Chapter 36
Robust Estimation of Heights of Moving People Using a Single Camera

Sang-Wook Park, Tae-Eun Kim and Jong-Soo Choi

Abstract In recent years, there has been increased interest in characterizing and extracting 3D information from video sequences for object tracking and identification. In this paper, we propose a single view-based framework for robust estimation of height and position. In this work, 2D features of a target object is back-projected into the 3D scene space where its coordinate system is given by a rectangular marker. Then the position and height are estimated in the 3D scene space. In addition, geometric error caused by an inaccurate projective mapping is corrected by using geometric constraints provided by the marker. The proposed framework is entirely non-iterative, and therefore is very fast. As the proposed framework uses a single camera, it can be directly embedded into conventional monocular camera-based surveillance/security systems. The accuracy and robustness of the proposed technique are verified on the experimental results of several real video sequences taken from outdoor environments.

Keywords Video surveillance · Smart surveillance · Security camera · Height estimation · Position estimation · Human tracking

36.1 Introduction

Vision-based human tracking is steadily gaining in importance due to the drive from various applications, such as smart video surveillance, human–machine interfaces, and ubiquitous computing. In recent years, there has been increased

S.-W. Park · J.-S. Choi
Department of Image Engineering, Chung-Ang University, 221 Heukseok-Dong, Dongjak-Gu, Seoul 156-756, Korea

T.-E. Kim
Department of Multimedia, Namseoul University, Cheonan 331-707, Korea

K. J. Kim and S. J. Ahn (eds.), *Proceedings of the International Conference on IT Convergence and Security 2011*, Lecture Notes in Electrical Engineering 120,
DOI: 10.1007/978-94-007-2911-7_36,

interest in characterizing and extracting 3D information from real-time video for human tracking. Emergent metrics are height, gait (an individual's walking style), and trajectory in 3D space [1–3]. Because they can be measured at a distance, and even from bad quality images, considerable research efforts have been devoted to use them for human identification and tracking. An important application is in security system, to measure dimensions of objects and people in images taken by surveillance cameras [4, 5, 6]. Because of bad quality of the image (taken by a cheap security camera), quite often it is not possible to recognize the face of a human or distinct features on his/her clothes. The height of the person may become, therefore, a very useful identification feature. Such a system is typically based on 3-dimensional metrology or reconstruction from two-dimensional images. Accordingly, it is extremely important to compute accurate 3-dimensional coordinates using projection of 3D scene space onto 2D image planes. In general, however, one view alone does not provide enough information for complete three-dimensional reconstruction. Moreover 2D to 3D projection, which is determined by the linear projective camera model, is defined up to an arbitrary scale; i.e. its scale factor is not defined by the projective camera model. Therefore, most single view-based approaches are achieved on the basis of geometric structures being resident in images, such as orthogonality, parallelism, and coplanarity. Vanishing points and vanishing lines are powerful cues, because they provide important information about the direction of lines and orientation of planes. Once these entities are identified in an image, it is then possible to make measurements on the original plane in three-dimensional space. In [4, 5, 6], excellent plane metrology algorithms to measure distances or length ratios on planar surfaces are presented. If an image contains sufficient information to compute a reference plane vanishing line and a vertical vanishing point, then it is possible to compute a transformation which maps identified vanishing points and lines to their canonical positions. The projective matrix which achieves this transformation allows reconstruction of affine structure of the perspectively imaged scene. By virtue of the affine properties, we can compute the relative ratio of lengths of straight line segments in the scene. This technique is relatively simple, and does not require that the camera calibration matrix or camera pose to be known. However, the geometric cues are not always available, and such methods cannot be applied in the absence of the scene structures. Alternatively, the position of an object on a planar surface in 3D space can be computed simply by using a planar homography. In this case, however, it is not possible to recover the original coordinates of a point which is not in contact with the reference plane in the scene. More popular approach to reconstruct three-dimensional structure is to employ multiple cameras [7–11]. By using multiple cameras, the area of surveillance is expanded and information from multiple views is quite helpful to handle issues such as occlusions. But the multiple camera-based approaches may bring some problems such as correspondence ambiguity between the cameras, inconsistency between images, and camera installation etc. For example, the feature points of an object extracted from different views may not correspond to the same 3D points in the world coordinate system. This may make the correspondence of feature point pairs ambiguous.

Furthermore, calibrations of multiple cameras are not a simple problem. In this paper, we propose a single view-based technique for the estimation of object height and position. Specifically, the target object is a human walking along the ground plane. Therefore a human body is assumed to be a vertical pole. Then we back-project the 2D coordinates of the imaged object into the three-dimensional scene to compute the height and position of the moving object. This framework requires a reference coordinate frame of the imaged scene. We use a rectangular marker to give the world coordinate frame. This marker is removed from the scene after the initialization phase. Finally, we apply a refinement approach to correct the estimated results by using geometric constraints provided by the marker. The proposed framework is entirely non-iterative, and is very fast. Therefore, the proposed method allows real-time acquisition of real position of a moving object as well as height in the 3D space. Moreover, as the projective mapping is estimated by using the marker, our method can be applied even in the absence of geometric cues. The remainder of this paper is structured in the following way: In Sect. 36.2, the proposed method is discussed, and experimental results are given in Sect. 36.3. The conclusions are drawn in Sect. 36.4.

36.2 Proposed Framework

36.2.1 Foreground Blob Extraction

An assumption throughout the proposed method is the linear projective camera model. This assumption is often violated by wide-angle lenses, which are frequently used in surveillance cameras. Those cameras tend to distort the image, especially near its boundaries, and this may affect metrology algorithm considerably. Therefore, we apply the radial distortion correction method introduced in [12] before the main process. After the preprocessing step, we are given a quartic polynomial function which transforms the distorted feature points into correct ones. In the proposed method, only the feature points are corrected because of the processing time. The foreground region is extracted by the statistical background subtraction technique presented in [13] which is robust to the presence of shadows. The main idea of this method is to learn the statistics of properties of each background pixels over N pre-captured background frames and obtain statistical values for the background. Based on this, the algorithm can classify each pixel into "moving foreground," "original background," "highlighted background," and "shaded background" after getting its new brightness and chromaticity values. The color model for the foreground extraction is illustrated in Fig. 36.2, which depicts that it separates the brightness from the chromaticity component (Fig. 36.1).

In Fig. 36.2, $I(i)$ is the color value of the ith pixel and $E(i)$ the expected color value of this pixel, for which coordinates ($\mu_R(i)$, $\mu_G(i)$, $\mu_B(i)$) are mean values of the RGB components of this pixel obtained during the training phase. $J(i)$ is the

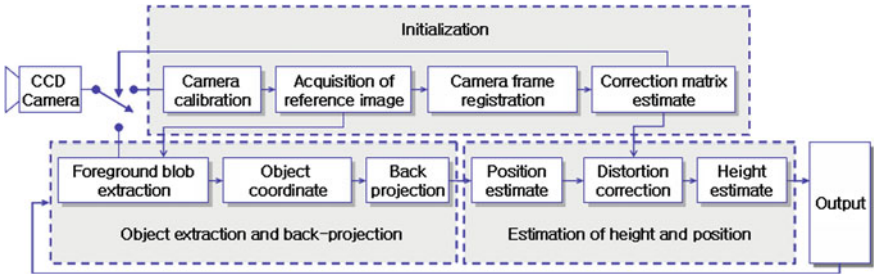


Fig. 36.1 Block diagram of the proposed method

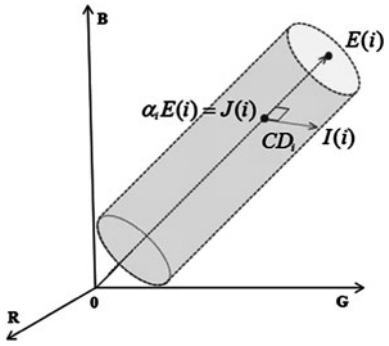


Fig. 36.2 The computational color model for the foreground extraction

projection of $I(i)$ onto the line $OE(i)$. The brightness and color distortion are computed as Eqs.(36.1) and (36.2), respectively.

$$\alpha_i = \arg\min_{\alpha_i} \left[\left(\frac{I_R(i) - \alpha_i \mu_R(i)}{\sigma_R(i)} \right)^2 + \left(\frac{I_G(i) - \alpha_i \mu_G(i)}{\sigma_G(i)} \right)^2 + \left(\frac{I_B(i) - \alpha_i \mu_B(i)}{\sigma_B(i)} \right)^2 \right], \tag{36.1}$$

$$CD_i = \sqrt{\left(\frac{I_R(i) - \alpha_i \mu_R(i)}{\sigma_R(i)} \right)^2 + \left(\frac{I_G(i) - \alpha_i \mu_G(i)}{\sigma_G(i)} \right)^2 + \left(\frac{I_B(i) - \alpha_i \mu_B(i)}{\sigma_B(i)} \right)^2}. \tag{36.2}$$

Here, $\sigma_R(i)$, $\sigma_G(i)$, and $\sigma_B(i)$ denote standard deviations of the ith pixel's RGB components, which are computed during the training phase. In our system, we only extract the foreground region from the rest. After the background subtraction, we use morphological operators to remove small misclassified blobs. Humans are roughly vertical while they stand or walk. In order to measure the height of a human in the scene, a vertical line should be detected from the image. However, the vertical line in the image may not be vertical to the ground plane in the real

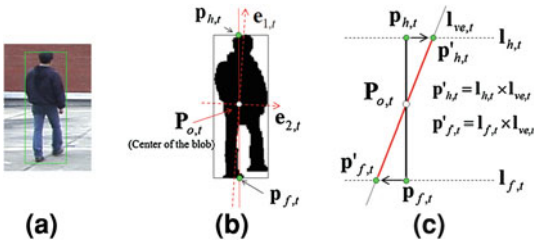


Fig. 36.3 Extraction of head and feet points; **a** captured image **b** estimation of principal axis using eigenvectors, and **c** extraction of the head and feet points

world space. Therefore, a human body is assumed to be a vertical pole that is a vertical principal axis of the foreground region.

We first compute the covariance matrix of the foreground region, and estimate two principal axes of the foreground blob. And a bounding rectangle of the foreground blob is detected. Then we compute intersections of the vertical principal axis and the vertical bounds of the foreground blob. These two intersections are considered as the apparent positions of the head and feet, which are back-projected for the estimation of the height and position. As shown in Fig. 36.3, let ($\mathbf{e}_{1,t}$, $\mathbf{e}_{2,t}$) be the first and second eigenvectors of the covariance matrix of the foreground region at frame t, respectively. Then, $\mathbf{e}_{1,\,t}$ and the center of the object blob $\mathbf{P}_{o,t}$ give the principal axis $\mathbf{l}_{ve,t}$ of the human body at the time step t. Given $\mathbf{l}_{ve,t}$, the intersections can be computed by cross products of each lines. The head and feet positions then are $\mathbf{p}'_{h,t}$ and $\mathbf{p}'_{f,t}$, respectively.

36.2.2 Back-Projection

In our method, the height and position are measured by using the back-projected features in three-dimensional scene space. Let $\tilde{\mathbf{M}} = [XYZ1]^{\mathrm{T}}$ be the 3D homogeneous coordinates of a world point and $\tilde{\mathbf{m}} = [xy1]^{\mathrm{T}}$ be the 2D homogeneous coordinates of its projection in the image plane. This mapping is defined by a linear projective transformation as follows.

$$\tilde{\mathbf{m}} = \lambda\tilde{\mathbf{P}}\tilde{\mathbf{M}} = \lambda\mathbf{K}[\mathbf{R}|\mathbf{t}]\tilde{\mathbf{M}} = \lambda\mathbf{K}[\mathbf{r_1}\mathbf{r_2}\mathbf{r_3}|\mathbf{t}]\tilde{\mathbf{M}}, \tag{36.3}$$

where λ is an arbitrary scale factor, and a 3×4 matrix $\tilde{\mathbf{P}}$ is called a projective camera matrix, which represents the projection of 3D scene space onto a 2D image plane. $\mathbf{R}$ is a 3×3 rotation matrix, and $\mathbf{t}$ denotes translation vector. And $\mathbf{r}_i$ means the ith column vector of the projection matrix. We use '$\sim$' notation for the homogeneous coordinate representation. The non-singular matrix $\mathbf{K}$ represents a camera calibration matrix, which consists of the intrinsic camera parameters. In our method, we employ the calibration method proposed by Zhang in [14]. This method computes the IAC (the image of absolute conic) $\boldsymbol{\omega}$ using the invariance of the circular points which are the intersections of a circle and the line at infinity $\mathbf{l}_{\infty}$.

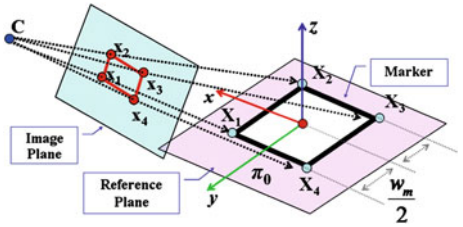


Fig. 36.4 Projective mapping between the marker and its image

Once the IAC $\boldsymbol{\omega}$ is computed, the calibration matrix $\mathbf{K}$ can be computed by $\boldsymbol{\omega}^{-1} = \mathbf{K}\mathbf{K}^{\mathrm{T}}$. Thus this method requires at least three images of a planar calibration pattern observed at three different orientations. From the calibrated camera matrix $\mathbf{K}$ and (36.3), the projective transformation between the 3D scene and its image can be determined. In particular, the projective transformation between a plane in the 3D scene and the image plane can be defined by 2D homography. Consequently, if four points on the world plane and their images are known, then it is possible to compute the projection matrix $\tilde{\mathbf{P}}$. Suppose that $\boldsymbol{\pi}_0$ is the XY-plane of the world coordinate frame in the scene, so that points on the scene plane have zero Z-coordinate. If four points $\tilde{\mathbf{X}}_1 \sim \tilde{\mathbf{X}}_4$ of the world plane are mapped onto their image points $\tilde{\mathbf{X}}_1 \sim \tilde{\mathbf{X}}_4$, then the mapping between and $\tilde{\mathrm{m}}_p = [\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_2\tilde{\mathbf{X}}_3\tilde{\mathbf{X}}_4]$ which consist of $\tilde{\mathbf{X}}_n = [X_n Y_n 01]^{\mathrm{T}}$ and $\tilde{\mathbf{x}}_n = [x_n y_n 1]^{\mathrm{T}}$ respectively is given by

$$\tilde{\mathbf{m}}_p = \mathbf{K}[\mathbf{R}|\mathbf{t}]\tilde{\mathbf{M}}_p = [\mathbf{p_1}\mathbf{p_2}\mathbf{p_3}\mathbf{p_4}]\tilde{\mathbf{M}}_p. \tag{36.4}$$

Here, $\mathbf{p}_i$ is ith column of the projection matrix. In this paper, $\tilde{\mathbf{X}}_n$ is given by four vertices of the rectangular marker. From the vertex points and (36.4), we have

$$\mathbf{K}^{-1}\begin{bmatrix} x_n \\ y_n \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11}X_n + r_{12}Y_n + t_x \\ r_{21}X_n + r_{22}Y_n + t_y \\ r_{31}X_n + r_{32}Y_n + t_z \end{bmatrix}, \tag{36.5}$$

where (x_n, y_n) is nth vertex detected from the image. And r_{ij} represents the element of the rotation matrix $\mathbf{R}$, t_x, t_y, and t_z the elements of the translation vector $\mathbf{t}$. From (36.5) and the four vertices, we obtain the translation vector $\mathbf{t}$ and the elements of the rotation matrix r_{ij}. By the property of the rotation matrix, the third column of $\mathbf{R}$ is computed by $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$. Assuming that the rectangular marker is a square whose sides have length w_{m}, and defining $\tilde{\mathbf{M}}_p$ as (36.6), the origin of the world coordinate frame is the center point of the square marker. In addition, the global scale of the world coordinate frame is determined by w_{m}. The geometry of this procedure is shown in Fig. 36.4.

$$\tilde{\mathbf{M}}_p = \begin{bmatrix} w_m/2 & w_m/2 & -w_m/2 & -w_m/2 \\ w_m/2 & -w_m/2 & -w_m/2 & w_m/2 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{36.6}$$

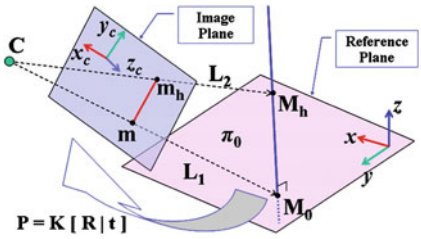


Fig. 36.5 Back-projection of 2D features

In general, the computed rotation matrix $\mathbf{R}$ does not satisfy with the properties of rotation matrix. Let the singular value decomposition of $\mathbf{R}$ be $\mathbf{U\Sigma V}^{\mathrm{T}}$, where $\mathbf{\Sigma} = diag(\sigma_1, \sigma_2, \sigma_3)$. Since a pure rotation matrix has $\mathbf{\Sigma} = diag(1, 1, 1)$, we set $\mathbf{R} = \mathbf{UV}^{\mathrm{T}}$ which is the best approximation matrix to the estimated rotation matrix [15]. An image point $m = (x, y)$ back-projects to a ray in 3D space, and this ray passes through the camera center as shown in Fig. 36.5. Given the camera projection matrix $\tilde{\mathbf{P}} = [\mathbf{P} \quad \tilde{\mathbf{p}}]$, where $\mathbf{P}$ is a 3×3 submatrix, the camera center is denoted by $\mathbf{C} = -\mathbf{P}^{-1}\tilde{\mathbf{p}}$. And the direction of the line $\mathbf{L}$ formed by the join of $\mathbf{C}$ and m can be determined by its point at infinity $\tilde{\mathbf{D}}$ as follows.

$$\tilde{\mathbf{P}}\tilde{\mathbf{D}} = \tilde{\mathbf{m}},\ \tilde{\mathbf{D}} = [\mathbf{D}\,0]^T, \tag{36.7}$$

$$\mathbf{D} = \mathbf{P}^{-1}\tilde{\mathbf{m}},\ \tilde{\mathbf{m}} = [\mathbf{m}^T 1]^T. \tag{36.8}$$

Then, we have the back-projection of m given by

$$\mathbf{L} = -\mathbf{P}^{-1}\tilde{\mathbf{p}} + \lambda\mathbf{P}^{-1}\tilde{\mathbf{m}} = \mathbf{C} + \lambda\mathbf{D},\ -\infty < \lambda < \infty. \tag{36.9}$$

36.2.3 Estimation of Height and Position

In our method, a human body is assumed to be a vertical pole. As shown in Fig. 36.5, the height of the object is the distance between $\mathbf{M}_0$ and $\mathbf{M}_{\mathrm{h}}$, and its position is $\mathbf{M}_0$ which is the intersection of the reference plane $\boldsymbol{\pi}_0$ and the line $\mathbf{L}_1$. Assuming that the line segment $\mathbf{M}_0 \sim \mathbf{M}_{\mathrm{h}}$ is mapped onto its image $\mathbf{m}_0 \sim \mathbf{m}_{\mathrm{h}}$, the intersection can be denoted as $\mathbf{M}_0 = \mathbf{C} + \lambda_0\mathbf{P}^{-1}\tilde{\mathbf{m}}0$, where λ_0 is a scale coefficient at the intersection point. Since $\mathbf{M}_0$ is always located on the reference plane $\boldsymbol{\pi}_0$, we have

$$\tilde{\pi}_0^T\tilde{\mathbf{M}}_0 = 0,\ \tilde{\pi}_0 = [0\,0\,1\,0]^T,\ \tilde{\mathbf{M}}_0 = [\mathbf{M}_0\,1]^T. \tag{36.10}$$

Then, from $\pi_0^{\mathrm{T}}\mathbf{M}_0 = \pi_0^{\mathrm{T}}(\mathbf{C} + \lambda_0\mathbf{P}^{-1}\tilde{\mathbf{m}}_0)$, we can uniquely determine λ_0 as following:

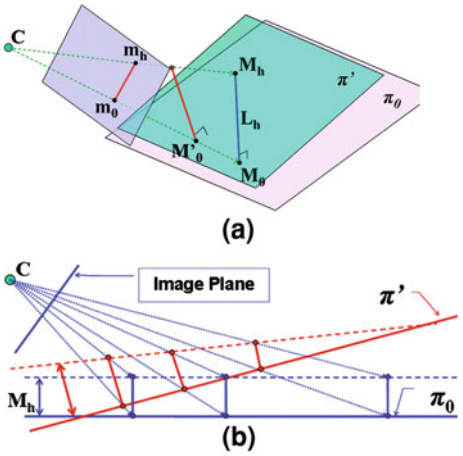


Fig. 36.6 Distortion of 2D–3D projective mapping due to inaccurate camera calibration: **a** projective relationship, **b** side view of **a**

$$\lambda_0 = -\frac{\boldsymbol{\pi}_0^T \mathbf{C}}{\boldsymbol{\pi}_0^T \mathbf{P}^{-1} \tilde{\mathbf{m}}_0}. \tag{36.11}$$

The height of the object is given by the length of $\mathbf{M}_0 \sim \mathbf{M}_\mathrm{h}$, and $\mathbf{M}_\mathrm{h}$ is the intersection of the vertical pole $\mathbf{L}_\mathrm{h}$ and the line $\mathbf{L}_2$ passing through $\mathbf{m}_\mathrm{h}$. The line $\mathbf{L}_2$ and the vertical pole $\mathbf{L}_\mathrm{h}$ can be denoted as follows.

$$\mathbf{L}_2 = -\mathbf{P}^{-1}\tilde{\mathbf{p}} + \lambda \mathbf{P}^{-1}\tilde{\mathbf{m}}_h = \mathbf{C} + \lambda \mathbf{D}_h, \; -\infty < \lambda < \infty, \tag{36.12}$$

$$\tilde{\mathbf{L}}_h = \tilde{\mathbf{M}}_0 + \mu \tilde{\mathbf{D}}_v, \; \tilde{\mathbf{D}}_v = [0\,0\,1\,0]^T, \; -\infty < \mu < \infty. \tag{36.13}$$

From $\mathbf{L}_\mathrm{h} = \mathbf{L}_2 = \mathbf{M}_\mathrm{h}$, we obtain

$$\mathbf{M}_0 + \mu \mathbf{D}_v = \mathbf{C} + \lambda \mathbf{D}_h. \tag{36.14}$$

We rearrange (36.14) so that a set of linear equations on λ and μ is given as follows.

$$\begin{bmatrix} m_1 - c_1 \\ m_2 - c_2 \\ m_3 - c_3 \end{bmatrix} = \begin{bmatrix} d_{h1} & -d_{v1} \\ d_{h2} & -d_{v2} \\ d_{h3} & -d_{v3} \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix}. \tag{36.15}$$

Here, m_i, c_i, d_{hi}, and d_{vi} represent the ith row's element of $\mathbf{M}_0$, $\mathbf{C}$, $\mathbf{D}_h$, and $\mathbf{D}_v$ respectively. Finally, we obtain the height and position from Eqs.36.12, 36.13.

36.2.4 Correction of Back-Projection Error

Inaccurate projective mapping, which is often caused by the inaccurate estimation of camera projection matrix, affects the 3D point estimation and consequently the

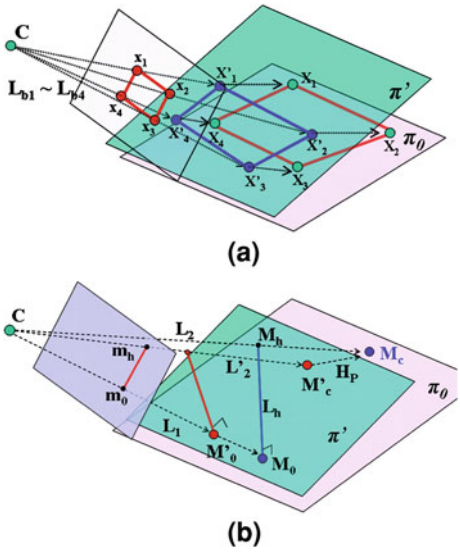


Fig. 36.7 Correction of geometric distortion using vertices of the marker

measurement results as well. Figure 36.6 shows an example of the back-projection error. Suppose that the camera is fixed and $\boldsymbol{\pi}_0$ is the ideal reference plane. In general, the detected plane $\boldsymbol{\pi}'$ does not coincide with $\boldsymbol{\pi}_0$ perfectly because of the back-projection error. Figure 36.6b is the side view of Fig. 36.6a, which illustrates that the measurements are significantly affected by perspective distortions. This problem is often solved by implementing nonlinear optimization algorithm such as the Levenberg–Marquardt iteration. However, there normally exists a significant trade-off between processing time and reliability of final result. In order to correct this perspective distortion, therefore, we employ four reference points on the rectangular marker, as illustrated in Fig. 36.7. Assuming that projective mapping is ideal, $\mathbf{x}_1 \sim \mathbf{x}_4$ is mapped to $\mathbf{X}_1 \sim \mathbf{X}_4$ of the ideal plane. In practice, however, the vertex images are back-projected onto $\mathbf{X}'_1 \sim \mathbf{X}'_4$ of $\boldsymbol{\pi}'$. From $\mathbf{X}'_1 \sim \mathbf{X}'_4$ and $\mathbf{X}_1 \sim \mathbf{X}_4$, we can estimate the homography which transforms the points of $\boldsymbol{\pi}'$ to those of $\boldsymbol{\pi}_0$. The measured position of the object can then be corrected simply by applying the homography. On the other hand, the height of the object cannot be corrected in this way because the intersection $\mathbf{M}_h$ is not in contact with the reference plane. Therefore, we rectify the measured height as follows.

1. Compute the intersection $\mathbf{M}_C{}'$ of $\mathbf{L}_2{}'$ and $\boldsymbol{\pi}'$ as follows.

$$\mathbf{M}'_C = \mathbf{P}^{-1}(-\tilde{\mathbf{p}} + \lambda_C \tilde{\mathbf{m}}_h), \text{ and } \lambda_C = \frac{\boldsymbol{\pi}_0^T \mathbf{C}}{\boldsymbol{\pi}_0^T \mathbf{P}^{-1} \tilde{\mathbf{m}}_h}. \tag{36.16}$$

2. Transform $\mathbf{M}_C{}'$ to $\mathbf{M}_C$ of $\boldsymbol{\pi}_0$ by applying the homography $\mathbf{H}_p$.

$$\tilde{\mathbf{M}}_C = \mathbf{H}_p \tilde{\mathbf{M}}'_C, \; \tilde{\mathbf{M}}_C = [\tilde{\mathbf{M}}_C 1]^{\mathrm{T}}, \tag{36.17}$$

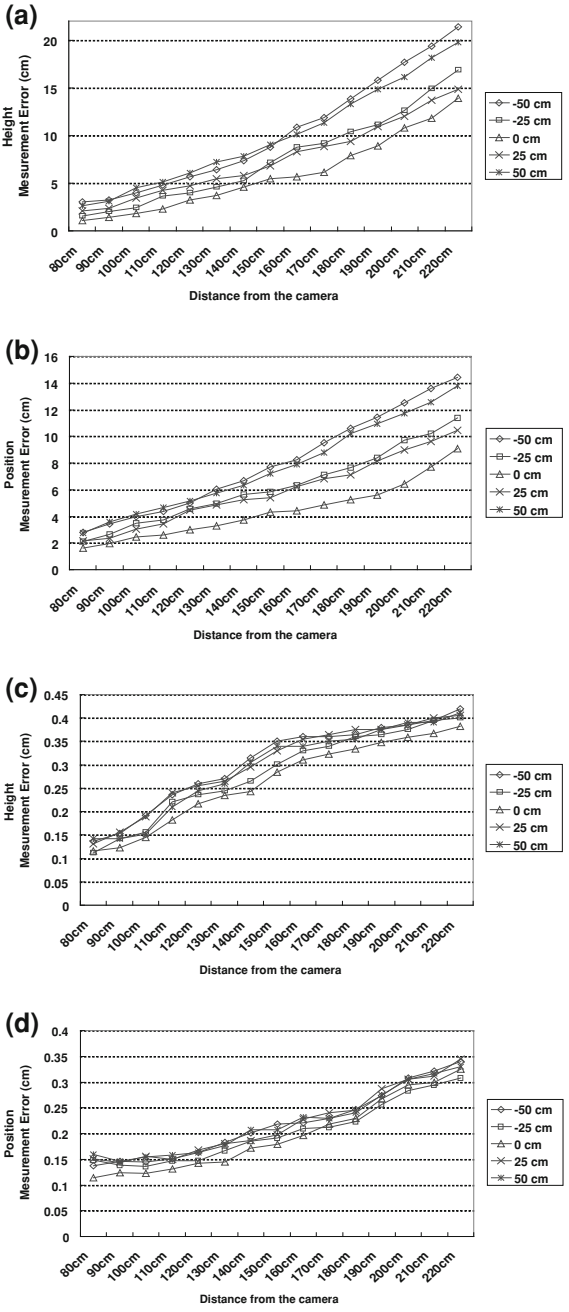


Fig. 36.8 Measurement errors: **a** and **b** height and position estimation errors before the distortion compensation; **c** and **d** the corresponding errors after the distortion compensation

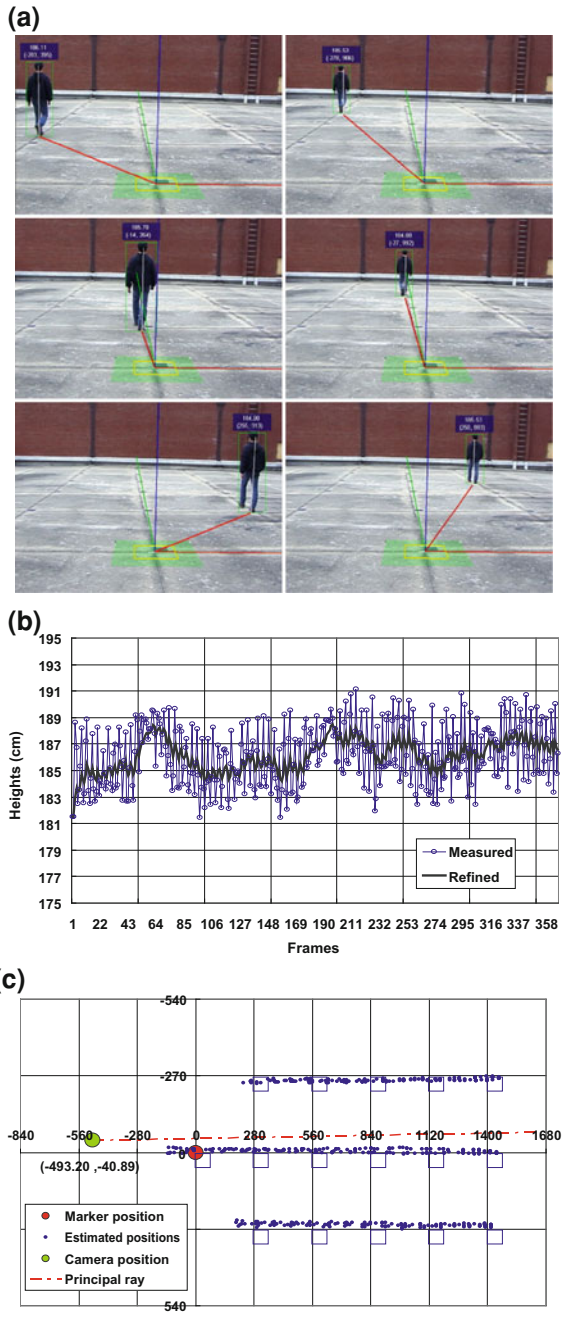


Fig. 36.9 Experiment #1: **a** input video stream; **b** estimated heights; and **c** bird's eye view which illustrates estimated positions

Table 36.1 Height estimation results

		Real Height (cm)	Mean (cm)	SD (cm)	Median (cm)
Experiment 1	Path 1	185.00	184.83	2.56	184.89
	Path 2		185.88	2.33	185.79
	Path 3		185.58	2.15	185.47
Experiment 2		168.00	170.08	3.08	169.68
Experiment 3		176.00	178.24	2.46	178.19

3. Finally, estimate $\mathbf{M}_h$ which is the intersection of the vertical pole $\mathbf{L}_h$ and $\mathbf{L}_2$ formed by the join of $\mathbf{C}$ and $\mathbf{M}_c$. Therefore height is obtained from $h = \| \mathbf{M}_h - \mathbf{M}_0 \|$.

36.3 Experimental Results

To evaluate the performance of the proposed method, a set of experiments are conducted. The first experiment is carried out under an ideal condition in the laboratory. And we validate the proposed method on outdoor video sequences. All experiments are performed with a CCD camera which produces 720 × 480 image sequences in 30 FPS. The first experiment is performed in the following way. In a uniform background, we locate and move a stick which has length of 30 cm. And then, at every 25 cm along the horizontal direction, and at every 10 cm from the camera, we measure its position and height. To give the reference coordinate, we used a square marker whose sides have length $w_m = 30$ cm. The measurement errors are shown in Fig. 36.8. Figure 36.8a and b illustrate that the measurements are significantly affected by the perspective distortion. However, Fig. 36.8c and d verify that the results are fairly improved by applying the distortion correction algorithm.

We note that the measurement errors grow as the distance in each direction is increased. Considering the dimension of the object and the distance from the camera, however, the measurement errors can be regarded as relatively small values. Therefore, our method achieves reliable estimation of the height and position without critical error. The second experiment is carried out using several outdoor video sequences. For the outdoor experiments, we preset an experimental environment. On every rectangular area of size 280 × 270 cm, we place a square landmark. During the experiment, a participant walks along preset paths, and the heights and positions are measured at each frame. The reference coordinate system is given by a square marker whose sides have length $w_m = 60$ cm. Figure 36.9a illustrate the input video streams, which also illustrate the measured height and position, the reference coordinate frame, and a vector pointing to the human. Figure 36.9b shows the measured heights at each frame. In general, human walking involves periodic up-and-down displacement. The maximum height

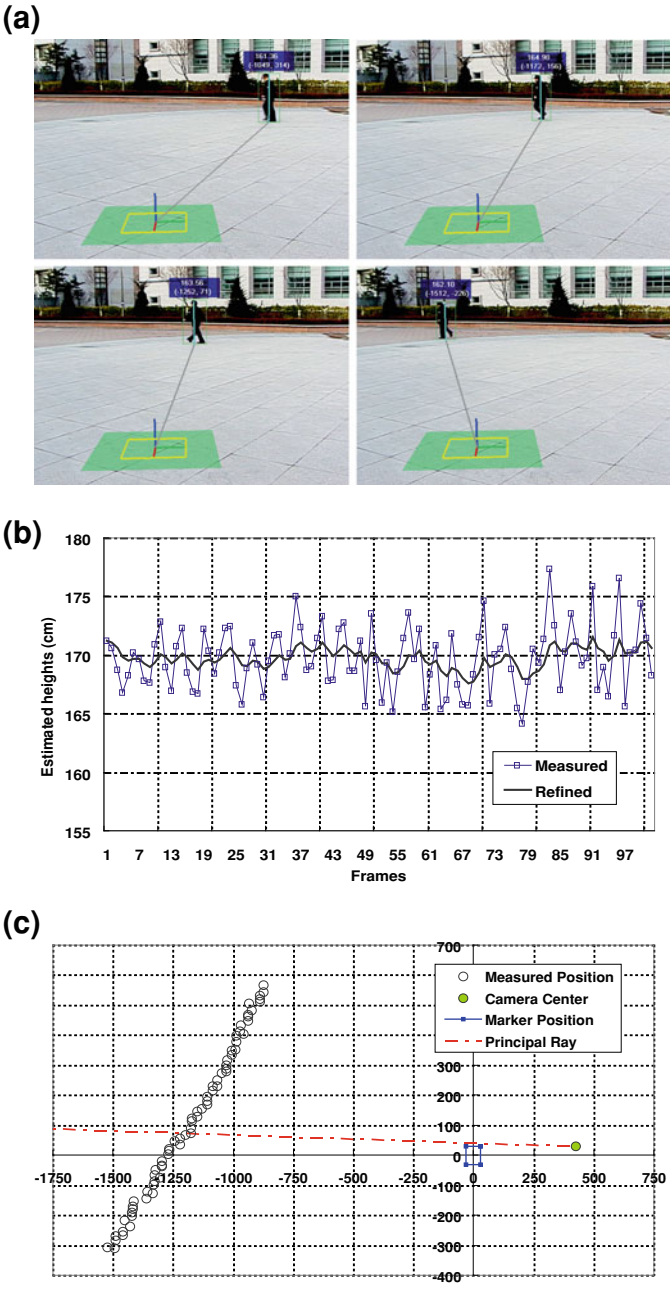
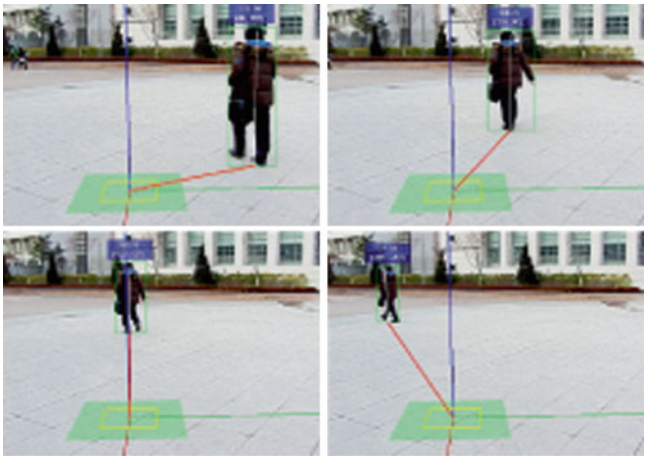
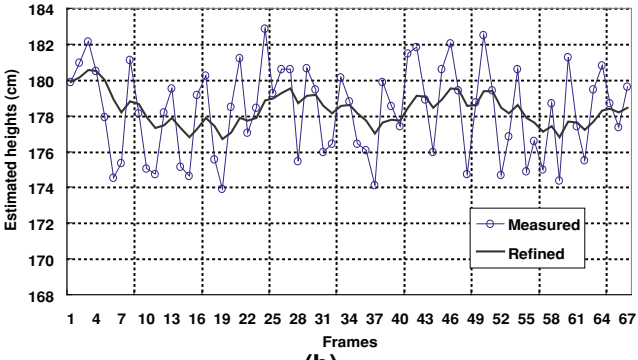


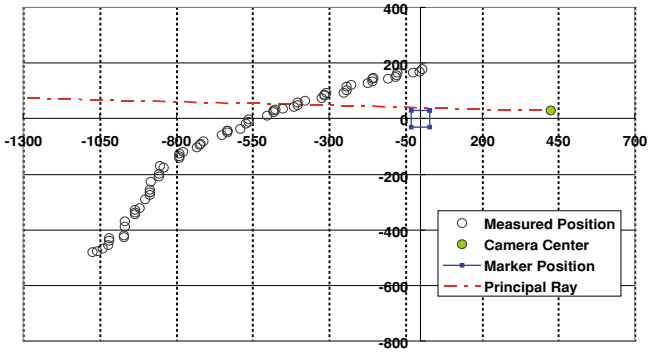
Fig. 36.10 Experiment #2: **a** input video stream; **b** height estimates; and **c** bird's eye view of **a** which illustrates measured positions



(a)



(b)



(c)

Fig. 36.11 Experiment #3: **a** input video stream; **b** height estimates; and **c** bird's eye view of **a** which illustrates measured positions

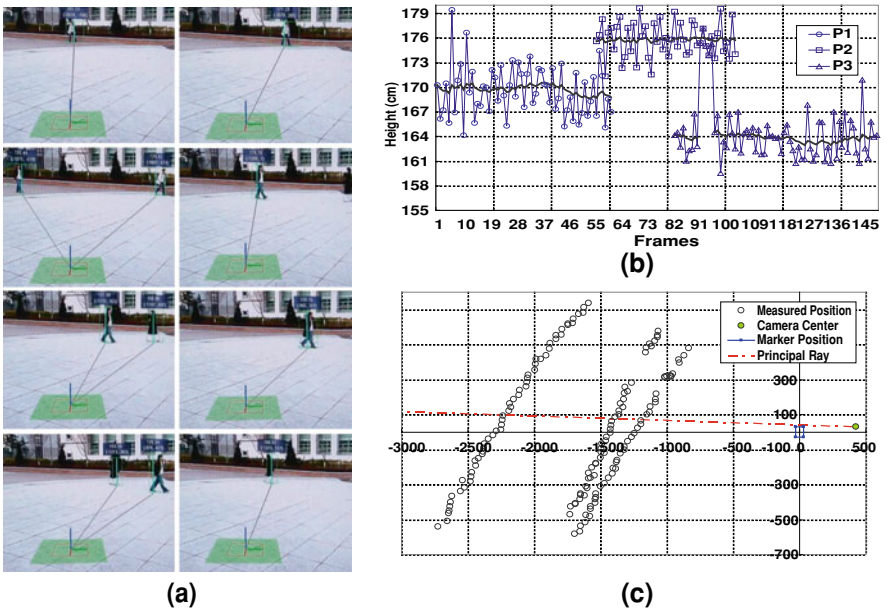


Fig. 36.12 Experiment #4: **a** input video stream **b** height estimates **c** bird's eye view of **a** which illustrates measured positions

occurs at the leg-crossing phase of walking, while the minimum occurs when the legs are furthest apart. Therefore we refine the results through moving average filters. As presented in Table 36.1, the height estimates are accurate to within $\sigma = 2.15 \sim 2.56$ cm. Figure 36.9c demonstrates a bird's eye view of the scene, which illustrates trajectory of the human, principal ray, position of the camera, and the position of the reference marker.

The trajectory which exactly coincides with the land marks clearly means that our method can recover the original position of the moving individual accurately. Similarly, Figs. 36.10 and 36.11 show the results on several outdoor scenes, which also confirm the accuracy and the robustness of the proposed method. Figure 36.12 demonstrates the experimental results of multiple targets. In this case, P3 is occluded by P2 between frame 92 and 98. As shown in Figs. 36.12b 36.12c, this occlusion may affect the estimates of P2 and P3. This problem can, however, be avoided by using a prediction algorithm, and we hope to report on this in the near future. The processing speed of the proposed method is roughly 12 frames/seconds, but this may be dependent on image quality and number of targets in the scene. In summary, the experimental results suggest that the proposed method allows accurate estimation of the trajectories and height.

36.4 Conclusion

We presented a single view-based framework for robust and real-time estimation of human height and position. In the proposed method, a human body is assumed to be a vertical pole. And 2D features of the imaged object are back-projected into the real-world scene to compute the height and position of the moving object.

To give the reference coordinate frame, a reference marker with a rectangular pattern is used. In addition, a refinement approach is employed to correct the estimated result by using the geometric constraints of the marker. The accuracy and robustness of our technique are verified on the experimental results of several real video sequences from outdoor environments. We believe that the proposed framework can be used as a useful tool in the monocular camera-based surveillance systems. As a future work, we will conduct further study to apply a tracking algorithm to the proposed framework in order to develop a robust multi-target tracking system.

References

1. Benabdelkader C, Cutler R, Davis L (2002) Person identification using automatic height and stride estimation. In: Proceedings European conference computer vision, 155–158, June 2002
2. Havasi L, Szlávik Z, Szirányi T (2007) Detection of gait characteristics for scene registration in video surveillance system. IEEE Trans Image Process 16(2):503–510
3. Liu Z, Sarkar S (2006) Improved gait recognition by gait dynamics normalization. IEEE Trans Pattern Anal Mach Intell 28(6):863–876
4. Leibowitz D, Criminisi A, Zisserman A (1999) Creating architectural models from images. Proc EuroGraphics'99, 18(3) Sep
5. Criminisi A, Reid I, Zisserman A (2000) Single view metrology. Int J Comput Vision 40(2):123–148
6. Criminisi A (2002) Single-view metrology: algorithms and application. In: Proceedings the 24th DAGM symposium on pattern recognition
7. Lee L, Romano R, Stein G (2000) Monitoring activities from multiple video streams: establishing a common coordinate frame. IEEE Trans Pattern Anal Mach Intell 22(8):758–769 Aug
8. Hu W, Hu M, Zhou X, Tan T, Lou J, Maybank S (2000) Principal axis-based correspondence between multiple cameras for people tracking. IEEE Trans Pattern Anal Mach Intell 28(4):663–671 Apr
9. K. Kim and L. Davis, "Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering," Proc. European Conf. Computer Vision, Part III, pp. 98-109, May 2006
10. Khan S, Shah M (2006) A multiple view approach to tracking people in crowded scenes using a planar homography constraint. In: Proceedings European conference computer vision, Part IV, 133–146, May 2006
11. Khan S, Shah M (2003) Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Trans Pattern Anal Mach Intell 25(10):1355–1361 Oct
12. Lee SH, Lee SK, Choi JS (2009) Correction of radial distortion using a planar checkerboard pattern and its image. IEEE Trans Consum Electron 55(1):27–33 Feb

13. Elgammel A, Harwood D, Davis L (2000) Non-parametric model for back ground subtraction. In: Proceedings European conference computer vision, Part II, 751–767, Jun 2000
14. Zhang Z (2000) Flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 19(7):1330–1334 Nov
15. Golub G, Loan C (1996) Matrix computations, 3rd edn. Johns Hopkins Univ Press, Baltimore
16. Faugeras O (1993) Three-dimensional computer vision. MIT Press, Cambridge
17. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge University Press, Cambridge
18. Criminisi A (2001) Accurate visual metrology from single and multiple uncalibrated images. Springer, Berlin
19. Hu W, Tan T, Wang L, Maybank S (2004) A survey on visual surveillance of object motion and behaviors. IEEE Trans Pattern Anal Mach Intell 34(3):334–353 Aug
20. Haritaoglu I, Harwood D, Davis L (2000) W4 Real-time surveillance of people. IEEE Trans Pattern Anal Mach Intell 22(8):809–830 Aug
21. Mckenna S, Jabri S, Duric J, Wechsler H, Rosenfeld A (2000) Tracking groups of people. Comput Vision Image Understand 80:42–56
22. Liang B, Chen Z, Pears N (2004) Uncalibrated two-view metrology.In: Proceedings international conference pattern recognition, vol. 1. 96–99, Aug 2004