보행자 기반의 변분 베이지안 감시 카메라 자가 보정

임종빈*

Pedestrian-Based Variational Bayesian Self-Calibration of Surveillance Cameras

Jong-Bin Yim*

*Principal Research Engineer, Sequin Department, ITXM2M Co., LTD, Seoul, 08502 Korea

요 약

보행자 기반의 카메라 자가 보정 방법들은 복잡한 보정 장치나 절차가 필요하지 않기 때문에 비디오 감시 시스템에 적합하다. 하지만 일의 보행자들을 보정 대상으로 사용하는 경우 보행자들의 기름 모르기 때문에 보정 정확도가 저하될 수 있다. 본 논문은 실제 감시 환경에서 이 문제를 해결하기 위한 베이지안 보정 방법을 제안한다. 제안하는 방법에서는 감시 지역 사람들이 기에 대한 통계가 있다고 가정하고, 발-머리 호몰로지(foot-head homology)를 사용하여, 발과 머리의 좌표와 보행자 기의 불확실성을 모두 고려하는 확률 모델을 구성한다. 이 확률 모델을 직접 도는 것은 난해하므로, 본 연구에서는 근사적 방법인 변분 베이지안 추론(variational Bayesian inference)을 사용한다. 따라서, 이를 통해 관측된 보행자들의 기를 추정함과 동시에 정확한 카메라 파라미터를 구할 수 있다. 다양한 실험을 통해 제안된 방법이 노이즈에 강하며, 보정에 대한 정확한 신뢰도를 제공함을 보였다.

ABSTRACT

Pedestrian-based camera self-calibration methods are suitable for video surveillance systems since they do not require complex calibration devices or procedures. However, using arbitrary pedestrians as calibration targets may result in poor calibration accuracy due to the unknown height of each pedestrian. To solve this problem in the real surveillance environments, this paper proposes a novel Bayesian approach. By assuming known statistics on the height of pedestrians, we construct a probabilistic model that takes into account uncertainties in both the foot/head locations and the pedestrian heights, using foot-head homology. Since solving the model directly is infeasible, we use variational Bayesian inference, an approximate inference algorithm. Accordingly, this makes it possible to estimate the height of pedestrians and to obtain accurate camera parameters simultaneously. Experimental results show that the proposed algorithm is robust to noise and provides accurate confidence in the calibration.

키워드 : 카메라 자가 보정, 호몰로지, 베이지안 추정, 변분 추론

Keywords : Camera self-calibration, Planar homology, Bayesian estimation, Variational inference

Received 9 July 2019, Revised 15 July 2019, Accepted 19 July 2019

* Corresponding Author Jong-Bin Yim(E-mail:gameboy@postech.ac.kr, Tel:+82-2-2082-8500)
Principal Research Engineer, Sequin Department, ITXM2M Co., LTD, Seoul, 08502 Korea

http://doi.org/10.6109/jkiice.2019.23.9.1060

Copyright © The Korea Institute of Information and Communication Engineering.
I. Introduction

Camera calibration is essential to obtaining 3D metric information from observed images in video surveillance systems. Knowledge of the camera parameters is indispensable to many computer vision tasks. Also it helps to improve performance of tracking or detection [1]. Among the various calibration techniques, pedestrian-based self-calibration methods are attractive to most stationary surveillance cameras, since they do not require complex apparatuses or procedures, moreover, the cameras mainly monitor pedestrians.

Several pedestrian-based self-calibration methods have been proposed [1-7]. In general, these methods use a human as a calibration target under the assumption that humans are perpendicular to the ground plane. The methods usually utilize the vanishing points, horizon, or homologies, and presume that the humans are the same height. Krahnstoever and Mendonça [2] proposed a Bayesian method to take account of observation noise on the image plane. Liu et al. [3] proposed a method to utilize the pedestrian height distribution for challenging environments. Brouwers et al. [4] proposed a method to use accurate head/feet detectors, and delicate pre- and post-processes. Huang et al. [5] proposed a method to recover vanishing points from periodic motion of a pedestrian to compute the intrinsic parameters.

This paper proposes a Bayesian framework for pedestrian-based camera calibration taking account of the uncertainties in the pedestrian heights as well as in the image locations, using foot-head homology. Since the observations of the head/foot locations contain noises, we construct a Bayesian framework similar to [2] to take account of uncertainties on the image plane. For pedestrian-based self-calibration methods, the height of each pedestrian should be known for accuracy since the height is the only 3D metric information. The height of each pedestrian, however, is unknown if we observe arbitrary pedestrians under uncontrolled calibration procedures. Such uncertainty can be resolved by modeling each height as a random variable, unlike [3]. Furthermore, we can precisely estimate the camera parameters as well as each of heights, if we have statistics on the height and enough observations of pedestrians, since the sample distribution of the height converges to the true distribution as the number of observations increases.

Finding the posterior distribution of the camera parameters, given observations and prior distribution of the camera parameters requires the marginalization over the latent variables, which represent the uncertainties. Sampling method can be used when considering uncertainties only in the image plane [2]. Since we also consider the uncertainties in the pedestrian heights, the sampling method is infeasible due to the large space of latent variables. For this reason, we use variational inference [8], an approximate inference algorithm.

To summarize, we propose a pedestrian-based variational Bayesian self-calibration algorithm taking account of the uncertainties both in the image locations and the pedestrian heights. Initially the proposed algorithm computes the camera parameters by using the foot-head homology and constraints, and subsequently variational Bayesian inference refines the camera parameters. As a result, our algorithm gives not only accurate results but also computational efficiencies. We rigorously test the proposed algorithm both on real data and synthesized data, to show the accurate estimation of the camera parameters and their confidences.

II. Background

2.1. Camera Geometry

Under the assumption of zero skew, unit aspect ratio, and the principal point at the image center, the focal length $f$ is the only intrinsic parameter to be estimated. Accordingly, the camera calibration matrix $K$ is given by $K=\text{diag}(f, f, 1)$). Without loss of generality, as illustrated in Fig. 1, we arbitrarily set the origin of the world coordinate system $O$, to the orthogonal projection of the camera center $C$ onto the ground plane so that $C = [0, 0, -h]^T$, where $h$ is the height of camera. We also assume
zero pan angle so that the y-axis of the world coordinate system and the z-axis of the camera coordinate system are coplanar. Therefore, the rotation matrix $R$ is defined by the tilt $\theta$ around x-axis and the roll $\rho$ around z-axis as $R = R_xR_z$. The camera projection matrix $P$ is given by $P = [p_1, p_2, p_3] = KR$, where $t=RC$ is the translation vector. Note that $\theta$ ranges from $-\pi/2$ to 0 in our setting. Additionally, we define $\theta_i = \theta + \pi/2 (0 < \theta_i < \pi/2)$ for simplicity. Hereafter, for a collection of the camera parameters we define $m$ as

$$m = [h, \theta, \rho, f]^T. \quad (1)$$

![Fig. 1 Geometric setting of camera](image)

2.2, Foot-head Homology

If we assume that pedestrians have the same height $l$, all head locations are on a single plane. Also, since the ground plane and head plane are parallel to each other, the homography between the two planes is given by a planar homology [9], which has five degrees of freedom, three for the horizon line $l$ and two for the vertical vanishing point $v$. Then the homology $H$ is parameterized as $H=I+lg$, where $I$ is the identity matrix and $G=-\frac{1}{l}v^T$ which depends only on $m$.

Obviously, $G$ is a rank one matrix and has single non-zero singular value $s_0$, and its left and right singular vectors associated with $s_0$ are $v$ and $I$, respectively.

The vanishing points in $x$, $y$, and $z$ directions are $p_x$, $p_y$, and $p_z$, respectively [9]. Thus, the vertical vanishing point $v$ corresponds to $p_0$, and the horizon line $l$ is given by $p_x+p_y$ since the vanishing points in $x$ and $y$ directions are both on the horizon line. By using our projection model, we obtain the representations of $v$ and $l$ in terms of the camera parameters as

$$\hat{v} = [f\sin\phi\sin\theta, -f\cos\phi\sin\theta, \cos\phi]^T, \quad \hat{l} = [\sin\phi\theta, -\cos\phi\sin\theta, f\cos\phi]^T. \quad (2)$$

Therefore, the camera parameters denoted by $m$ are easily computed from the estimated homology $H$ if the height of pedestrians $l$ is known.

However, in reality, the heights of pedestrians are all different and their homologies are also different since each pedestrian has its own head plane. In this case, we parameterize the homology of each person $H_n$ as

$$H_n = I + l_nG, \quad (3)$$

where $l_n$ is height of the $n$-th pedestrian. The homology $H_n$ maps a foot location $z_n=[u_n,v_n,1]^T$ to a head location $z'_n=[u_n,v_n,1]^T$.

2.3, Initial Camera Parameter Estimation

We assume that pedestrians have the same height $l$, which is known in advance, and compute the matrix $G$ by using a standard direct linear transformation (DLT) [9] from a set of foot and head pairs. We then employ the rank constraint on $G$ for valid homologies by using the singular value decomposition. Together with the rank constraint, we impose an additional constraint on the vertical vanishing point and the horizon line, which is derived from the pole-polar relationship [2].

Once $G$ is estimated, the camera parameters are straightforwardly obtained from $v$ and $l$ using (2) as

$$h = -\frac{1}{l}(g_1 + g_2 + g_3), \quad \theta = -\arctan\frac{g_0}{g_3 + g_4},$$
$$\rho = -\arctan\frac{g_0}{g_3 + g_4}, \quad f = \sqrt{(g_3 + g_4)(g_3 + g_4)}.$$  

where $g_i$ is the $i$-th element of $G$.

Further, we normalize the foot and head locations separately, and compute their associated normalization transforms $T$ and $T'$; for stable results [9]. Each transform consists of translation and scaling; the centroid of the transformed points is the origin and their average
distance from the origin is $\sqrt{2}$. After we estimate $\hat{G}$ for normalized data, the final $G$ is obtained by unnormalizing $\hat{G}$. In summary, the procedure of the initial estimation of the camera parameters is as follows:

1. Normalize the foot and head locations.
2. Compute $G$ for normalized data.
3. Unnormalize $G$ by $G = T^{-1}\hat{G}T$.
4. Enforce the constraints on $G$.
5. Compute the camera parameters $m$ using (4).

### III. Variational Bayesian Calibration

The crux of our Bayesian self-calibration is computing the posterior distribution over the camera parameters $m$, which is given by $p(m|D) \propto p(D|m)p(m)$ for the likelihood $p(D|m)$ and the prior $p(m)$. Note that $D = \{z_n, z'_n\}_{n=1,N}$ is a set of observation variables, where $Z = \{z_n\}_{n=1,N}$ and $Z' = \{z'_n\}_{n=1,N}$ denote noisy observations of foot and head locations, respectively. In this problem, we assume that the noise models of foot and head locations are based on Gaussian distributions, whose covariances are measurable. We also assume that we can observe a sufficient number of pedestrians, whose heights are described by a Gaussian distribution.

---

Fig. 2 Bayesian model for self-calibration

3.1. Bayesian Model

To deal with the uncertainties, we introduce latent variables $X = \{x_n\}_{n=1,N}$ for the true foot locations in the image, and $l = \{l_n\}_{n=1,N}$ for the true pedestrian heights. The true head locations $X' = \{x'_n\}_{n=1,N}$ depend on the foot locations $X$, the heights $l$, and the homologies $\{H_n\}_{n=1,N}$. For each pedestrian, we have observations $(z_n, z'_n)$ with uncertainties $(\Sigma_{z_n}, \Sigma_{z'_n})$, and corresponding latent variables $x_n$ and $l_n$. We assume Gaussian noise model, therefore, our observations $(z_n, z'_n)$ are given by the true locations $(x_n, x'_n)$ with additive Gaussian noises so that $z_n = x_n + \epsilon_n$ and $z'_n = x'_n + \gamma_n$, where $\epsilon_n$ and $\gamma_n$ are zero mean Gaussian random variables with covariances $\Sigma_{z_n}$ and $\Sigma_{z'_n}$, respectively. Thus we can write

$$
p(z_n|x_n) = N(z_n|x_n, \Sigma_{z_n}),
$$

$$
p(z'_n|x'_n) = N(z'_n|x'_n, \Sigma_{z'_n}),
$$

where $N(\cdot)$ denotes a Gaussian distribution, and $x'_n = H(x_n, l_n, m)$ is a transformation from the foot location to the head location in the image, which is derived from (3) and can be written in terms of $x'_n, l_n$, and $G$ as

$$
x'_n = \frac{1}{1 + l_n g_{x,x'_n} \left| \frac{g_{z,x'_n}}{g_{z,x'_n}} \right|} x_n + l_n \left| \frac{g_{z,x'_n}}{g_{z,x'_n}} \right|.
$$

Fig. 2(a) illustrates the latent variables and the uncertainties on the observations in the image. By assuming independence of the observations $D$, the likelihood $p(D|X, L, m) = p(Z|X) p(Z|X, L, m)$ is given by

$$
p(D|X, L, m) = \prod_{n=1}^{N} p(z_n | x_n) p(z'_n | x'_n, l_n, m).
$$

Fig. 2(b) illustrates the probabilistic graphical model of our framework.

We use an isotropic Gaussian prior over each latent foot location $x_n$ governed by a single precision parameter $\alpha$, given by

$$
p(x_n) = N(x_n | \mu_{x|x}, \alpha^{-1}I),
$$

where $\mu_{x|x}$ is the initial location, which may be set to the observed foot location. Also, we use a Gaussian prior over each latent height $l_n$, which is given by

$$
p(l_n) = N(l_n | \mu_0, \sigma_0^2),
$$

where $\mu_0$ and $\sigma_0$ are known in advance, and common for all $l_n$'s. The prior of head location $p(x'_n)$ is implicitly
given by the priors of foot location and height. By assuming independence of \( X \) and \( L \), we have
\[
p(X) = \prod_{n=1}^{N} p(x_n), \quad p(L) = \prod_{n=1}^{N} p(l_n).
\]
(10)
We use a Gaussian prior over \( m \), which is given by
\[
p(m|\mu_m, \Lambda^{-1}) = \mathcal{N}(m|\mu_m, \Lambda^{-1}),
\]
(11)
where \( \mu_m \) is the initial parameter set, and \( \Lambda \) is a hyperparameter for the precision matrix of the Gaussian. A Wishart distribution is chosen as the conjugate hyperprior over \( \Lambda \), which is given by
\[
p(\Lambda) = W(\Lambda|\nu_0, \mathbf{W}_0),
\]
(12)
where \( W(\cdot) \) denotes a Wishart distribution with a symmetric positive definite matrix \( \mathbf{W}_0 \) and a scalar \( \nu_0 \).

As all factors are specified, the joint distribution of all of the random variables in our model is given by
\[
p(D, \Phi) = p(D|X, L, m)p(X|L, m, \Lambda)p(L)p(\Lambda)p(D|\Phi) + \text{const}.
\]
(13)
where we defined a set of all of the random variables except for the observations as \( \Phi = \{X, L, m, \Lambda\} \). The posterior distribution \( p(m|D) \) is obtained by marginalizing out all of the latent variables as follows:
\[
p(m|D) = \frac{1}{p(D)} \int_{X, L, \Lambda} p(D, \Phi)dXdLd\Lambda.
\]
(14)
Seeking an analytic solution to \( p(m|D) \), however, is infeasible due to the intractability of the integration and the nonlinearity in the transformation \( H(x_n, l_n, m) \).

3.2. Variational Distribution

Our goal is to find the posterior distribution \( p(m|D) \), which is approximated by the factor \( q(m) \) of the variational distribution. Because each of optimal factors depends on moments evaluated with respect to the other distributions, we consider the variational distribution over all random variables and iteratively update the optimal factors. For a tractable solution, we assume the factorization on the variational distribution \( q \) as
\[
q(X, L, m, \Lambda) = q(X)q(L)|q(m)|q(\Lambda),
\]
(15)
where \( q \) distributions only depend on the arguments. Then, by making use of the general result of the variational inference [8], the log of the optimized factor for a variable \( S \) in \( \Phi \) is given by
\[
\text{ln} q^*(S) = \mathbb{E}_q[\text{ln} p(D|\Phi)] + \text{const},
\]
(16)
where \( S^q \) denotes variables in \( \Phi \) except for \( S \).

For a tractable variational Bayesian inference, we approximate the nonlinear transformation \( H \) in Equation (6) so that the expectations of the likelihood term for the head location are expressed analytically. Since our variational Bayesian inference is iterative method, at each iteration we use a linear approximation of the transformation by using the parameters computed in the previous iteration. As a result, the approximation gives analytically tractable posterior distributions over the random variables.

Since we get rid of the nonlinearity in the likelihood, the optimized factors can easily be derived by making use of the joint distribution (13) followed by completing the square for (16). The resulting optimized factors are given by
\[
q^* (X) = \mathcal{N}(x_n|\mu_x, \Sigma_x),
\]
(17)
\[
q^* (l_n) = \mathcal{N}(l_n|\mu_l, \sigma_l^2),
\]
\[
q^* (m) = \mathcal{N}(m|\mu_m, \Sigma_m),
\]
\[
q^* (\Lambda) = \mathcal{W}(\Lambda|\nu_0, \mathbf{W}_0),
\]
where
\[
\Sigma^{-1}_x = \Sigma_{xx}^{-1} + \sum_{n=1}^{N} J_{x}^{T} J_{x} \nu_0 + \sum_{n=1}^{N} J_{m}^{T} J_{m} \nu_0,
\]
the Jacobians of $x_n'$ with respect to $x_n$, $l_n$ and $m$, respectively, evaluated at $(\hat{x}_n, \hat{l}_n, \hat{m}) = (\mu_{x_n}, \mu_{l_n}, \mu_{m})$. We omit detailed derivations of the optimized factors.

3.3. Variational Lower Bound and Algorithm Outline

For the proposed Bayesian model, the variational lower bound $L_q$ is given by

$$ L_q = \mathbb{E}_q[\ln p(D, \Phi)] - \mathbb{E}_q[\ln q(\Phi)]. \quad \text{(19)} $$

Since variational Bayesian inference maximizes $L_q$ by iteratively updating the variational distribution, $L_q$ should not decrease during the iteration. Thus we can utilize $L_q$ not only to check the convergence, but also to verify the correctness of both the mathematical derivations and their software implementation [8].

In summary, we begin the iteration by initializing the parameters of the prior distributions; we need just $\mu_{\nu_0}$, $\nu_0$, and $W_0$, since $p(X)$ is non-informative and $p(L)$ is known priorly. $\mu_{\nu_0}$ is set to the initially estimated camera parameters $m_0$ described in Section 2.3. We set $W_0^{-1} = \text{diag}(\{m_0^2/2, \pi/2, m_0^2\})$, which implies nearly flat prior, and $\nu_0 = 3,001$ so that the covariance matrix of the posterior over $m$ merely depends on the observations. We then alternately update the parameters of each factor in turn using (18) until convergence. The convergence can be checked by monitoring $L_q$ or $q(m)$.

IV. Experiments

We rigorously test the proposed algorithm (VBI) on real data as well as synthesized data. We also compare VBI to others: our implementation of K05 [2] and L11 [3] for synthesized data, and B16 [4] for real data.

Table. 1 Ground truth parameters for synthesized data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$h$ (m)</th>
<th>$\theta$ (°)</th>
<th>$\rho$ (°)</th>
<th>$f$(pix.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>15</td>
<td>5</td>
<td>600</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>30</td>
<td>-2</td>
<td>800</td>
</tr>
</tbody>
</table>

Fig. 3 Noise-free synthesized data for 64 targets: lines and numbers represent the principal axes and the true height, respectively: dataset1 (left) and dataset2 (right)

4.1. Synthesized Data

We test the proposed algorithm on synthesized data, generated from the true parameters as shown in Table 1, which are the common settings for surveillance. For this purpose, for each set of camera parameters, we randomly generate noise-free foot/head locations on the 640×360 image plane as shown in Fig. 3. A Gaussian height distribution with mean of 1.67m and standard deviation of 0.1m is used. When generating a target, its foot location is uniformly drawn on the image and its height is drawn from the height distribution, and corresponding head location is computed by using a homology determined by the height and the true camera parameters. Finally, zero-mean Gaussian noises are added to the foot/head locations followed by estimating the camera parameters. This process is repeated 100 times. We measure root-mean-square error (RMSE) of the estimated camera parameters, by varying the noise level (standard deviation) and the number of targets.

4.1.1. Estimation Error vs. Noise Level

We measure RMSE as the noise level increases from 1 to 7, for the fixed 512 targets. Fig. 4 shows the results for the synthesized datasets for 4 algorithms: Init (Section 2.3), K05 [2], L11 [3], and VBI(proposed). For both datasets, the RMSE increases as the noise level increases. It can be seen that our algorithm finds the optimal solution start from the unstable parameters, and consistently outperforms all others as the noise level varies. In other words, it means that the proposed algorithm is more robust to the noise on the image plane than other algorithms, where the noise is induced from foot/head localization, such as object detection.
4.1.2. Estimation Error vs. Number of Targets

Along with the previous experiment, we measure RMSE as the number of targets varies from 64 to 2048, for the noise level 3. As shown in Fig. 5, the RMSE decreases as the number of targets increases for both datasets. It can be seen that our algorithm consistently outperforms all other algorithms as the number of targets varies. In brief, this means that the proposed algorithm requires fewer targets than other algorithms to acquire the camera parameters of similar quality. Ultimately, the proposed algorithm is superior to others in terms of the error and the stability.

4.1.3. Estimation of the Pedestrian Heights

One of the reason why the proposed algorithm is more accurate than other algorithms is that VBI deals with noise in the target heights by computing the posterior distributions of the pedestrian heights as well. To verify that, we analyze the relation between the true height and the estimated mean of the height, when the noise level is 1 and 7, for 2048 targets, as shown in Fig. 6. For the lower noise level, the pedestrian heights are accurately estimated, which implies an accurate calibration. On the contrary, estimation of the pedestrian heights suffers from the higher noise level, and it may result in inaccurate estimation of the camera parameters.

4.2. Real Data

We use two real sequences from publicly available datasets: Terrace0 from EPFL [10] and Outdoor from VPTZ [11]. Both sequences are typical surveillance scenes as shown in Fig. 7. The Terrace0 sequence has resolution of 720×576, and shows 9 pedestrians walking freely. The Outdoor sequence has resolution of 1280×960, and shows many people on the ground. Since actual height distribution is unknown, we arbitrary set the mean and standard deviation to 1.75m and 0.1m, respectively, for both sequences.

4.2.1. Extraction of Foot and Head

Similar to [4], we trained a head/foot detector based on YOLO [12], where input resolution of the network is 288×288. The detector is trained for two classes, head and feet (when the legs cross), and achieved mAP of 97.94% for our test data. When extracting foot/head, we
first execute a pre-trained human detector (also based on YOLO) then we run our head/feet detector on each of the detected human regions (Fig. 8(b)), followed by computing foot/foot locations and the uncertainties for each detection. Also, while collecting foot/foot pairs at a regular interval of time, the foot/foot locations are checked to ensure no duplication.

For each of head/foot detections, we use center of the bounding box as the foot/foot location. Uncertainties are obtained from the detected bounding boxes as $\Sigma_w = \text{diag}([w/2, h/2]^2)$ and $\Sigma_h = \text{diag}([w'/2, h'/2]^2)$, where $(w, h)$ and $(w', h')$ are the (width, height) of the bounding box for the feet and head, respectively.

In addition, we eliminate outliers by using regression on the foot/foot distance $d$, and the principal axis angle $a$, where $d$ is distance between the foot and the head, and $a$ is angle between the principal axis and the $y$-axis, as shown in Fig. 8(a). Intuitively, $d$ and $a$ of a pedestrian somewhat depend on the image location. Therefore, outliers can be removed by inspecting the dependencies with linear regression and then checking the deviations from the regression lines. Fig. 9 shows the statistics of candidates (Outdoor), where the green and red lines represent the regression lines and the ranges for inliers respectively.

As a result, 2459 and 1811 candidates are detected for each sequence, with 1724 and 1387 targets remaining after outlier elimination. Fig. 10 shows the foot/foot locations and the uncertainties of the inliers as red dots and blue ellipses, respectively.

4.2.2. Calibration Results

The calibration results are shown in Table 2, where the ground truth parameters are compared with the results of the algorithms: K05 [2], L11 [3], B16 [4] and VBI(ours). For both sequences, it can be seen that our VBI surpasses others. L11 and B16 show comparable results, since they employ specialized outlier elimination algorithms and post-processing stages. The inaccuracy in the estimated parameters may arise from the inconsistent assumptions on intrinsic parameters (ex. the principal point on the image center), lens distortion, or incorrect statistics for the pedestrian height.

Together with the parameter estimation, we measure the uncertainty of the estimated camera parameters, which is represented by the covariance $\Sigma_{\Theta}$. Table 3 shows the uncertainties of the estimated camera parameters. It can be seen that the true parameters are within 1-stddv of the estimated camera parameters except for the $\rho$ of the Terrace0 and the $h$ of the Outdoor. The inaccuracies may arise from incorrect prior distribution of the pedestrian height (1.75±0.1) or inaccurate head/foot localization. Nevertheless, most of the estimated parameters provide reliable confidence that has benefited from our Bayesian framework.
Table. 2 The estimated camera parameters

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>$h$ (m)</th>
<th>$\theta$ (°)</th>
<th>$\rho$ (°)</th>
<th>$f$(pix.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrace0</td>
<td>GT</td>
<td>1.90</td>
<td>15.60</td>
<td>1.11</td>
<td>876.61</td>
</tr>
<tr>
<td></td>
<td>K05</td>
<td>1.81</td>
<td>9.03</td>
<td>1.86</td>
<td>1408.68</td>
</tr>
<tr>
<td></td>
<td>L11</td>
<td>1.75</td>
<td>14.00</td>
<td>1.31</td>
<td>925.74</td>
</tr>
<tr>
<td></td>
<td>B16</td>
<td>1.93</td>
<td>15.50</td>
<td>1.82</td>
<td>850.00</td>
</tr>
<tr>
<td></td>
<td>VBI</td>
<td>1.90</td>
<td>15.33</td>
<td>0.67</td>
<td>849.83</td>
</tr>
<tr>
<td>Outdoor</td>
<td>GT</td>
<td>8.78</td>
<td>19.11</td>
<td>0.96</td>
<td>1197.80</td>
</tr>
<tr>
<td></td>
<td>K05</td>
<td>8.31</td>
<td>17.98</td>
<td>1.28</td>
<td>1084.40</td>
</tr>
<tr>
<td></td>
<td>L11</td>
<td>8.40</td>
<td>19.04</td>
<td>1.87</td>
<td>1108.51</td>
</tr>
<tr>
<td></td>
<td>B16</td>
<td>8.83</td>
<td>18.89</td>
<td>-0.18</td>
<td>1019.00</td>
</tr>
<tr>
<td></td>
<td>VBI</td>
<td>9.72</td>
<td>18.88</td>
<td>0.91</td>
<td>1188.84</td>
</tr>
</tbody>
</table>

Table. 3 Uncertainty of the estimation for the real data

<table>
<thead>
<tr>
<th>Data</th>
<th>$h$ (m)</th>
<th>$\theta$ (°)</th>
<th>$\rho$ (°)</th>
<th>$f$(pix.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrace0</td>
<td>0.024</td>
<td>0.473</td>
<td>0.159</td>
<td>28.05</td>
</tr>
<tr>
<td>Outdoor</td>
<td>0.206</td>
<td>0.514</td>
<td>0.234</td>
<td>43.83</td>
</tr>
</tbody>
</table>

V. Conclusion

In this paper, we have presented a novel pedestrian-based variational Bayesian self-calibration algorithm taking account of the uncertainties both on the foot/head locations in the image and the pedestrian heights in the real world. The proposed algorithm requires neither exact heights of pedestrians nor controlled calibration procedures. Instead, it uses statistics of the pedestrian heights, which is commonly known. To conclude, our contribution is two fold: first, we take account of the uncertainty in the pedestrian heights and construct the probabilistic model, second we use variational Bayesian inference to solve the posterior probabilities of the random variables including the camera parameters in analytic forms. Rigorous test have verified that the proposed algorithm estimate the camera parameters and their confidences accurately for the noisy foot/head observations of arbitrary pedestrians. Finally, our future work includes (1) taking account of other intrinsic parameters and lens distortion, (2) improving performance of object detectors by using the estimated camera parameters.

REFERENCES


