Accurate self-calibration of two cameras by observations of a moving person on a ground plane

Tsuhan Chen

Advanced Multimedia Processing Lab
Carnegie Mellon University
Pittsburgh, PA, U.S.A.

Alberto Del Bimbo, Federico Pernici
and Giuseppe Serra

Dipartimento Sistemi e Informatica
University of Florence
Florence, Italy

Abstract

A calibration algorithm of two cameras using observations of a moving person is presented. Similar methods have been proposed for self-calibration with a single camera, but internal parameter estimation is only limited to the focal length. Recently it has been demonstrated that principal point supposed in the center of the image causes inaccuracy of all estimated parameters. Our method exploits two cameras, using image points of head and foot locations of a moving person, to determine for both cameras the focal length and the principal point. Moreover with the increasing number of cameras there is a demand of procedures to determine their relative placements. In this paper we also describe a method to find the relative position and orientation of two cameras: the rotation matrix and the translation vector which describe the rigid motion between the coordinate frames fixed in two cameras. Results in synthetic and real scenes are presented to evaluate the performance of the proposed method.

1. Introduction

The observation and recognition of human activity is one of the most important problem in visual surveillance. This problem is greatly simplified when cameras are calibrated, namely when the internal and external parameters of cameras are available.

For example, an accurate camera calibration is useful to compute the distance from a protected zone or to determine if people are walking or running. Moreover camera calibration has been used in tracking systems to accommodate change in object scale and to infer the depth-order of multiple objects in occlusion [1]. Standard methods of camera calibration use a calibration object or measurements of a sufficient number of 3D points in the scene [3, 4]. Unfortunately, such measurements are rarely available and difficult to obtain and this has inspired research in self-calibration methods. Vanishing points of parallel lines in 3D have also been used for this task [7, 6, 5]. These methods exploit vanishing points from static scene structures, such as buildings, but they can not be applied in scenes without structures. Recently a method that computes an approximate camera calibration observing moving objects was proposed by Stauffer *et al.* [9]. Criminisi *et al.* [10] describe how to perform 3D measurements in world coordinate from a *single camera* using a person in a planar scene; this method does not determine explicitly the internal and external parameters. Lv *et al.* [11] describe an algorithm to extract internal and external parameters from *single camera* exploiting the periodic motion of a walking human. This calibration method supposes zero skew, unit aspect ratio and that the y-coordinate of the principal point is half of the image height. Whereas zero skew and unit aspect ratio are two reasonable assumptions, the hypothesis that one coordinate of principal point is in the center of the image causes inaccuracy of all estimated parameters. Indeed, Hartley *et al.* [12] have proved that the determination of the focal length of the camera is tied very closely to the estimate of the principal point. Moreover, small changes in the estimated (sometimes merely guessed) principal point can cause very large changes in the estimated focal length.

Figure 1: The proposed method uses only corresponding image points of foot and head location between cameras to compute their internal parameters and their relative position.

The works of Nils *et al.* [19] and Junejo *et al.* [24] propose two methods for self-calibration of a *single camera* exploiting the homology between head and foot image points of a

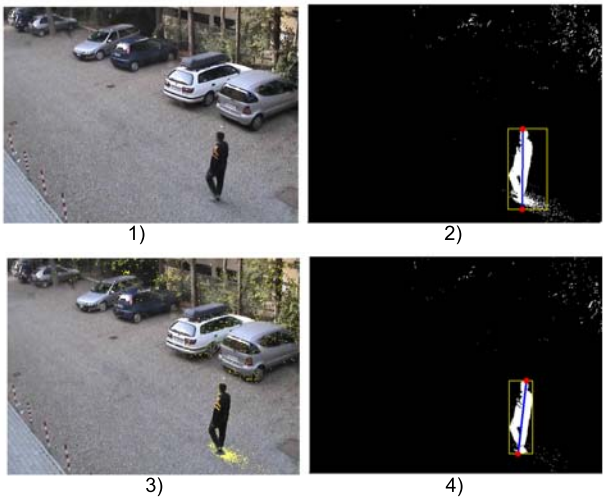


Figure 2: 1) Original image 2) Foot and head localization with shadow 3) Shadow detection 4) Foot and head localization after shadow removal.

moving person on a ground plane; only the focal length is computed.

Self-calibration of multiple views from observations of people has been studied as well. Liebowitz *et al.* [21] and Tresadern *et al.* [20] propose two self-calibration cameras methods using constraints of the articulated human structure, specifically the constant length between rotation joints over time. The main limit of these methods is the joints localizations accuracy. Sinha *et al.* [23] propose a method for self-calibration of cameras networks using silhouette of a person, but they need three or more views.

In this paper we present a method for calibration of two cameras based on features of a moving person in their common field of view. We use only the image of foot and head locations and we show how these points and their geometric relationship between cameras give enough information to find their relative position and orientation and the internal parameters of each camera, (*i.e.* the focal length and the principal point). We assume, like the previous methods, that the cameras have zero skew and unit aspect ratio.

The proposed self-calibration method works under the assumption that the scene needs to be modeled well with a dominant ground plane and the person is considered as a vertical segment of constant height. In particular the infinite homography, obtained from the fundamental matrix and the projective proprieties of the foot and head image locations, is exploited to transfer a linear constraint to one camera to the other. The height of the moving person must be identified in order to fix the scale factor of the translation vector. When multiple people are present in the scene, calibration can be performed individually on each of them.

The rest of the paper is organized as follows. The algorithm to determine the foot and head location is presented in Section 2. The method to compute the vanishing point, vanishing line and infinite homography is presented in Section 3. The approach to obtain internal parameters is presented in Section 4. The approach to obtain relative position of two cameras is presented in Section 5. Results are presented in Section 6 and the paper is concluded in Section 7.

2 Foot and head localization

To determine the moving objects in the scene *the Gaussian Mixture Model (GMM) for background subtraction* [13] is used. After localization, each moving blob is processed by the detector described by Saptharishi *et al.* [15] to determine if it is a person or not. Once the person is identified, the head and foot positions can be found by intersecting the smallest box around the person with its main axis (computed by the second order moment). Besides, for the save of effectiveness we use a shadow detection method [16] that converts the image from RGB color space to HSV color space and checks the chromaticity and the luminance (fig 2).

3 Vanishing point, Vanishing line and Infinite Homography

Considering different positions of a person on a ground plane, the lines passing through the head and the foot locations of each position are parallel and they intersect in a point at infinity. The image location of this point is a *vanishing point*, $\mathbf{v}_\infty$. Thus the vanishing point is computed intersecting all the image lines.

In addition, since the height of a person is the same all frames, for each couple of different person positions the line passing through heads and the line passing through feet are parallel and intersect in a vanishing line at infinity. The image of it is defined as the *vanishing line*, $\mathbf{l}_\infty$.

To ensure robustness of the vanishing points and lines estimation the RANSAC algorithm is used.

Moreover, for each frame we exploit the histogram distance defined in [17] to determine the matching person between two cameras. Thus, we have corresponding image points of foot and head between the cameras. The Fundamental matrix can be computed using corresponding image points of foot and head. In our situation, the head and foot locations form two planes respectively and thus the Fundamental Matrix can be computed by 6 image point correspondences and not 8 correspondences [18]. We can use 4 foot correspondences and 2 head correspondences or 4 head correspondences and 2 foot correspondences. Moreover we suppose that our cameras are synchronized (if the cameras are not synchronized it is possible to use the algorithm proposed by Caspi *et al.* [14]).

Once the vanishing points and vanishing lines of both cam-

eras and the Fundamental Matrix are recovered, the Infinite Homography matrix can be computed [18].

4 Computation of internal camera parameters

Constraints on intrinsic camera parameters are enforced in terms of the absolute conic ω ($\omega = \mathtt{K}^{-T}\mathtt{K}^{-1}$, where $\mathtt{K}$ is the intrinsic parameters matrix). Once ω is known, $\mathtt{K}$ can be computed by *Cholesky decomposition* of ω.
The matrix ω is symmetric, 3×3 matrix, defined up to scale, with five degrees of freedom. At least five constraints are needed to determine ω uniquely.
The zero skew assumption gives one linear constraint in ω, $\omega_{12} = \omega_{21}$ (where the low characters identify the row and column of the element into the matrix ω respectively). The square pixel assumption gives another linear constraint $\omega_{11} = \omega_{22}$ [18]. Furthermore, two linear constraints on ω result from the vanishing point and vanishing line arising from the ground plane and his vertical direction, $\mathbf{l}_\infty = \omega\mathbf{v}_\infty$.
Once the Infinite Homography $\mathtt{H}_\infty$ is known (sec. 3), the following equation can be used[18]:

$$\omega' = \mathtt{H}_\infty^{-T}\omega\mathtt{H}_\infty^{-1} \tag{1}$$

where ω and ω' are the image of the absolute conic in the first and in the second camera respectively. This equation gives a linear relation between ω and ω'; then zero skew assumption in the second camera gives one linear constraint. Finally ω can be computed linearly using the equations found before. Of course the same method can be used to find ω'.

5 Relative position and orientation of two cameras

Once the Fundamental matrix and the intrinsic camera matrix for each camera are known the Essential matrix can be computed. The relative position and relative orientation, namely the rotation matrix $\mathtt{R}$ and the translation vector $\mathbf{t}$ which describe the rigid motion between the coordinate frames fixed in two cameras, are extracted from the essential matrix [18]. It's known that this method finds four solutions. The correct solution is obtained by testing a single point $\mathbf{X}$ located in front of both cameras. The intersection of the plane G (see fig. 3) and the a ray back-projected from one image point of the first camera gives the searched 3D point. Note that the image point of the first camera can not be chosen randomly, because the 3D point may be not located in front of the second camera. According to this, one image foot correspondence is chosen.
We start considering two coordinate frames, one fixed

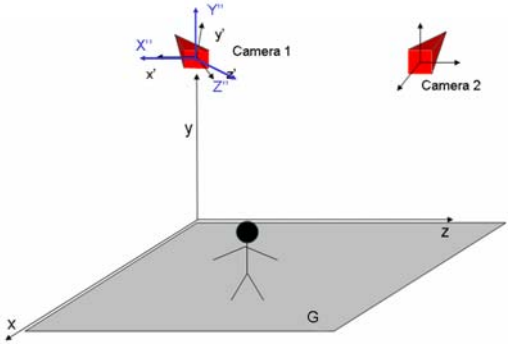


Figure 3: The figure shows all used coordinate frames, in particular it shows the first-camera coordinate frame (x', y', z'), world cooridnate frame (x, y, z) and the new coordinate frame (x'', y'', z'').

with the origin in the first camera center and the z-axis posed on principal axis (first-camera coordinate frame $C \equiv (x', y', z')$) and the other one posed with x-axis and z-axis on the ground plane and the y-axis passing through the origin of C (world coordinate frame $W \equiv (x, y, z)$). The fig. 3 shows these coordinate frames. First of all we determine the rotation and translation of the first camera respect to W. The first camera projection matrix wrt the world coordinate frame is:

$$\mathtt{P} = \mathtt{K}_1[\mathtt{R}_1|\mathbf{t}_1] \tag{2}$$

where

$$\mathtt{K}_1 = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

$$\mathtt{R}_1 = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\beta) & -\sin(\beta) \\ 0 & \sin(\beta) & \cos(\beta) \end{bmatrix} \begin{bmatrix} \cos(\alpha) & 0 & -\sin(\alpha) \\ 0 & 1 & 0 \\ \sin(\alpha) & 0 & \cos(\alpha) \end{bmatrix} \tag{4}$$

$$\mathbf{t}_1 = \begin{bmatrix} 0 \\ -h_1 \\ 0 \end{bmatrix} \tag{5}$$

where $\mathtt{K}_1$ is the internal matrix of the first camera, $\mathtt{R}_1$ is the rotation matrix from W to C, $\mathbf{t}_1$ is the translation vector from the origin of W to the origin of C and h_1 is the height of the camera from the ground plane.

To determine γ and β we note that the vanishing point $\mathbf{v}_\infty$ is the image point of $\mathbf{V}_\infty = [0, 0, 1, 0]^T$. Given the projective matrix $\mathtt{P} = [\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3]$ (where $\mathbf{p}_i$ is the i-th column of $\mathtt{P}$), we have $\mathbf{v}_\infty = [v_1, v_2, v_3]^T = \mathtt{P}\mathbf{V}_\infty = \mathbf{p}_2 = [p_1, p_2, p_3]^T$. This gives the following equation:

$$\frac{v_2}{v_3} = \frac{p_2}{p_3} = v_0 + f_1\cos(\gamma)\cot(\beta). \tag{6}$$

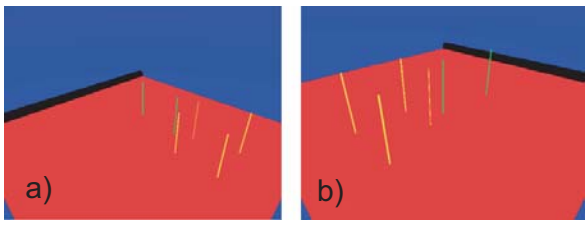


Figure 4: Example of two images taken of two cameras in the synthetic scene. a) First camera b) Second camera.

The vanishing line $\mathbf{l}_\infty$ is the image of the line at infinity of the planes parallel to the plane $z = 0$. The image point of $\mathbf{V}_0 = [1, 0, 0, 0]^T$ and $\mathbf{V}_1 = [0, 1, 0, 0]^T$ are both on the vanishing line and give $\mathbf{p}_0$ and $\mathbf{p}_1$ respectively. The vanishing line is hence given by $\mathbf{l}_\infty = [l_1, l_2, l_3]^T = \mathbf{p}_0 \times \mathbf{p}_1$, and so we can have:

$$\frac{l_1}{l_2} = -\tan(\gamma). \tag{7}$$

From the equations (6) and (7) we can easily find γ and β. Note that, in our case α is unknown, because the vanishing point and the vanishing line do not give any constraints of it.

The distance h_1 between the first camera and the ground plane of the eq. 5 is estimated by:

$$h_1 = \frac{1}{N} \sum_i \left[h / \left(1 - \frac{d(\mathbf{h}_i, \mathbf{c}_i) d(\mathbf{f}_i, \mathbf{v}_\infty)}{d(\mathbf{f}_i, \mathbf{c}_i) d(\mathbf{h}_i, \mathbf{v}_\infty)} \right) \right] \tag{8}$$

where N is the number of the foot collection data, h is the referent height of a person, $\mathbf{h}_i$, and $\mathbf{f}_i$ are the image point of head and foot of the i-th position of the moving person, $\mathbf{c}_i$ is the intersection of the line passing through $\mathbf{h}_i$ and $\mathbf{f}_i$ with the vanishing line $\mathbf{l}_\infty$. The equation 8 can be derived from the cross ratio invariance [10]. The equation of the plane G with respect to C can be determined. A point in the world coordinate frame $\mathbf{Q}_W$ and the same point in the first-camera coordinate frame $\mathbf{Q}_C$ are related by the following equation:

$$\mathbf{Q}_C = \mathtt{R}_1(\mathbf{Q}_W + \mathbf{t}_1). \tag{9}$$

To determine the equation of the plane we take the origin of the W, $\mathbf{X}_W = [0, 0, 0]^T$. We choose this particular point because it is independent wrt α and it belongs to the plane. So using equation (9) we find $\mathbf{X}_C$. The normal plane in the first-camera coordinate frame is $\mathbf{n} = \mathtt{K}^T \mathbf{l}_\infty$. Once the normal plane and a point on it are known the equation of the plane is recovered.

Furthermore taking an image point correspondence of foot, $\mathbf{q}$ and $\mathbf{q}'$, the back-projected ray of $\mathbf{q}$ intersects the plane G in the searched 3D point. Using this point we can find the correct solution.

It's remain to fix the scale factor of the translation vector $\mathbf{t}$. To this end, we consider a new coordinate frame (x'', y'', z'') with the origin in the center of C and the y-axes aligned to the y-axes of W (fig. 3). We can express the position of the second camera $\mathbf{C}_2$ ($\mathbf{C}_2 = -\mathtt{R}^T\mathbf{t}$) to the new coordinate frame as $\mathbf{C}_2^{''} = \mathtt{R}_1\mathbf{C}_2$. At this point to determine the scale of $\mathbf{C}_2^{''}$ we constraint the y-coordinate value to be equal to the difference between h_2 (height of the second camera wrt the ground plane) and h_1. One $\mathbf{C}_2^{''}$ is fixed, we can transfer back it to the first-camera coordinate frame, $\tilde{\mathbf{C}}_2$ ($\tilde{\mathbf{C}}_2 = \mathtt{R}_1^T\mathbf{C}_2^{''}$). Finally we have $\mathbf{t} = -\mathtt{R}\tilde{\mathbf{C}}_2$.

6 Experiments and Results

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

Synthetic data: To examine the performance of the proposed calibration algorithm we created a synthetic scene (fig. 4). The first and the second camera were located respectively 3 and 2 meters above a ground plane with a tilt angle of $\pi/6$. Both cameras had a focal length of $f = 480$, unit aspect ratio, zero skew and principal point at (320,240). The image resolution was 640 $\times$ 480 pixels.

In the scene randomly generated vertical segments of height 1,7 meters were inserted. The "foot" segments were positioned on the ground plane.

First, the performance of the estimation of internal and external parameters from data with noise was evaluated. Gaussian noise with zero mean and different standard deviations were added to heads and feet separately. The number of segments used was 1000, and for each value of the standard deviation the experiments were repeated 2000 times. Results in Table 1 show the root mean square errors between the estimated values and the true values of all estimated parameters for different values of the standard deviation. Very large standard deviation values were chosen in order to be as close as possible to real world conditions.

The focal length estimation is compared between our estimation algorithm and with the algorithm with the hypothesis of the principal point locate in the center of the image. Figure 5 shows the mean and standard deviation absolute error of focal length estimation for synthetic cameras where the principal point was moved out of the image center (the standard deviation of gaussian noise added to heads and feet was 0.5). It can be observed that the absolute mean error of our algorithm remains constant whereas the other case grows more than linear. The fig 5 also shows that small changes in the estimated of the principal point location cause large errors in the estimated focal length.

Real data: The proposed algorithm has been tested on multiple real sequences with a different walking person each. The image sequences had a resolution of 320 $\times$ 240 pixels and were captured at multiple locations and orientations. Three of them are shown in fig. 6 and 7. To evaluate the performance of our algorithm we tested its ability to recover

	Camera 1			Camera 2			Relative position					
							Translation			Rotation		
	f	u	v	f	u	v	x	y	z	α	β	γ
True value	480	320	240	480	320	240	-4.854m	-2.629m	2.554m	-58.80°	-72.42°	88.391°
Std. Gaus. Error.												
0,5	1.331	0.312	2.385	2.162	0.353	3.514	0.0296m	0.0131m	0.0114m	0.17°	0.30°	0.38°
1	2.314	0.601	4.085	4.027	0.688	6.504	0.0492m	0.0248m	0.0197m	0.29°	0.53°	0.72°
2	5.267	1.176	9.576	8.929	1.427	15.189	0.1261m	0.0512m	0.0453m	0.71°	1.17°	1.50°
3	8.601	2.101	15.976	14.807	1.954	26.851	0.2215m	0.0759m	0.0722m	1.25°	1.83°	2.17°
5	15.885	2.974	28.477	31.694	3.482	66.196	0.4212m	0.1522m	0.1194m	2.60°	3.15°	4.21°

Table 1: The top of the table shows the true values of the internal parameters of both cameras, f is the focal length, $[u, v]$ are the principal point coordinates, the values of translation vector $\mathtt{t} = [x, y, z]^T$ and rotation parameters α, β, γ. The bottom part shows the root mean square errors between the estimated values and the true values of all estimated parameters.

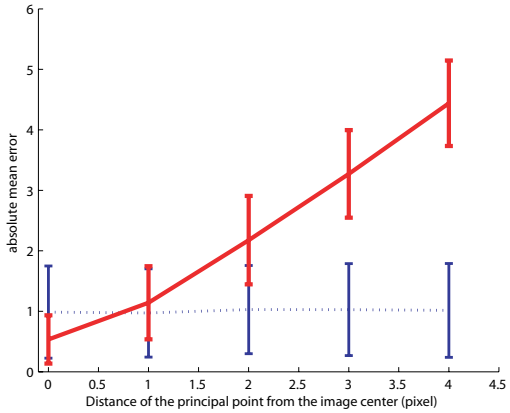


Figure 5: The dash line shows the absolute mean error of focal length estimation with our algorithm. The solid line shows the absolute mean error of the estimation of focal length supposing the principal point in the center of image. Their vertical lines identify the standard deviation errors. The gaussian error used for heads and feet has zero mean and 0.5 standard deviation.

distances between objects or the height of an object.
In the first sequence (fig. 6 see the left example) the heads and feet data were acquired and 70 measurements of distances were collected and five of which are shown in fig. 6 where their estimated values and the true values are reported. The mean error of 0,04m over all 70 measurements was estimated.
In the other sequence (fig. 6 see the right example) the heads and feet data were captured and 80 measurements of distances were exploited for evaluating the performance (Six measurements examples are shown in fig. 6). The mean error of all 80 measurements is 0,02m.
Fig. 7 shows the last example. The heads and feet data were captured and the performed measurements were about 70 and the estimation of the mean error is 0,05m.
Results show that our method is robust with regard to various viewing angles and camera positions. The experiments cover a wide range of tilt angles and relative positions of typical surveillance camera applications. Results on different walking people (four in total in all experiments) show that our method is also insensitive with regard to various subjects and various object lengths.

7 Conclusions

This paper proposed a novel method for self-calibration consisting of two different cameras. In particular the presented work is able to obtain calibration parameters completely automatically (*i.e.* focal length, principal point and relative position and orientation for both cameras). The method combines single and two view geometry of cameras viewing a common scene plane with a walking person.

Synthetic experiments show satisfactory results with regard to various levels of noise. The real experiments report good behavior in recovering the distances and object lengths.
An interesting direction for future research is to extend our approach in the case of three or four cameras. The special scene structure (a plane and vertical segments) can be exploited using trifocal and quadrifocal tensors.

Acknowledgment The authors would like thank Ted Square for many helpful discussions in the first part of this work.

References

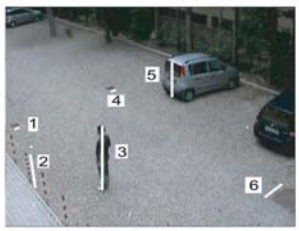
[1] T. Zhao, R. Nevatia, "Tracking Multiple Humans in Crowded Environment", *Int. Conf. on Computer Vision and Pattern Recognition*, 2004.

[2] B. Triggs, "Autocalibration and the absolute quadric", *Int. Conf. on Computer Vision and Pattern Recognition*, 1997.

[3] O. Faugeras, "Three-dimensional computer vision: a geometric viewpoint", *MIT Press*, 1993.

[4] Z. Zhang, "A flexible new technique for camera calibration", *Transactions on Pattern Analysis and Machine Intelligence*, 2000.

Measurements	Estimated values	True values
1	0,948m	0,980m
2	1,792m	1,800m
3	0,971m	0,940m
4	0,769m	0,800m
5	2,796m	2,820m



Measurements	Estimated values	True values
1	0,308m	0,330m
2	0,941m	0,940m
3	1,902m	1,880m
4	0,311m	0,330m
5	1,547m	1,550m
6	0,721m	0,750m

Figure 6: Examples of measurements show the performance to the algorithm in two different sequences.

Measurements	Estimated values	True values
1	1,923m	1,960m
2	1,864m	1,880m
3	0,31m	0,330m
4	3,521m	3,460m

Figure 7: Examples of measurements show the performance to the algorithm in an other different sequence.

[5] R. Cipolla and T. Drummond and D. Robertson, "Camera calibration from vanishing points in images of architectural scenes", *British Machine Vision Conf.*, 1999.

[6] D. Liebowitz and A. Zisserman, "Metric Rectification for Perspective Images of Planes", *Int. Conf. on Computer Vision and Pattern Recognition*, 1998.

[7] B. Caprile and V. Torre, "Using vanishing points for camera calibration", em Int. Journal of Computer Vision, 1990

[8] B. Bose and E. Grimson, "Ground Plane Rectification by Tracking Moving Objects", *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[9] C. Stauffer, K. Tieu and L. Lee, "Robust automated planar normalization of tracking data", *International Workshop on Visual Surveillance and Performance Evaluation in Tracking and Surveillance*, 2003.

[10] A. Criminisi, I. Reid and A. Zisserman, "Single View Metrology", *Int. Journal of Computer Vision*, 2000.

[11] F. Lv and T. Zhao and R. Nevatia, "Self-Calibration of a Camera from Video of a Walking Human", *Int. Conf. on Computer Vision and Pattern Recognition*, 2002.

[12] R. Hartley and R. Kaucic, "Sensitivity of Calibration to Principal Point Position", *European Conference on Computer Vision*, 2002.

[13] C. Stauffer And W. Grimson, "Adaptive background mixture model for real time tracking", *Int. Conf. on Computer Vision and Pattern Recognition*, 1999.

[14] Y. Caspi and D. Simakov and M. Irani, "Feature-Based Sequence-to-Sequence Matching", *Int. Journal of Computer Vision*, 2006.

[15] M. Saptharishi, J. B. Hampshire, and P. K. Khosla, "Agent-based moving object correspondence using differential discriminative diagnosis", *Int. Conf. on Conputer Vision and Pattern Recognition*, 2000.

[16] R. Cucchiara, C. Grana, M. Piccardi and A. Patri, "Improving shadow suppression in moving object detection with HSV color information", *Intelligent transportation System Conference Proceedings*, 2001.

[17] D. Comaniciu, V. Ramesh and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift", *Int. Conf. on Computer Vision and Pattern Recognition* , 2000.

[18] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision", *Cambridge University Press*, 2004.

[19] N. Krahnstoever and P. R. S. Mendonça, "Bayesian Autocalibration for Surveillance", *International Conference on Computer Vision*, 2005.

[20] P. Tresadern and I. Reid, "Uncalibrated and Unsynchronized Human Motion Capture : A Stereo Factorization Approach", *Int. Conf. on Computer Vision and Pattern Recognition*, 2004.

[21] D. Liebowitz and S. Carlsson, "Uncalibrated Motion Capture Exploiting Articulated Structure Constraints", *Int. Journal of Computer Vision*, 2003.

[22] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections", *Nature*, 1981.

[23] S. Sinha, M. Pollefeys and L. McMillan, "Camera Network Calibration from Dynamic Silhouettes", *Int. Conf. on Computer Vision and Pattern Recognition*, 2004.

[24] I. Junejo and H. Foroosh, "Robust Auto-Calibration from Pedestrians", *International Conference on Video and Signal Based Surveillance*, 2006.