Height and Position Estimation of Moving Objects using a Single Camera

Seok-Han Lee, Jae-Young Lee, Bu-Gyeom Kim, and Jong-Soo Choi

Graduate School of Advanced Imaging Science, Multimedia, and Film,
Chung-Ang University
Seoul, Korea
E-mail: {ichthus, evs0, kbs211, jschoi}@imagelab.cau.ac.kr

ABSTRACT

In recent years, there has been increased interest in characterizing and extracting 3D information from 2D images for human tracking and identification. In this paper, we propose a single view-based framework for robust estimation of height and position. In the proposed method, 2D features of target object is back-projected into the 3D scene space where its coordinate system is given by a rectangular marker. Then the position and the height are estimated in the 3D space. In addition, geometric error caused by inaccurate projective mapping is corrected by using geometric constraints provided by the marker. The accuracy and the robustness of our technique are verified on the experimental results of several real video sequences from outdoor environments.

Keywords: Human tracking, Height and position estimation, Visual metrology.

1. INTRODUCTION

Vision-based human tracking is steadily gaining in importance due to the drive from many applications, such as smart video surveillance, human-machine interfaces, and ubiquitous computing. In recent years, there has been increased interest in characterizing and extracting 3D information from 2D images for human tracking. Emergent features are height, gait(an individual's walking style), and trajectory in 3D space. Because they can be measured at a distance, and from coarse images, considerable research efforts have been devoted to use them for human identification and tracking. An important application is in forensic science, to measure dimensions of objects and people in images taken by surveillance cameras [1, 2]. Because bad quality of the image (taken by cheap security camera), quite often it is not possible to recognize the face of the suspect 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 upon 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 this paper, we propose a single view-based technique for the estimation of human height and position. In our method, 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 location 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. Finally, we apply a refinement approach to correct the estimated result by using geometric constraints provided by the marker. The proposed method allows real-time acquisition of the position of a moving object as well as the height in 3D space. Moreover, as the projective camera mapping is estimated by using the marker, our method is applicable even in the absence of geometric cues. The remainder of this paper is structured in the following way: In Section 2, the proposed method is discussed in Section 3, and experimental results are given in Section 4. The conclusions are drawn in Section 5.

2. PROPOSED METHOD

2.1 Foreground Blob Extraction

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 world space. Therefore, 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 in the image is detected. Then we compute intersections of the vertical principal axis and the bounding box. 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. 2, 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 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.

2.2 Back-projection.

In our method, the height and position are measured by using the back-projected features in three-dimensional scene. Let $\tilde{\mathbf{M}} = [X\ Y\ Z\ 1]^{\mathrm{T}}$ be the 3D homogeneous

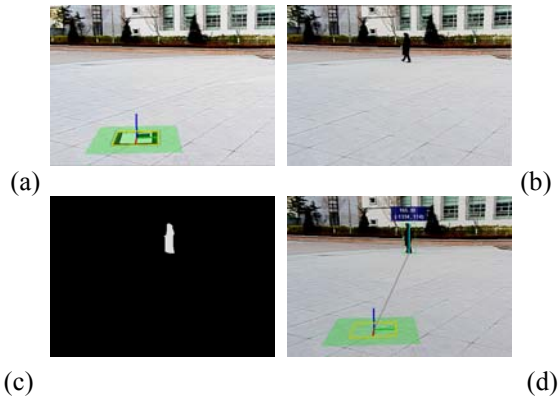


Fig. 1. An example of the procedure (a) Estimation of the projective camera matrix using a marker (b) Real-time input image (c) Extraction of the object (d) Final result

coordinates of a world point and $\tilde{\mathbf{m}} = [x\ y\ 1]^T$ be the 2D homogeneous coordinates of its projection in the image plane. This 2D-3D mapping is defined by a linear projective transformation as follows.

$$\tilde{\mathbf{m}} = \lambda\tilde{\mathbf{P}}\tilde{\mathbf{M}} = \lambda\mathbf{K}[\mathbf{R}\,|\,\mathbf{t}]\tilde{\mathbf{M}} = \lambda\mathbf{K}[\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3\,|\,\mathbf{t}]\tilde{\mathbf{M}}, \quad (1)$$

where λ is an arbitrary scale factor, and the 3 x 4 matrix $\tilde{\mathbf{P}}$ is the projective camera matrix, which represents the projection of 3D scene space onto a 2D image. **R** is a 3 x 3 rotation matrix, and **t** denotes translation vector of the camera. And $\mathbf{r}_i$ means *i*-th column vector of the projection matrix. We use '~' notation to denote the homogeneous coordinate representation. The non-singular matrix **K** represents the camera calibration matrix, which consists of the intrinsic parameters of the camera. In our method, we employ the calibration method proposed by Zhang in [6]. This method computes the IAC (the image of absolute conic) ω by using the invariance of the circular points which are the intersections of a circle and the line at infinity $\mathbf{l}_\infty$. Once the IAC ω is computed, the calibration matrix can be **K** computed by $\omega^{-1}=\mathbf{K}\mathbf{K}^T$. Thus this method requires at least three images of a planar calibration pattern observed at three different orientations. From the calibrated camera matrix **K** and (1), the projective transformation between 3D scene and its image can be determined. In particular, the projective transformation between a plane of 3D scene and the image plane can be defined by a general 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 π_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 $\tilde{\mathbf{M}}_p=[\tilde{\mathbf{X}}_1\ \tilde{\mathbf{X}}_2\ \tilde{\mathbf{X}}_3\ \tilde{\mathbf{X}}_4]$ and $\tilde{\mathbf{m}}_p=[\tilde{\mathbf{x}}_1\ \tilde{\mathbf{x}}_2\ \tilde{\mathbf{x}}_3\ \tilde{\mathbf{x}}_4]$ which consist of $\tilde{\mathbf{X}}_n=[X_n\ Y_n\ 0\ 1]^T$ and $\tilde{\mathbf{x}}_n=[x_n\ y_n\ 1]^T$ respectively is given by

$$\tilde{\mathbf{m}}_p = \mathbf{K}[\mathbf{R}\,|\,\mathbf{t}]\tilde{\mathbf{M}}_p = [\mathbf{p}_1\ \mathbf{p}_2\ \mathbf{p}_3\ \mathbf{p}_4]\tilde{\mathbf{M}}_p. \quad (2)$$

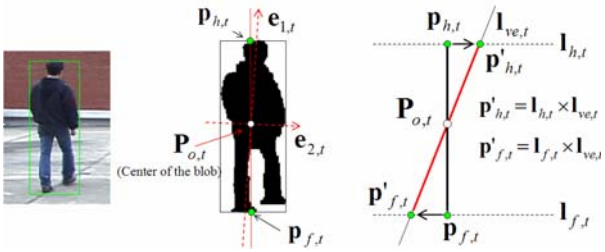


Fig. 2. Extraction of head and feet locations (a) Captured image (b) Estimation of principal axis using eigenvectors (c) Extraction of the head and feet points

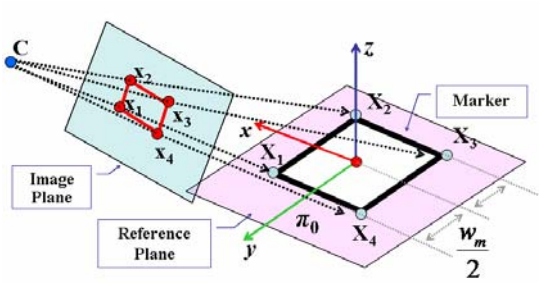


Fig. 3. Projective mapping between the marker and its image

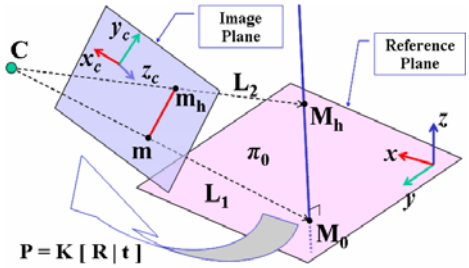


Fig. 4. Back-projection of 2D features

Here, $\mathbf{p}_i$ is *i*-th 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 (2), we have

$$\mathbf{K}^{-1}\begin{bmatrix} x_n \\ y_n \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11}X_n + r_{12}Y_n + t_x \\ r_{21}X_n + r_{22}Y_n + t_y \\ r_{31}X_n + r_{32}Y_n + t_z \end{bmatrix}, \quad (3)$$

where (x_n, y_n) is *n*-th vertex detected from the image. From (3) and the four vertices, we obtain the translation vector t and the elements of the rotation matrix $\mathbf{r}_{ij}$. By the property of the rotation matrix, the third column of **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 (4), 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. 3.

$$\tilde{\mathbf{M}}_p = \begin{bmatrix} w_m/2 & w_m/2 & -w_m/2 & -w_m/2 \\ w_m/2 & -w_m/2 & -w_m/2 & w_m/2 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \quad (4)$$

In general, the computed rotation matrix **R** does not satisfy with the properties of a rotation matrix. Let the singular value decomposition of **R** be $\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$, where $\Sigma = diag(\sigma_1, \sigma_2, \sigma_3)$. Since a pure rotation matrix has $\Sigma = diag(1, 1, 1)$, we set $\mathbf{R} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ which is the best approximation matrix to the estimated rotation matrix. 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. 4. Given the camera projection matrix $\tilde{\mathbf{P}} = [\mathbf{P} \ \tilde{\mathbf{p}}]$, where **P** is a 3 x 3 submatrix, the camera center is denoted by $\mathbf{C} = -\mathbf{P}^{-1}\tilde{\mathbf{p}}$. And the direction of the line **L** which formed by the join of **C** and m can be determined by its point at infinity $\tilde{\mathbf{D}}$ as follows

$$\tilde{\mathbf{P}}\tilde{\mathbf{D}} = \tilde{\mathbf{m}}, \tilde{\mathbf{D}} = [\mathbf{D} \ 0]^{\mathrm{T}}, \quad (5)$$

$$\mathbf{D} = \mathbf{P}^{-1}\tilde{\mathbf{m}}, \tilde{\mathbf{m}} = [\mathbf{m}^{\mathrm{T}} \ 1]^{\mathrm{T}}. \quad (6)$$

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

$$\mathbf{L} = -\mathbf{P}^{-1}\tilde{\mathbf{p}} + \lambda\mathbf{P}^{-1}\tilde{\mathbf{m}} = \mathbf{C} + \lambda\mathbf{D}, \ -\infty < \lambda < \infty. \quad (7)$$

2.3 Estimation of Height and Position

In our method, a human body is approximated as a vertical pole. As shown in Fig. 4, the height of the object is the distance between $\mathbf{M}_0$ and $\mathbf{M}_h$, and its position is $\mathbf{M}_0$ which is the intersection of the reference plane π_0 and the line $\mathbf{L}_1$. Assuming that the line segment $\mathbf{M}_0 \sim \mathbf{M}_h$ is mapped onto its image $\mathbf{m}_0 \sim \mathbf{m}_h$, the intersection is 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. As $\mathbf{M}_0$ is always located on the reference plane π_0, we have

$$\tilde{\boldsymbol{\pi}}_0^{\mathrm{T}}\tilde{\mathbf{M}}_0 = 0, \tilde{\boldsymbol{\pi}}_0 = [0\ 0\ 1\ 0]^{\mathrm{T}}, \tilde{\mathbf{M}}_0 = [\mathbf{M}_0\ 1]^{\mathrm{T}}. \quad (8)$$

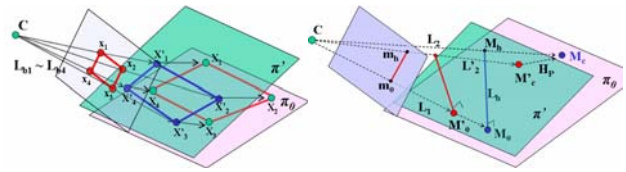
Then, from $\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{M}_0 = \boldsymbol{\pi}_0^{\mathrm{T}}(\mathbf{C} + \lambda_0 \mathbf{P}^{-1}\tilde{\mathbf{m}}_0)$, we can uniquely determine λ_0 as follows

$$\lambda_0 = -\frac{\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{C}}{\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{P}^{-1}\tilde{\mathbf{m}}_0}. \quad (9)$$

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

$$\mathbf{L}_2 = -\mathbf{P}^{-1}\tilde{\mathbf{p}} + \lambda\mathbf{P}^{-1}\tilde{\mathbf{m}}_h = \mathbf{C} + \lambda\mathbf{D}_h, \ -\infty < \lambda < \infty, \quad (10)$$

$$\tilde{\mathbf{L}}_h = \tilde{\mathbf{M}}_0 + \mu\tilde{\mathbf{D}}_v, \tilde{\mathbf{D}}_v = [0\ 0\ 1\ 0]^{\mathrm{T}}, -\infty < \mu < \infty. \quad (11)$$



(a) (b)

Fig. 5. Correction of geometric distortion using vertices of the marker

From $\mathbf{L}_h = \mathbf{L}_2 = \mathbf{M}_h$, we obtain

$$\mathbf{M}_0 + \mu\mathbf{D}_v = \mathbf{C} + \lambda\mathbf{P}^{-1}\tilde{\mathbf{m}}. \quad (12)$$

We rearrange (12), so that a set of linear equations on λ and μ is given as follows

$$\begin{bmatrix} m_1 - c_1 \\ m_2 - c_2 \\ m_3 - c_3 \end{bmatrix} = \begin{bmatrix} d_{h1} & -d_{v1} \\ d_{h2} & -d_{v2} \\ d_{h3} & -d_{v3} \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix}. \quad (13)$$

Without difficulty, this can be solved via simple linear-squared estimation. Finally, from (10) and (11), we obtain the height and position. Inaccurate projective mapping often affects the estimation of 3D points and consequently the measurement results. This problem is often solved by implementing nonlinear optimization algorithm such as the Levenberg -Marquardt iteration. However, there normally exist a significant trade-off between processing time and the reliability of the result. In order to correct this perspective distortion, therefore, we use the four vertices of the square marker as shown in Fig. 5. Assuming that the projective mapping is ideal, x1 ~ x4 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 π'. From $\mathbf{X}'_1 \sim \mathbf{X}'_4$ and $\mathbf{X}_1 \sim \mathbf{X}_4$, we can estimate the homography which trans-forms the points of π' to those of π_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 can not 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 π' as follows

$$\mathbf{M}'_C = \mathbf{P}^{-1}(-\tilde{\mathbf{p}} + \lambda_C\tilde{\mathbf{m}}_h), \lambda_C = \frac{\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{C}}{\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{P}^{-1}\tilde{\mathbf{m}}_h}. \quad (14)$$

2) Transform $\mathbf{M}_C'$ to $\mathbf{M}_C$ of π_0 by applying the homography

$$\tilde{\mathbf{M}}_C = \mathbf{H}_p\tilde{\mathbf{M}}'_C, \tilde{\mathbf{M}}_C = [\tilde{\mathbf{M}}_C \ 1]^{\mathrm{T}}, \quad (15)$$

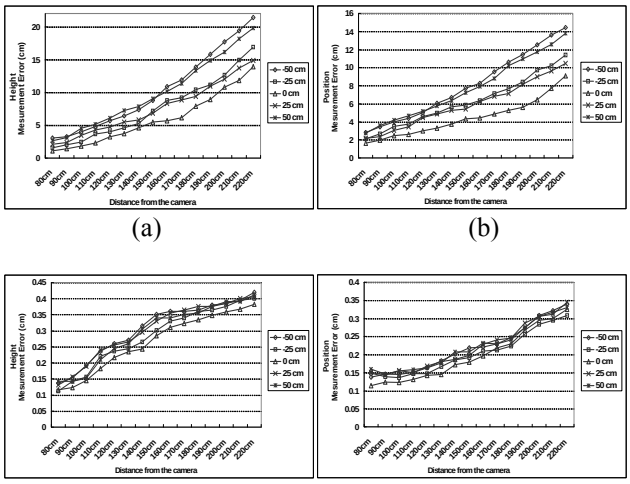


Fig. 6. Measurement errors (a), (b) Height and position estimation errors before the distortion compensation (c), (d) After the distortion compensation

where $\mathbf{H}_p$ denotes the homography defined by the quadruple point pairs.

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 $\mathbf{M}_c$. The height is obtained from $\mathbf{h} = \| \mathbf{M}_h - \mathbf{M}_0 \|$.

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, two sets of experiments are conducted. The first experiment is carried out under ideal condition in laboratory. And we validate the proposed method on outdoor videos sequences. All experiments are performed with a CCD camera which produces 720 x 480 image sequences in 30 FPS. The first experiment is performed in following way. In a uniform background, we locate and move a rod which has length 30cm. And then, at every 25cm along horizontal direction and at every 10cm from the camera, we measure its position and height. To give the reference coordinate, we used a square marker whose sides have length wm = 30cm. The measurement errors are shown in Fig. 6. Fig. 6(a) and Fig. 6(b) illustrate that the results are affected significantly by the perspective distortion. From Fig. 6(c) and Fig. 6(d), however, we verify that the measurements are fairly improved by applying the correction algorithm. We note that the measurement error grows 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. Therefore, we can conclude that our method achieves reliable estimation of the height and position without critical error. The second experiment is carried out using several outdoor videos sequences. For the outdoor experiments, we preset an experimental environment. On every rectangular area which has dimension of 280cm x 270cm, we place a recognizable landmark. During the experiment, a participant walks along preset paths, and the

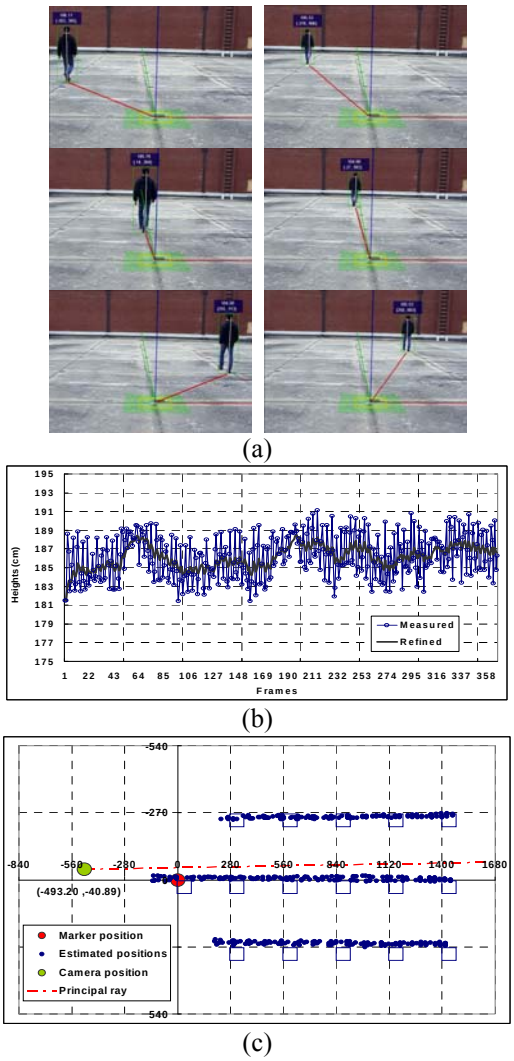
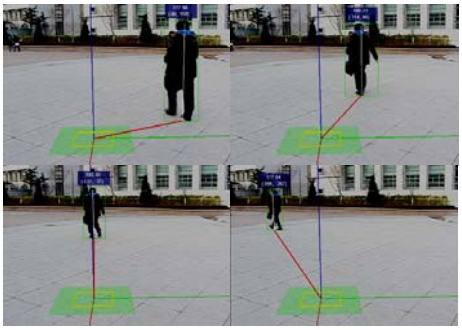
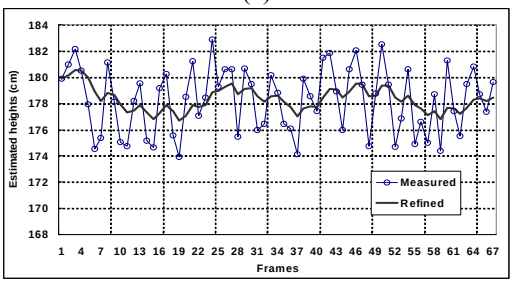


Fig. 7. Experiment #1 (a) Input video stream (b) Estimated heights (c) Bird's eye view which illustrates estimated positions

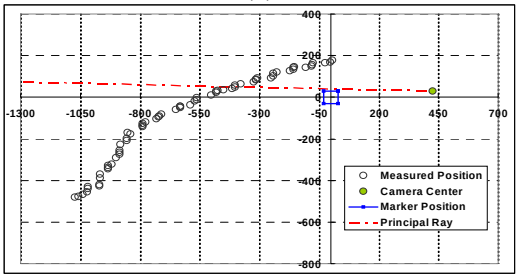
height and position are measured at each frame. The reference coordinate system is given by a square marker whose sides have length wm = 60cm. Fig. 7(a) illustrate the input video streams, which also show the measured height and position, the reference coordinate frame, and a vector pointing to the individual. Fig. 7(b) shows the measured heights at each frame. In general, human walking involves periodic up-and -down displacement. The maximum height occurs at the leg-crossing phase of walking, while the minimum occurs when the legs are furthest apart. Therefore we refine the results through running average filter. As shown in Table 1, the height estimate is accurate to within σ =2.15 ~ 2.56cm. Fig. 7(c) shows a bird's eye view of the scene, which illustrates trajectory of the individual, principal ray, position of the camera, and position of the 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. Fig. 8 and Fig. 9 show results on several outdoor scenes, which also confirm the accuracy and the robustness of the proposed method. Fig. 9 demonstrates experimental results of multiple targets. In this case, P3 is occluded by P2 between frame 92 and 98.



(a)



(b)



(c)

Fig. 8. Experiment #3 (a) Input video stream (b) Height estimates (c) Bird's eye view of (a) which illustrates measured positions

Table 1. Height estimation results

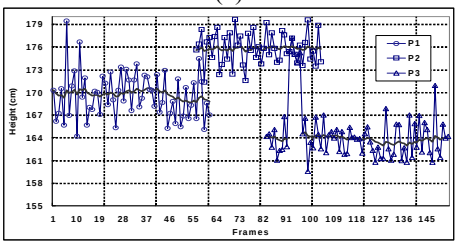
		Real Height (cm)	Mean (cm)	Std. Dev. (cm)	Median (cm)
Experiment 1	Path 1	185.00	184.83	2.56	184.89
	Path 2		185.88	2.33	185.79
	Path 3		185.58	2.15	185.47
Experiment 2		168.00	170.08	3.08	169.65
Experiment 3		176.00	178.24	2.46	178.19

As shown in Fig. 9(b) and Fig. 9(c), 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 12frames/sec., 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 recovering the motion trajectories and height with high accuracy.

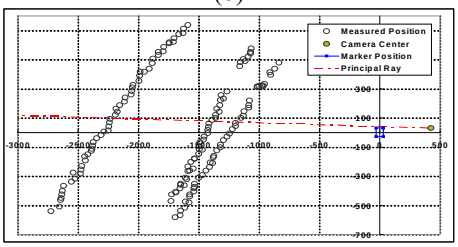
4. CONCLUSION

We have presented a single view-based framework for robust and real-time estimation of human height and position. In the proposed method, human body is assumed to be a vertical pole. And the 2D features of the imaged

(a)



(b)



(c)

Fig. 9. Experiment #4 (a) Input video stream (b) Height estimates (c) Bird's eye view of (a) which illustrates measured positions

object are back-projected into the real-world scene to compute the height and location of the moving object. To give the reference coordinate frame, a rectangular marker 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 was verified on the experimental results of several real video sequences from outdoor environments. The proposed method is applicable to surveillance/security systems which employ a simple monocular camera.

Acknowledgment. This work was supported by Korean Research Foundation under BK21 project and SFCC Cluster established by Seoul R&BD Program.

5. REFERENCES

[1]Leibowitz, D., Criminisi, A., Zisserman, A.: Creating Architectural Models from Images. Proc. EuroGraphics'99. 18 (1999) 39-50
[2] A. Criminisi, I. Reid, and A. Zisserman, "Single View Metrology," Int'l J. Computer Vision, vol. 40, no. 2, pp. 123-148, 2000.
[3]Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge Univ. Press (2003)
[4]O. Faugeras, "Three-Dimensional Computer Vision," The MIT Press, 1993.
[5]Criminisi, A.: Accurate Visual Metrology from Single and Multiple uncalibrated Images. Springer-Verlag (2001)
[6] Zhang, Z.: Flexible New Technique for Camera Calibration. IEEE Trans. Pattern Analysis and Machine Intelligence. 19 (2000) 1330-1334