for this sequences are given in Table 1(a). The standard deviation is low and the estimated focal length is $f = 2332.084 \pm 83.08$. **Seq** $\#2$ is a another sequence used for testing, a couple of instances are shown in Figure 5f-g. The estimated focal lengths are very close to each other, as shown in Table 1(b). The error in the results can be attributed to many factors. One of the main reason is that only a few frames are used per sequence. If a large data sequence is used, the system of equations (i.e. Eq. (10)) becomes more stable and thus better results may be obtained. The standard deviation in f for all our experiments is found to be less than the results reported by [11].

6 Conclusion

This paper presents a robust and a more general solution to camera calibration by observing pedestrians. Compared to existing methods, the solution does not assume any special kind of pedestrian motion. We recognize the special geometry of the problem and present formulation different from existing method. Two harmonic homologies are extracted from a pair of images containing instances of a pedestrian. Using unique properties of these homologies, linear constraints are derived to obtain the unknown camera parameters. The detected head/feet locations are used to robustly estimate the unknown camera parameters. We successfully demonstrate the proposed method on synthetic as well as on real data.

References

[1] L. D. Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zooming cameras. *Int. J. Comput. Vision*, 45(2):107–127, 2001.

[2] B. Caprile and V. Torre. Using vanishing points for camera calibration. *Int. J. Comput. Vision*, 4(2):127–140, 1990.

[3] R. Cipolla, T. Drummond, and D. Robertson. Camera calibration from vanishing points in images of architectural scenes. In *Proc. of BMVC*, pages 382–391, 1999.

[4] G. Golub and C. V. Loan. *Matrix Computations*. John Hopkins Press, 1989.

[5] W. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998.

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

[7] S. V. Huffel and P. Lemmerling. *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002.

[8] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In *the seventh European Conference on Computer Vision*, 2002.

[9] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *Proc. of British Machine Vision Conference (BMVC)*, 1995.

[10] I. Junejo, O. Javed, and M. Shah. Multi feature path modeling for video surveillance. In *17th conference of the International Conference on Pattern Recognition (ICPR)*, 2004.

[11] N. Krahnstoever and P. R. S. Mendonca. Bayesian autocalibration for surveillance. In *Tenth IEEE International Conference on Computer Vision*, 2005.

[12] D. Liebowitz and A. Zisserman. Combining scene and autocalibration constraints. In *Proc. IEEE ICCV*, pages 293–300, 1999.

[13] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In *IEEE International Conference of Pattern Recognition*, 2002.

[14] D. Makris and J. T.J. Ellis. Bridging the gaps between cameras. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2004.

[15] M. Pollefeys, R. Koch, and L. V. Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *Int. J. Comput. Vision*, 32(1):7–25, 1999.

[16] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[17] J. G. Semple and G. T. Kneebone. *Algebraic Projective Geometry*. Oxford Classic Texts in the Physical Sciences, 1979.

[18] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *Proc. IEEE ICCV*, 1998.

[19] K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *International Conference on Computer Vision*, 2005.

[20] Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334, 2000.