ROBUST AUTO-CALIBRATION USING FUNDAMENTAL MATRICES INDUCED BY PEDESTRIANS

Imran N. Junejo *Nazim Ashraf* *Yuping Shen* *Hassan Foroosh*

Computational Imaging Lab (CIL), 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 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 were 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 our formulation lies in recognizing two fundamental matrices present in the geometry obtained by observing pedestrians, and then using their properties to impose linear constraints on the unknown camera parameters. Experiments with synthetic as well as real data are presented - indicating the practicality of the proposed system.

***Index Terms*—** Video Surveillance, Camera Calibration, Fundamental Matrix.

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. Most of the video surveillance inovles, for instance, monitoring an area of interest (e.g. a building entrance, or an embassy) using stationary cameras where the intent is to monitor as large an area as possible. The goal for such a system can be to model the behavior of objects (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. [1]) and even establish relations between the camera system [2]. 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.

This paper proposes a robust auto-calibration method to estimate camera intrinsics and extrinsics by observing pedestrians in a scene. Many camera calibration techniques exits for different scenarios [3] but we limit ourselves with related work on camera auto-calibration from observing pedestrians. Lv et al. [4] were the first to propose calibration by recovering the horizon line and the vanishing points from observed walking humans. However, their formulation does not handle robustness issues. Recently Krahnstoever and Mendonça [5] proposed a Bayesian approach for auto-calibration by observing pedestrians. Foot-to-head homology is decomposed to extract the vanishing point and the horizon line for calibration. They also incorporate measurement uncertainties and outlier models. However, their method requires prior knowledge about some unknown calibration parameters and prior knowledge about the location of people; and their algorithm is also non-linear.

We propose a robust linear solution to estimate camera intrinsic and extrinsic parameters by observing pedestrians. See Fig. 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 fundamental matrices: *horizontal*- where the epipole lies on the horizon line, and *vertical* - the epipole is the vertical vanishing point. Linear constraints on the unknown camera parameters are obtained by using properties of these matrices. The noise in 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 [6].

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. BACKGROUND

The projection of a 3**D** 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 [7, 8], 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.

The aim of camera calibration is to determine the calibration matrix $\mathbf{K}$. Instead of directly determining $\mathbf{K}$, it is a common practice to compute the symmetric matrix $\boldsymbol{\omega} = \mathbf{K}^{-\mathbf{T}}\mathbf{K}^{-1}$ referred to as

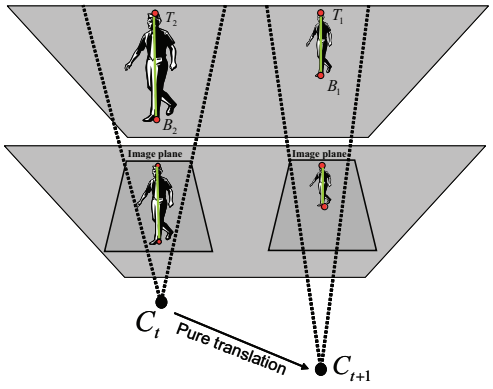


Fig. 1. **Observing pedestrians:** Instead of looking at the movement of a pedestrian, one can equivalently assume the world to be stationary and the camera to be translating. The two locations of the camera are denoted by $\mathbf{C}_t$ and $\mathbf{C}_{t+1}$ at time instance t and $t+1$, respectively. The epipole for such a translating camera lies at infinity. See text for more details.

Image of the Absolute Conic (**IAC**) [3]. **IAC** is then decomposed uniquely using the Cholesky Decomposition [9] to obtain $\mathbf{K}$.

Image of a family of parallel lines pass through a common point in the image. This point is referred to as the *vanishing point*. 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:

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

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 [3] or by known world information, for instance the maximum rotation possible for the camera.

3. FUNDAMENTAL MATRICES INDUCED FROM PEDESTRIANS

Our auto-calibration method is based on exploiting the fundamental matrices induced from pedestrians in a scene. The fundamental matrix satisfies the condition that for any pair of corresponding points $x \longleftrightarrow x'$ (in two images):

$$\mathbf{x'^T F x} = 0 \quad (3)$$

where the point $\mathbf{x'^T}$ is mapped to a line $\mathbf{l} = \mathbf{Fx}$ in the other image such that $\mathbf{x'^T l} = \mathbf{0}$ [3].

$\mathbf{F}$ is a rank 2 homogenous matrix with 7 d.o.f. In order to compute $\mathbf{F}$, in general at least 7 point correspondences are required. An important concept is the **epipole** - the image in one view of the camera center of the other view (cf. Fig 2a). It is also the vanishing point of the baseline (i.e. the line joining the two camera centers) direction. The epipole $\mathbf{e}$ is given as the right null-vector of $\mathbf{F}$: $\mathbf{Fe} = \mathbf{0}$. Similarly, $\mathbf{e}'$ is the left null-vector of $\mathbf{F}$.

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. As shown in Fig. 1, the head of the pedestrian is labeled as $\mathbf{T}_i$, while the feet as $\mathbf{B}_i$, where $i = 1, 2, \ldots, n$; n being the number of frames in which the pedestrian is visible. Without loss of generality, for a simple case of two frames, this head-to-feet correspondence can be mapped by a fundamental matrix.

Typically, the concept of fundamental matrix arises between multiple views taken from a camera, or equivalently from multiple cameras with overlapping field of view. In this paper, we focus on a special kind of a fundamental matrix that is *induced* by the pedestrian movements in a scene. *The key idea is*: instead of considering translation of the pedestrians (any two instances can be considered as being translating), one may equivalently consider the situation in which the camera undergoes translation, and the world is stationary. This is as depicted in Fig. 1. Thus the camera is considered non-stationary. This *re-formulation* of the problem allows us to introduce the concept of fundamental matrix, as described above, into our problem. Each instance of a pedestrian (head and feet location) can be treated as one single image. Thus any two instance of pedestrian induces a fundamental matrix between them.

When the motion of the camera is pure translational, the fundamental matrix has the form:

$$\mathbf{F_h} = [\mathbf{e}']_\times \mathbf{K}\mathbf{R}\mathbf{K}^{-1} = [\mathbf{e}']_\times \quad (4)$$

where $\mathbf{R} = \mathbf{I}$, $[\mathbf{e}']_\times$ is the skew-symmetric matrix representation of the epipole and $\mathbf{F_h}$ is defined as $\mathbf{T}_i^\mathbf{T}\mathbf{F_h}\mathbf{T}_j = 0$, where $i \neq j$. Note that $\mathbf{F_h}$ now has only 2 d.o.f., instead of 7 [3], which correspond to the position of the epipole. Therefore, only two point correspondences, $\mathbf{T}_i \longleftrightarrow \mathbf{T}_j, \mathbf{B}_i \longleftrightarrow \mathbf{B}_j$ where $i \neq j$, should be sufficient to compute $\mathbf{F_h}$. The two epipoles $\mathbf{e}$ and $\mathbf{e}'$ are also collinear.

$\mathbf{F_h}$ can be considered as mapping points in a direction parallel to the ground plane or *horizontally*. We can also introduce another fundamental matrix for the *vertical* direction such that $\mathbf{T}_i^\mathbf{T}\mathbf{F_v}\mathbf{B}_i = 0$. Thus now we are looking at the correspondences $\mathbf{T}_i \longleftrightarrow \mathbf{B}_i$. This is shown in Fig. 2b. The special epipolar geometry arising for a pure translating camera is depicted in Fig. 2a. As this figure shows, the intersection of the baseline with the image plane is at infinity. That is, the epipole lies at infinity or the epipole becomes a vanishing point.

Fig. 2b depicts the unique geometry induced by pedestrians. For any two instances of a pedestrian, the 2 d.o.f. $\mathbf{F_h}$ can be estimated by solving the following two linear equations:

$$\mathbf{T}_1^\mathbf{T}\mathbf{F_h}\mathbf{T}_2 = 0 \quad (5)$$

$$\mathbf{B}_1^\mathbf{T}\mathbf{F_h}\mathbf{B}_2 = 0 \quad (6)$$

Similarly, $\mathbf{F_v}$ can be estimated by solving:

$$\mathbf{T}_1^\mathbf{T}\mathbf{F_v}\mathbf{B}_1 = 0 \quad (7)$$

$$\mathbf{T}_2^\mathbf{T}\mathbf{F_v}\mathbf{B}_2 = 0 \quad (8)$$

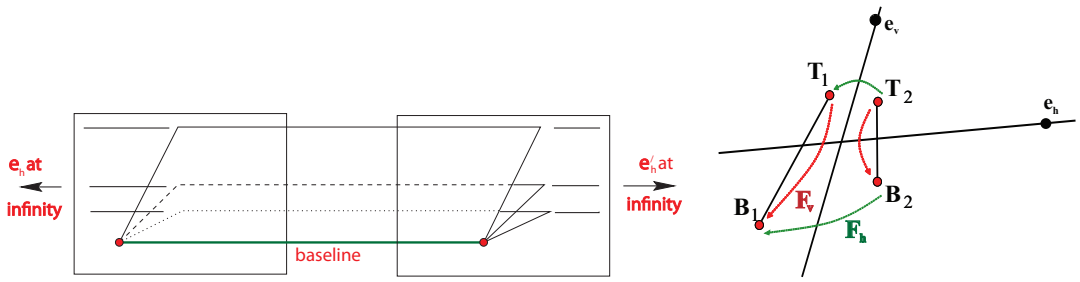
Once the fundamental matrix is determined, the epipole is computed as the null-vector of the fundamental matrix, as described above.

As Fig. 2a shows, the epipole $\mathbf{e_h}$ for $\mathbf{F_h}$ lies on the plane at infinity i.e. it is a vanishing point. Similarly $\mathbf{e_v}$ is also a vanishing point. These two vanishing points represent mutually orthogonal (horizontal and vertical) directions in world. Therefore, these points are used to enforce orthogonality constraint [3] on the IAC $\boldsymbol{\omega}$:

$$\mathbf{e_v^T}\boldsymbol{\omega}\mathbf{e_h} = 0 \quad (9)$$

Eq. (9) is a linear equation with an unknown parameter w_{11} of $\boldsymbol{\omega}$. Once w_{11} is determined, Cholesky decomposition is applied to $\boldsymbol{\omega}$ to obtain the camera calibration matrix $\mathbf{K}$.

Determining head/foot locations: The proposed method requires point correspondences, which are head/feet positions of the



(a) (b)

Fig. 2. (a) Epipolar geometry for pure translating camera (courtesy of [3]). The epipoles lie at infinity. (b) The two fundamental matrices, $\mathbf{F_v}$ and $\mathbf{F_h}$, induced by pedestrians.

pedestrians. Moving foreground objects (or regions of interest), with shadows removed, can be extracted and tracked fairly accurately with statistical background models (for e.g. [10, 1]). Lv et al.[4] perform eigendecomposition of the detected blob to extract head/feet locations. A simpler approach can be adopted: the head and feet 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 [5].

4. ROBUST AUTO-CALIBRATION

The main issue with camera auto-calibration by observing pedestrians is that head/feet detection is noisy. For example, a pedestrian 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 and noise on *true* data points so that accurate results may be obtained.

Eq. (9) provides only one constraint on $\boldsymbol{\omega}$. Unless we have more information, we can only solve for one unknown in $\boldsymbol{\omega} = \text{diag}(\omega_{11}, \omega_{11}, 1)$. Fortunately, this equation is linear and therefore can be simplified to the form: $a_i w_{11} + b_i = 0$, where the subscript i indicates the frame number. Thus from each image pair we obtain one equation with one unknown. Equations obtained from a sequence are used to construct an over-determined system of equations:

$$\underbrace{\begin{bmatrix} a_1 & b_1 \\ \vdots & \vdots \\ a_n & b_n \end{bmatrix}}_{\mathbf{Q}} \begin{bmatrix} w_{11} \\ 1 \end{bmatrix} = 0 \qquad (10)$$

The main goals of robust statistics is to recover the best structure that fits the majority of the model while rejecting the outliers. Thus, we need to recover the best w_{11} such that $\mathbf{K}$ is closest to the actual calibration matrix. The popular standard least squares (LS) estimation is extremely sensitive to outliers i.e. it has a breakdown point of zero. Therefore, Total Least Squares (TLS) method is adopted to solve the system of Eqs (10). Given an over-determined system of equations, TLS problem is to find the smallest perturbation to the data and the observation matrix to make the system of equations compatible. A suitable function also needs to be selected that is less forgiving to outliers, one such example is the *truncated quadratic* [11], commonly used in computer vision. The errors are weighted up to a fixed threshold, but beyond that, errors receive constant penalty. Thus the influence of outliers goes to zero beyond the threshold.

We use the truncated Rayleigh quotient to remove outlier influence. The quotients are estimated as:

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

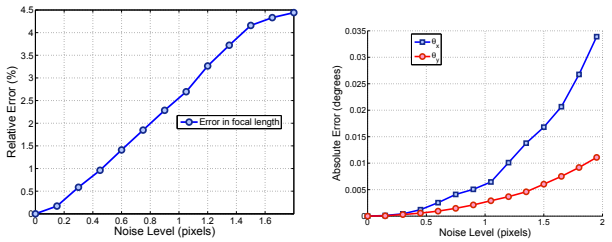
where $\mathbf{x} = \begin{bmatrix} w_{11} \\ 1 \end{bmatrix}$, $A = \begin{bmatrix} a_i^j & b_i^j \end{bmatrix}^{\mathbf{T}} \begin{bmatrix} a_i^j & b_i^j \end{bmatrix}$ and ξ is the threshold. The Rayleigh quotients are estimated from the observation points and the residual errors are estimated. The threshold ξ is set to the median of all the residual errors. Observation points obtained from Eq. 10 having residual errors greater than ξ are removed as outliers. After outlier removal, the *outlier-free* remaining observation points $\mathbf{Q}$ are used to construct the over-determined system of Eqs. (10). The system is then solved using the Singular Value Decomposition (SVD). The correct solution is the eigenvector corresponding to the smallest eigenvalue.

In summary, in order to minimize the influence of noise on our observation matrix $\mathbf{Q}$, we apply the Rayleigh quotient to *filter* out the noisy data points. Once the outliers are removed, the Total Least Squares method is applied to the remaining observation points to estimate the unknown parameter w_{11} of the IAC.

5. EXPERIMENTS AND RESULTS

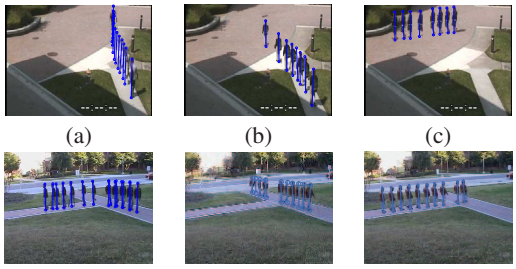
In order to estimate the accuracy of the proposed method, we experimented with synthetic and the real data.

Synthetic data: We rigourously tested the proposed method for estimating the camera parameter i.e. f. Eleven vertical lines of same height but random locations were generated to represent pedestrians 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 $\sigma \leq 2$ 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 $\mathbf{F_h}$ and $\mathbf{F_v}$. Vanishing points $\mathbf{e_v}$ and $\mathbf{e_h}$ are substituted in to Eq. (9) to construct the over-determined system of equations. While varying the noise from 0.1 to 2 pixel level, we perform 1000 independent trials for each noise level, the results are shown in Figure 3. The relative error in f increases almost linearly with respect to the noise level. For a maximum noise of 2 pixels, we found that the error was under 5%. The absolute error in the rotation



(a) (b)

Fig. 3. Performance of auto-calibration method VS. Noise level in pixels: (a) error in focal length. (b) error in the estimated angles.



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

Fig. 4. The figure depicts instances of the data set used for testing the proposed method. The estimated head and foot locations are marked with a circle. Different frames are super-imposed on the background image to better visualize the test data.

Seq #1	Recovered Focal Length (f)
Fig. 4a	$f = 955.31$
Fig. 4b	$f = 938.87$
Fig. 4c	$f = 952.05$
Seq #2	Recovered Focal Length (f)
Fig. 4d	$f = 1019.74$
Fig. 4e	$f = 976.09$
Fig. 4f	$f = 980.24$

Table 1. The recovered focal lengths for (*starting from top*) **Seq** #1 and **Seq** #2. See text for more details.

angles increases linearly and is well under 0.5 degrees.

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. Different pedestrians from a single sequences are used to estimate the camera parameters. Then, as reported by [12], 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.

Two video sequences are used for testing. **Seq** #1 contains less than 5 minutes of data. As shown in Fig. 4a-c, different pedestrians are chosen for auto-calibration. The results for this sequence are given in Table 1. The estimated focal length is $f = 948.74$ with a low standard deviation of $\sigma = 8.7$. **Seq** #2 is another sequence used for testing, and three instances are shown in Figure 4d-f. The estimated focal lengths are very close to each other: $f = 992.02$ with standard deviation of $\sigma = 24.09$, as shown in the table. 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 [5].

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 fundamental matrices are extracted from a pair of images containing instances of a pedestrian. Using unique properties of these matrices, 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] W.E.L. 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.

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

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

[4] Fengjun Lv, Tao Zhao, and Ramakant Nevatia, "Self-calibration of a camera from video of a walking human," in *IEEE International Conference of Pattern Recognition*, 2002.

[5] Nils Krahnstoever and Paulo R. S. Mendonca, "Bayesian auto-calibration for surveillance," in *Tenth IEEE International Conference on Computer Vision*, 2005.

[6] G.H. Golub and C.F. Van Loan, *Matrix Computations*, John Hopkins Press, 1989.

[7] L. De Agapito, E. Hayman, and I. Reid, "Self-calibration of rotating and zooming cameras," *Int. J. Comput. Vision*, vol. 45, no. 2, pp. 107–127, 2001.

[8] 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*, vol. 32, no. 1, pp. 7–25, 1999.

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

[10] Omar Javed and Mubarak Shah, "Tracking and object classification for automated surveillance," in *the seventh European Conference on Computer Vision*, 2002.

[11] Michael J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Journal of Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, January 1996.

[12] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.