Image and Vision Computing 28 (2010) 367–375

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Self-calibration of spherical rectification for a PTZ-stereo system ☆

Dingrui Wan, Jie Zhou *

Department of Automation, TNList, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 31 July 2008
Received in revised form 24 February 2009
Accepted 3 June 2009

Keywords:
PTZ-stereo system
Spherical rectification
Self-calibration

ABSTRACT

Since PTZ (pan–tilt–zoom) camera is able to obtain multi-view-angle and multi-resolution information, PTZ-stereo system using two PTZ cameras has much higher capability and flexibility compared with traditional stereo system. In this paper, we propose a self-calibration framework to deal with the calibration of spherical rectification, which can be deemed as a kind of relative pose estimation, for a PTZ-stereo system. The goal of this calibration is to guarantee high performance of stereo rectification, so that stereo matching can be achieved more efficiently and accurately. In this framework, we assume two PTZ cameras are fully calibrated, i.e., the focal length and the local camera orientation can be computed by given pan–tilt–zoom values. This approach, which is based on point matches, aims at finding uniformly distributed point matches in an iterative way. At each iteration, according to the distribution of previously used point matches, the system could automatically guide two cameras to move to collect a new match. Point matching is firstly performed for the lowest zoom setting (widest field of view). Once a candidate match is chosen, each camera is then controlled to zoom in on corresponding point to get a refined match with high spatial resolution. The final match will be added into the estimation to update the calibration parameters. Compared with previous researches, the proposed framework has the following advantages: (1) Neither manual interaction nor calibration object is needed. Calibration samples (point matches) will be added and removed in each stage automatically. (2) The distribution of calibration samples is as uniform as possible so that biased estimation could be avoided to some extent. (3) The accuracy of calibration can be controlled and improved when iteration goes on. These advantages make the proposed framework more practicable in applications. Experimental results illustrate its accuracy.

1. Introduction

In recent years, automatic visual surveillance techniques have been widely applied in many fields. One trend of the hardware configuration is to use ‘multiple’ and ‘active’ cameras as a visual unit. Multi-camera unit could supply multi-viewpoint images and stereo information, while the active-camera unit could obtain multi-resolution and multi-view-angle images. In our study, we use two PTZ (pan–tilt–zoom) cameras to constitute a dual-PTZ-camera system which is one of the simplest multi-active-camera systems, so the research on such system is significant and promising [1].

The significance of stereo vision is well known. PTZ stereo can be utilized to obtain depth map of local region with high precision and depth map of a large scene, by changing pan–tilt–zoom values [2]. This depth information could be useful in scene understanding. Standard stereo vision assumes that the stereo system satisfy nonverged geometry [3], i.e., the epipolar lines are parallel to each other. However , this assumption does not hold for many stereo systems. Stereo rectification is a way to make arbitrary stereo image pairs to become nonverged geometry, which makes the searching scope of stereo matching to be one dimension [4]. Many rectification algorithms have been reported in the literature [5–9]. For a PTZ-stereo system, since the camera parameters are alterable, a feasible way is to build a rectification system based on camera parameters (such as pan–tilt–zoom values) instead of image content. There are few literatures on the rectification-related problem for dual-PTZ-camera system except for our earlier works [10], and we call it “spherical rectification”. However, the calibration of the spherical rectification model still needs improvement because of some flaws. This problem can be seemed as a kind of relative pose estimation between two PTZ cameras. In this paper, we propose a self-calibration framework to deal with this problem.

In our study, we assume each PTZ camera is already calibrated, i.e., the focal length and the local camera orientation can be computed by given pan–tilt–zoom values. Some relevant research about single PTZ camera calibration can be found in the literature,

☆ This work was supported by Natural Science Foundation of China under Grant Nos. 60673106, 60721003 and 60573062, and the Specialized Research Fund for the Doctoral Program of Higher Education.

* Corresponding author. Address: Department of Automation, TNList, Tsinghua University, Qinhuayuan 1, Haidian District, Beijing 100084, PR China. Tel.: +86 10 62796881.

E-mail addresses: wandingrui00@mails.tsinghua.edu.cn (D. Wan), jzhou@tsinghua.edu.cn (J. Zhou).

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.06.003

for example [11,12]. The calibration of spherical rectification model includes epipoles and zero longitudes estimation for two cameras, and this problem can be seemed as one specific application of motion estimation problem in computer vision [13]. Compared with estimating relative pose, our problem has one less degree of freedom, i.e., the distance between two cameras. So we directly estimate the spherical rectification model, instead of solving a standard motion estimation problem. However, if the relative pose of the two PTZ cameras is known, the spherical rectification model can be directly computed. No matter what kind of problem we regard this estimation to be, two aspects should be considered for a point-matches based approach: (1) How to control the distribution of point matches? (2) How to improve the accuracy of point matches?

In order to compute the spherical rectification model, we have proposed a two-step calibration approach in [10]: firstly, use the fundamental matrix to estimate two epipoles by the epipolar constraint; secondly, calculate the two zero longitudes by minimizing the longitude difference of point matches with fixed epipoles. However, the accuracy of calibration is restricted, because the error in estimating epipole might become a bottleneck of the whole calibration. Furthermore, since the calibration is based on point matches, better performance needs well distributed and accurate point matches as well, but how to collect such matches is not concerned in this study. Fujiki et al. [14] proposed calibration approach to deal with central-omnidirectional cameras system which also uses the spherical stereo model. This approach uses an iterative way to achieve a more accurate calibration. Compared with PTZ-stereo system, the main difference is that the calibration can be accomplished using one single spherical image pair. So, all point matches have fixed spatial resolution, and there is no need to consider automatically gathering calibration point matches.

In our study, we proposed a novel self-calibration framework to deal with this problem. This approach is based on point matches and works iteratively. These calibration point matches are collected with high spatial resolution by automatically changing PTZ parameters of two cameras, so this calibration is convenient and reliable. For simplicity, we define *sample* as a point match; *sampling* as collecting samples for calibration. All samples used for calibration form a *sample set*. In the proposed framework, the following three key problems need to solve:

(1) An effective optimization algorithm is needed to estimate target calibration parameters for a given sample set.
(2) An automatic sample management mechanism is needed, so that the calibration samples can be dynamically added and removed.
(3) When adding a new sample, we need a proper camera control strategy to ensure that a high-spatial-precision sample is likely to be found.

Compared with previous work [10], this framework has following advantages:

(1) The framework works in an iterative way. In each stage, calibration samples can be automatically added and removed with neither manual interaction nor calibration object.
(2) New sample is added to make the distribution of samples, which is determined by current calibration sample set and current estimated parameters, to be as uniform as possible, so that biased calibration caused by the non-uniform distribution of samples can be avoided to some extent.
(3) The accuracy of calibration can be controlled. Since the accuracy can be improved after adding suitable samples stage by stage, a desired accuracy can be achieved by controlling the number of stages.

These advantages will greatly increase the practicability of the calibration. Experimental results show that the performance of calibration is much better than the method used in Ref. [10].

The remainder of this paper is organized as follows: in Section 2, the stereo model of dual-PTZ-camera system is depicted. In Section 3, we introduce the proposed self-calibration framework. Experimental results will be shown in Section 4. In Section 5, we summarize the paper.

2. Dual-PTZ-camera stereo model

2.1. Single PTZ camera model

The pin-hole camera model is used in our study:

$$\tilde{x}_n = K_0^{-1}\tilde{x} = \kappa Z(z)R(p,t)X, \tag{1}$$

where x and X are image coordinates and world coordinates, respectively; symbol '$\sim$' means homogeneous coordinates. x_n is the normalized image coordinates. p, t, z are the PTZ parameters supplied by camera. κ is a scaling factor. For simplicity, we set pixel aspect ratio to be 1, and skew be 0; and no radial distortion is considered.

K_0 is a 3×3 matrix to translate image origin to principal point (u_0, v_0). We assume that the rotation axes of pan and tilt are orthogonal and intersect at one point, which is chosen as the origin of the world coordinate system, and hence no translation factor is considered.

$R(p,t)$ is a rotation matrix determined by the 'pan' and 'tilt' angles. $Z(z) = \mathrm{diag}\{f(z), f(z), 1\}$ is a scaling matrix determined by 'zoom' parameter. We use the approach in [11] to estimate Z at several discrete zoom levels z, and then choose a proper model to fit these discrete values. In our study, we use an experiential function, $f(z) = a\exp(bz) + c\exp(dz)$ [1], where a, b, c, d are the parameters need calibration.

In our study, the SONY EVI D70 camera is in use. For the detailed calibration of single PTZ camera, please refer to Ref. [1].

2.2. Spherical stereo model

In the ideal dual-camera stereo model, two cameras have the same focal length, and the optical axes are parallel and perpendicular to the baseline. This model is always called the classical stereo model or nonverged stereo model, as the two optical axes are parallel [3]. Using this model, stereo matching can be easy fulfilled efficiently. However, in practice, this model does not always hold well for real applications, especially when two cameras have different PTZ parameters. Stereo rectification is a way to make the verged geometry to be nonverged. Rectification is able to reduce the searching scope in stereo matching from 2D to 1D, so both performance and efficiency of stereo matching can be improved. For dual-PTZ-camera system, we adopt the spherical rectification method which has been proposed in our previous work [10].

The main idea of the spherical rectification method is to bring in the longitude-latitude coordinate system, and the goal is to have the longitude components of corresponding points be the same after coordinates conversion. Fig. 1 briefly illustrates this method. The definition of longitude and latitude coordinates are shown in Fig. 2. The longitude, α_x, is defined as the elevation angle relative to the baseline, i.e., the angle from OM to $OX'(OM \perp EE')$, and $\alpha_x \in [-\pi, \pi)$; the latitude, β_x, is defined as the angle from Oe to OX, and $\beta_x \in [0, \pi]$. According to this definition, OM has a zero longitude value, so we call it zero-longitude vector.

We summarize the conversion between original image coordinates and rectified image coordinates in Fig. 3.

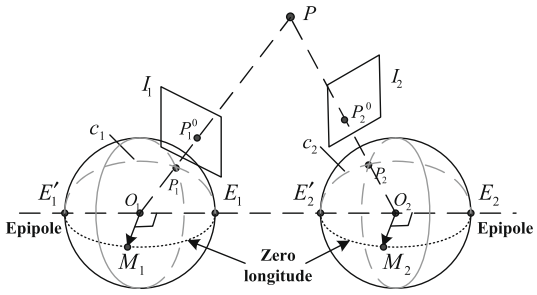

Fig. 1. Sketch map of spherical rectification.

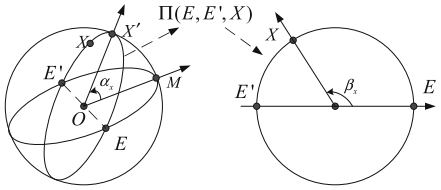

Fig. 2. Definition of longitude (α_x) and latitude (β_x) coordinates.

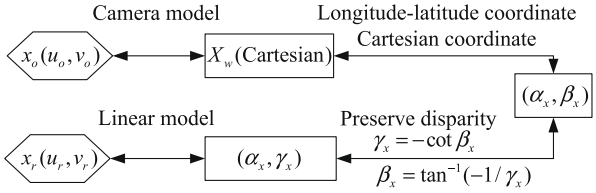

Fig. 3. Flow chart of coordinates conversion.

2.3. Calibration of spherical rectification

Calibration of spherical rectification is to construct the longitude-latitude coordinate system for each camera, which is determined by two epipoles (E_i, E_i') and zero longitudes $(O_iM_i$, where $O_iM_i \perp O_iE_i)$, where $i = 1, 2$ (see Fig. 1).

The calibration method used in [10] is based on point matches between two camera coordinate systems. It includes two steps: (1) use the fundamental (or essential) matrix to estimate epipoles by the epipolar constraint; (2) choose an arbitrary $O_1M_1 \perp O_1E_1$, and find $O_2M_2 (\perp O_2E_2)$ by minimizing the longitude difference of corresponding calibration points. A major shortcoming of this method is that the error in estimation of epipoles will affect the accuracy of O_iM_i, $i = 1, 2$. In order to solve this problem, we refer to the method proposed in [14], which directly estimates all parameters at the same time.

In order to build a self-calibration framework, besides specific parameter estimation algorithm, we need to consider how to obtain useful calibration samples as well. This problem is not concerned in [10,14]. In our situation, the following two aspects should be considered:

(1) *Sample management:* An evaluation mechanism should be designed to look for a desired sample which will be the best for calibration and to determine a sample which should be removed;
(2) *Camera control strategy:* There are two requirements: one is to ensure two cameras have overlapped FOV, and the other one is to use high zoom level to guarantee the precision of samples.

3. Automatic self-calibration framework

The flow chart of the proposed framework is shown in Fig. 4. In this framework, several initial samples and initial parameters are prepared in initialization module. Then, the procedure works in an iterative way. In each stage, the system evaluates a distribution of current calibration samples according to current estimated parameters, and determines a rough area that could generate a new sample which will make the distribution be more uniform. The two cameras are then automatically controlled towards this area and collect new sample with highest spatial resolution. After a sample is added into calibration sample set, the target parameters will be refined in the optimization module. Under some certain conditions, sample removing module will be triggered. Once a sample is regarded as an outlier, it will be ruled out. The stop conditions are used to determine when to terminate the procedure.

We first parameterize this calibration problem. Since the estimation target, epipoles (E_i, E_i') and zero longitudes $(O_iM_i,\ i = 1, 2)$ (see Fig. 1) are orthogonal, an orthonormal matrix (R_i^r) can be used to represent them, where O_iE_i and O_iM_i equal the first and second row of R_i^r, respectively. R_i^r is also a rotation matrix which can be parameterized by three euler angles. However, as the zero longitudes (O_1M_1 and O_2M_2) only have relative meaning, only five parameters are needed, and we denote them by $\theta = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5]^T$. So R_i^r can be determined by θ:

$$R_1^r(\theta_1, \theta_2) = \begin{bmatrix} s_1c_2 & -s_2 & c_1c_2 \\ c_1 & 0 & -s_1 \\ s_1s_2 & c_2 & c_1s_2 \end{bmatrix}, \tag{2}$$

and

$$R_2^r(\theta_3, \theta_4, \theta_5) = \begin{bmatrix} s_3c_4 & -s_4 & c_3c_4 \\ s_3s_5s_4 + c_3c_5 & s_5c_4 & c_3s_5s_4 - s_3c_5 \\ s_3c_5s_4 - c_3s_5 & c_5c_4 & c_3c_5s_4 + s_3s_5 \end{bmatrix}, \tag{3}$$

where $s_j = \sin\theta_j,\ c_j = \cos\theta_j,\ j = 1, 2, 3, 4, 5$.

In our study, this calibration is based on point matches between two PTZ cameras. We call these matches *samples*, and the *k*th sample is denoted by $\{X_1^k, X_2^k\}$, where X_i^k is the normalized camera coordinates of camera-*i* (i.e., $\|X_i^k\| = 1$). Assume x_i^k is the matched point

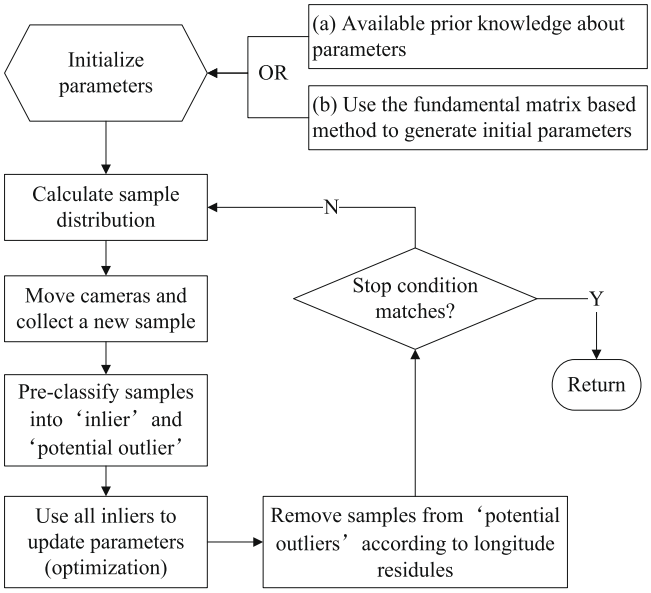

Fig. 4. Flow chart of self-calibration framework.

in camera-i's image, we back project x_i^k onto the unit sphere centered at camera-i's center, and X_i^k is the intersection point.

3.1. Initialization

In order to obtain an initial value of θ, we first estimate the essential matrix between two views from two PTZ cameras, respectively, and then the epipoles (E_i, E_i') can be calculated. Finally, the zero longitudes (O_iM_i) can be computed by the estimated E_i and point matches.

The two views are captured by manually controlling two PTZ cameras. In order to guarantee a better initial value, it is better that the two views have larger overlapping visual field. For each view, we extract SIFT [15] feature points and match them between two views (http://vision.ucla.edu/vedaldi/). Since the camera is calibrated, according to the camera model equation (1), the normalized camera coordinates, X_1^k and X_2^k, can be computed from the kth point match. X_1^k and X_2^k satisfy the epipolar constraint:

$$\left(X_2^k\right)^T H_{21} X_1^k = 0, \tag{4}$$

where H_{21} is the essential matrix.

We use the five-point algorithm as a hypothesis generator within a RANSAC scheme [16,17] to obtain the essential matrix (H_{21}) and collect inliers among all point matches. For each hypothesis, the cost is defined as sum of truncated absolute longitude residual errors of all samples:

$$\text{cost} = \sum_k \min\left\{|e^k|, T\right\}, \tag{5}$$

where e^k is the longitude residual error of the kth sample, and T is the threshold value (in our experiment $T = 0.1$). Inliers are defined as those samples satisfy $|e^k| < T$.

(1) The key part is to compute e^k through a given essential matrix, H_{21}: (1) compute E_i according to the epipolar constraint $(H_{21}E_1 = E_2^T H_{21} = 0)$ by SVD decomposition. The sign of E_i is determined through the latitude values of all samples, $\beta_{x,i}^k$ (see definition in Fig. 2). From the definition of epipoles in PTZ-stereo system, $E_i' = -E_i$. For ideal case, all the latitude values should satisfy $\beta_x^1 < \beta_x^2$, i.e., $\angle PO_1E_1 < \angle PO_2E_2$ in Fig. 1. For the four combinations: $\{E_1, E_2\}$, $\{-E_1, E_2\}$, $\{E_1, -E_2\}$ and $\{-E_1, -E_2\}$, we find the one with most samples satisfy $\beta_x^1 < \beta_x^2$ as the solution. If this number is smaller than 60% of total amount of samples, we directly reject this hypothesis.

(2) O_iM_i can be chosen after E_i and E_i' are determined. Since $O_iM_i \perp O_iE_i$, there is only one degree of freedom in computing O_1M_1 and O_2M_2. Given a arbitrary vector v, let $M_1 = v \times E_1$ and $M_2' = v \times E_2$, then the longitude coordinates, $\alpha_{x,i}$ (see definition in Fig. 2), can be obtained. Considering the five samples used for estimating H_{21}, the average difference of longitude coordinates is denoted by $\delta\alpha$; then the final M_2 is determined by rotating M_2' around O_2E_2 by $\delta\alpha$.

(3) The longitude coordinates of each sample in two camera system can be calculated. The residual error of the kth sample is: $e^k = \alpha_2^k - \alpha_1^k$. Then the cost of a given hypothesis can be computed by Eq. (5).

Finally, we choose the hypothesis with smallest cost. The corresponding E_i and M_i will be used to construct R_i^r, and then θ can be computed through Eq. (2). All the inliers will form the initial calibration sample set, and $\theta^0 = \theta$ will be the initial calibration parameter for further optimization.

If we have some prior knowledge about the calibration parameters, for example, the two cameras are placed side by side on a same plane, θ^0 can be directly given. In this case, this framework can totally work automatically. In our study, we assume no such prior knowledge is available. Note that, since these samples have low spatial resolution, when the number of new added samples with high spatial resolution is large enough, we force these initial samples to be removed.

3.2. Parameter optimization

In our study, we use point matches (calibration samples) to estimate θ by minimizing the sum of square difference of longitude coordinates. In Ref. [14], Fujiki et al. have proposed a similar method by using the Rodrigues' formula of rotation matrix to calibrate the spherical images for central-omnidirectional cameras system. In this paper, we follow this idea to optimize the five parameters in each stage.

Parameter optimization module is running after new samples are added in each stage. In this module, we denote θ^0 as the initial value of θ. In the first stage, θ^0 is generated from the initialization module, and in the later stage, θ^0 can be the optimization result in the previous stage.

Given a point X in the scene, and denote X_i as the camera coordinates of camera-i $(i = 1, 2)$. After spherical rectification, the spherical coordinates $X_i^r = R_i^r X_i$ (R_i^r can be computed from θ, see Eqs. (2) and (3)), and the longitude component can be represented by using the four-quadrant inverse tangent function: $\alpha_i = \text{atan2}(X_i^r(3), X_i^r(2))$, where $X_i^r(m)$ is the mth component of vector X_i^r. According to the definition of spherical rectification, the optimization target is to minimize

$$E = \sum_k \left(\alpha_2^k - \alpha_1^k\right)^2, \tag{6}$$

where k is the index of calibration samples. Consider one-order Taylor expansion,

$$E \doteq \sum_k \left(e^k + \left(J_2^k - J_1^k\right)^T \mathbf{d}\right)^2, \tag{7}$$

where $e^k = \alpha_2^k - \alpha_1^k$; $\mathbf{d}$ is the incremental parameter vector of θ; $J_i^k = \frac{\partial \alpha_i}{\partial \mathbf{d}}|_{X_i^k}$. In order to calculate J_i^k, let $C_i = X_i^r(3)$ and $B_i = X_i^r(2)$, then we have

$$\frac{\partial \alpha_i}{\partial \mathbf{d}} = \frac{\frac{\partial C_i}{\partial \mathbf{d}} B_i - C_i \frac{\partial B_i}{\partial \mathbf{d}}}{B_i^2 + C_i^2}, \tag{8}$$

where $\frac{\partial B_i}{\partial \mathbf{d}}$ and $\frac{\partial C_i}{\partial \mathbf{d}}$ can be easily calculated from Eqs. (2) and (3). The least square method could provide an estimation of the increment:

$$\mathbf{d} = -\left(\sum_k J^k (J^k)^T\right)^{-1} \left(\sum_k e^k J^k\right), \tag{9}$$

where $J^k = J_2^k - J_1^k$. The estimation of θ works in an iterative way. In each iteration we do above operations to update $\theta : \theta \leftarrow \theta + \mathbf{d}$, and finally, θ could be obtained. In our experiment, the stop conditions of this optimization module include: (i) exceeding maximum iteration number; (ii) $\|\mathbf{d}\| < th_\theta$; and (iii) $\frac{1}{N}\sum_k |e^k| < th_e$. If any one of above conditions is met, the optimization procedure stops. In our experiment, the maximum iteration number is set to be 20; $th_\theta = 10^{-5}$; and $th_e = 0.0005$.

3.3. Sample management

After θ is estimated, if the stop condition does not match, sample management module will be triggered to compute the distribution of current calibration samples, and prepare for collecting a new sample.

3.3.1. Add new samples

In θ's estimation, sufficient samples are very important to guarantee the accuracy of estimated θ. On the other hand, the distribution of samples is also significant, especially when the total number of samples is not large enough. If all samples have the same uncertainties, we hope all samples uniformly distribute in longitude coordinate space.

We define a *distribution factor circle*, $f(\alpha)$, where $-\pi \leqslant \alpha < \pi$. $f(\alpha)$ is determined by all calibration samples ($\{X_1^k, X_2^k\}$) with their calibration errors ($|e^k|$). Assume the spherical coordinates of X_i^k ($i = 1, 2$) are denoted by (α_i^k, β_i^k), we mark $\alpha^k = (\alpha_1^k + \alpha_2^k)/2$ and $\beta^k = (\beta_1^k + \beta_2^k)/2$. Then $f(\alpha)$ is defined as

$$f(\alpha) = \sum_k \sin\left(\beta^k\right) \exp\left(-(e^k)^2/\sigma_e^2\right) \exp\left(-\left(\alpha - \alpha^k\right)^2/\sigma_\alpha^2\right). \quad (10)$$

Remark 1. The first part, $\sin(\beta^k)$, is a factor to determine the amplitude of the kth sample's contribution. This term indicates that, if the average latitude (β^k) is more close to $\pi/2$, the contribution of this sample to the distribution circle will be greater.

Given a point on the unit sphere with longitude-latitude coordinates (α, β), because of the uncertainties in pan and tilt parameters, this point may have a displacement error on the unit sphere, $\Delta\varphi$, which is independent with α and β. $\Delta\varphi$ can be decomposed into longitude and latitude components, and the longitude error $\Delta\alpha \approx \Delta\varphi/\sin\beta$. The larger $\sin\beta$, the smaller the $\Delta\alpha$, and the more reliable the longitude residual (e^k), and so the corresponding sample will have higher weight. For simplicity, we directly choose the trigonometric function in Eq. (10).

Remark 2. The second part, $\exp\left(-(e^k)^2/\sigma_e^2\right)$, is another factor to determine the amplitude of the kth sample's contribution. This term indicates that, if the calibration error ($|e^k|$) is smaller, the contribution of this sample to the distribution circle will be greater.

This part can be regarded as a kind of weighting strategy according to corresponding $|e^k|$. A natural idea is to collect more samples around those samples with larger calibration error. This strategy will have more obvious effect for larger number of samples. For simplicity, we choose the Gaussian-like function in Eq. (10), where σ_e is an experiential parameter, and in our experiment we set $\sigma_e = \pi/180$.

Remark 3. The third part, $\exp\left(-\left(\alpha - \alpha^k\right)^2/\sigma_\alpha^2\right)$, determines the influence scope of the kth sample on α. This term indicates that, each sample will have a Gaussian-like effect on the distribution circle centered at its longitude value, α^k. σ_α is an experiential parameter, and in our experiment we set $\sigma_\alpha = \pi/32$.

Basic idea: The ideal situation for the distribution circle is that for all $\forall\alpha \in [-\pi, \pi)$, $f(\alpha)$ is a constant, which can be regarded as a uniform distribution. The principle of adding new samples is to collect samples whose longitude coordinates component is close to $\alpha_{\min} = \arg\min_\alpha f(\alpha)$ on the distribution circle.

In order to efficiently calculate $\alpha_{\min}$, we discrete the range of α into N_{bin} (e.g., 36 in our experiments) equal bins which are represented by their central values, α_i ($i = 1, 2, \ldots, N_{\text{bin}}$). Among these values, we find the one (α_j) satisfies $f(\alpha_j) \leqslant f(\alpha_i)$, and $\alpha_{\min} \approx \alpha_j$. The jth bin is called the candidate bin.

Determine camera parameters: In order to collect new samples with corresponding longitude coordinates close to $\alpha_{\min}$, we have to calculate suitable PTZ parameters of two cameras, so that such a sample will be visible in both views of two cameras.

Step 1. Determine pan and tilt parameters. Since longitude value alone is not enough to determine the camera orientation, two latitude values for two cameras (β_1 and β_2) are also needed. We decide them with the following two considerations:

(1) β_1 and β_2 should be close, because we have no idea about the depth of the scene, if the discrepancy between the orientations of two cameras is large, the common FOV could be small and so it degrade the chance to find point matches we need. In our study, we set $\beta = \beta_1 = \beta_2$.
(2) As we mentioned before, β should be close to $\pi/2$ so that the longitude uncertainty will be small, see Remark 1.

When $(\alpha_{\min}, \beta)$ is given, it is easy to find the camera coordinates on the unit sphere, X, and then, according to the camera model, pan and tilt parameters can be calculated. Considering that this candidate bin (associated with $\alpha_{\min}$) might be selected more than once, in order to avoid collecting reduplicate samples, before compute the pan and tilt parameters, we add a small random value on both $\alpha_{\min}$ and β, so that each time when adding new sample, the cameras' parameters will be different. If the calculated pan and tilt parameters are out of range, we deem that the sample collection at $\alpha_{\min}$ is failed, and this case will be discussed later.

Step 2. Determine zoom parameter. For both cameras, we use the lowest zoom level to improve the possibility that two cameras have larger common FOV. However, spatial resolution will be sacrificed, so we design a refining strategy for compensation.

After new PTZ parameters of both cameras are determined, two cameras will be controlled towards new positions. Feature points extracting and matching will be performed within two images. As the two images might have difference in rotation (or even small difference in scale), SIFT descriptor [15] is utilized. For all point matches, we use current estimated θ to calculate their corresponding longitude coordinates, $\{\alpha_1^j, \alpha_2^j; j = 1, 2, \ldots\}$. Denote $\alpha^j = (\alpha_1^j + \alpha_2^j)/2$ and $e^j = \alpha_1^j - \alpha_2^j$. Those matches whose α^j does not belong to the candidate bin or $|e^j|$ is too large will be removed. For the rest matches, we compute the distribution factor $f(\alpha^j)$, and choose three matches with smallest $f(\alpha)$ as the candidate samples, see an example in Fig. 5(a).

As we mentioned before, lower spatial resolution might cause larger error in points matching, so we design a refining strategy to solve this problem. We firstly choose one of the candidate samples, and calculate the corresponding camera coordinates (X_1, X_2) which will be used to compute new pan and tilt parameters for both cameras. At this time, it is sure that the two matched points will be visible for each camera under any zoom level, because after camera moves to the new PTZ position, if the errors in camera model and camera controlling are not considered, the matched points should be in the image centers (principle point) for both cameras. So we set a high zoom level for both cameras, and move two cameras to new PTZ position. Then, we again perform feature point extraction and matching near the image centers of the two high-spatial-resolution images, and select the point match with highest matching score. Finally, according to PTZ parameters and camera model equation (1), compute the normalized camera coordinates of the selected point match as the new sample, $\{X_1^{\text{new}}, X_2^{\text{new}}\}$.

For example, in Fig. 5, it is failed finding new sample with the first candidate sample, so we move on with the second one, and succeed. The corresponding high-zoom image pair is shown in Fig. 5(b), and the final selected point match is labeled by 'P'.

If it is failed to find new matched point pairs, we choose the next candidate sample and do this procedure again. If all the three candidate samples are run out, we deem that the sample collection in this candidate bin (associated with $\alpha_{\min}$) is failed.

Failure treatment: If the sample collection at $\alpha_{\min}$ is failed, the corresponding candidate bin will be forbidden for the next several stages to improve the efficiency of calibration and avoid endless

Fig. 5. An example of sample collection. (a) Low-zoom image pair with matched points. Three candidate points are labeled by numbers, '1', '2', and '3'. (b) High-zoom image pair with matched points near image center. The finally selected point is labeled by 'P'.

loop. Then we find $\alpha_{\min}$ again from the rest available bins and redo the same procedure, until a new sample is collected.

In our study, we only add one sample in each stage, because first, when a new sample is added into calibration, the distribution circle will change. If we add another sample at the same time, it will likely aggravate the unbalance of the distribution circle. Second, to collect more samples needs more camera movement, and that will occupy lost of time of the whole calibration.

3.3.2. Remove outlier samples

As all samples are delicately collected by the zoom refining strategy, false matching will hardly happen because if a false match is chosen as a candidate sample in low-resolution image pair, it is likely that no match will be found after two cameras move toward this match and zoom in. However, we still need to assume the existence of false matches. On the other hand, the calibration samples might have other errors which might be caused by the mechanical clearance in camera movement, and local point matching error, etc. In order to improve the accuracy of calibration, we allow removing outliers from the calibration data set.

When a new sample is collected, only if there are enough samples (e.g., the total number of samples $N > 15$), we execute the sample removing module. We also use the five-point algorithm as a hypothesis generator within a RANSAC scheme [16,17], which has already been used in initialization (see Section 3.1), to find inlier samples. The rest samples are labeled as 'potential outliers' which will not be removed immediately since the longitude residual computed in RANSAC loop might not be accurate enough. The inlier samples will be used to optimize θ using the algorithm in Section 3.2, so that the longitude residual errors ($|e^k|$) of all samples are recomputed according to the estimated θ. Assume $\bar{e}_{\text{inlier}}$ and σ_{inlier} are the mean and standard deviation of residuals of all the inliers, respectively. For each 'potential outlier', if the corresponding $|e^k| > \bar{e}_{\text{inlier}} + 3\sigma_{\text{inlier}}$, this sample will be removed; otherwise, we keep it as an inlier without recomputing θ.

3.4. Stop condition

In each stage, we need to check whether the stop condition is satisfied or not:

(1) The total number of samples $N > N_{\max}$.
(2) The total number of stages $N_s > N_{s\_\max}$.
(3) The total calibration error $\varepsilon < T_\varepsilon$ and $N > N_{\min}$ which is used to avoid local optimum.

If one of the above conditions is satisfied, the whole procedure will stop, and current estimated θ will be the final result.

4. Experimental results

We utilize two SONY EVI D70 cameras to compose a PTZ-stereo system. We assume no pre-knowledge for θ is available, so we use the method mentioned in initialization module (see Section 3.1) to generate an initial parameter, θ^0.

4.1. Experiment on real data

We use the proposed self-calibration framework to calibrate our PTZ-stereo system. Since it is difficult to measure the groundtruth of the parameters to estimate, we collect those samples used in previous calibration as the testing data (150 samples in total). The average absolute longitude residual ($|e^k|$) on testing data could indirectly show a rough trend of the performance along stages.

The left image in Fig. 6 shows the average absolute longitude residuals on training and testing data in each stage, where the training data are all current used calibration samples. The right image in Fig. 6 shows the final sample distribution. The broken line indicates $f(\alpha)$ for each bin ($N_{\text{bin}} = 36$), which is normalized by $\sum_{i=1}^{N_{\text{bin}}} f(\alpha_i) = 1$. The circle-spoke image indicates the longitude (α) distribution for all samples in $[-\pi, \pi)$. Note that in our experiment, the α distribution only covers about half space in $[-\pi, \pi)$, because: (1) for the used PTZ camera, the range of pan is from $-170°$ to $170°$, and tilt, from $-90°$ to $30°$, so some directions are unreachable; (2) in our system, the two cameras are appended on the top window frame, and few feature points can be detected in the upper half space with sky and inner roof whose corresponding rough longitude range is about $(-\pi, 0)$.

We also provide two original high-resolution image pairs from which two samples are collected, see Fig. 7. Using this calibration result, we rectify a image pair and show the result in Fig. 8.

4.2. Experiment on simulated data

In order to quantitatively compare the estimated parameters with groundtruth, we use some simulated data for experiment. Assume two cameras have only relative translation on x-coordinate, and $E_1^0 = E_2^0 = [1, 0, 0]^T$; $M_1^0 = M_2^0 = [0, 1, 0]^T$. Imitate the real system, we set baseline width $b = 0.75$ m. To generate a sample, we need the spherical coordinates of two cameras, (α_i, β_i), $i = 1, 2$. As the sample adding procedure, when the target α with smallest $f(\alpha)$ is given, we set $\alpha_1 = \alpha_2 = \alpha$, $\beta_1 = \pi/2 + n_\beta$ (n_β is a $N(0, \pi/36)$ Gaussian noise). Assume the depth of this virtual point is randomly chosen from $[20, 200](m)$, then β_2 can be determined. According to (α_i, β_i), the normalized camera coordinates X_i can be computed. Finally, a 3×1 Gaussian noise vector ($N(0, \sigma)$) is added on each X_i. In this experiment, we set $\sigma = 0.001$ which almost equals the uncertainty of pan and tilt values provided by cameras.

When θ is estimated, E_i and M_i can be computed. In order to compare the calibrated result with groundtruth, we define the following three errors:

(1) $\epsilon(E_1)$: the angle between O_1E_1 and $O_1E_1^0$, $< O_1E_1, O_1E_1^0 >$;
(2) $\epsilon(E_2)$: the angle between O_2E_2 and $O_2E_2^0$, $< O_2E_2, O_2E_2^0 >$;
(3) $\epsilon(M_{12})$: since the zero longitude has only relative meaning, we first project M_i onto the plane perpendicular to $O_iE_i^0$, and denote it by M_i'; then $\epsilon(M_{12}) =< O_1M_1', O_1M_2' >$.

Note that, $\epsilon(E_1)$ and $\epsilon(E_2)$ are dominant among the three kind of errors, only when both of them are very small, $\epsilon(M_{12})$ has significance.

We generate eight samples with random α values for initialization. In each stage, the three kind of calibration errors are shown in Fig. 9.

From this result, we can conclude that: (1) the three errors keep low in each stage, so that the estimation of θ is stable. (2) The three errors are basically descending while the number of stage increases, which testifies that the accuracy of estimated θ is improved stage by stage. (3) The distribution of calibration samples is almost uniform, so that the sample collecting strategy is testified.

4.3. Importance of samples' distribution

In order to how the distribution of calibration samples affects the accuracy of calibration, we design an experiment on the simu-

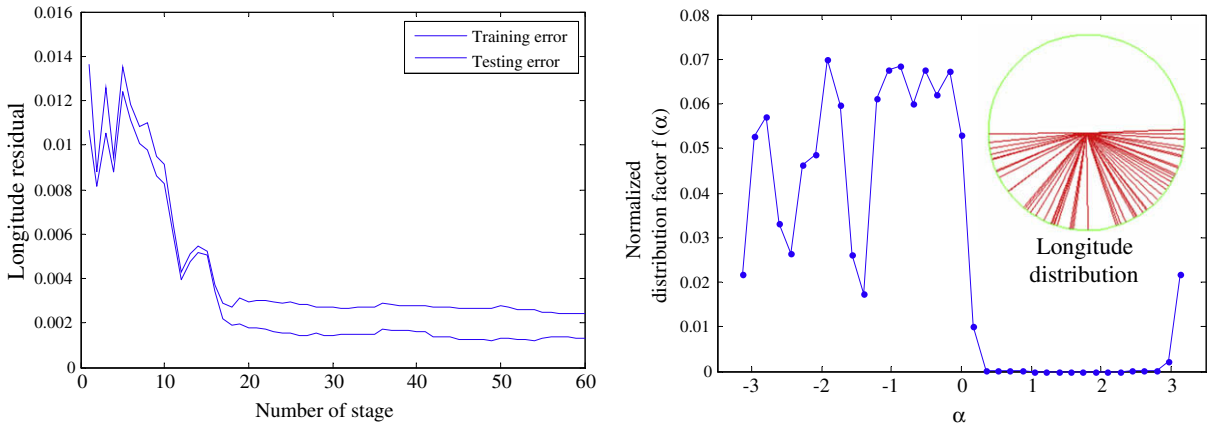

Fig. 6. Experiment on real data. The left image is longitude residuals on training and testing data in each stage. The right image is the final sample distribution: the broken line indicates $f(\alpha)$ for each bin ($N_{\text{bin}} = 36$); the circle-spoke image is α distribution for all samples in $[-\pi, \pi)$.

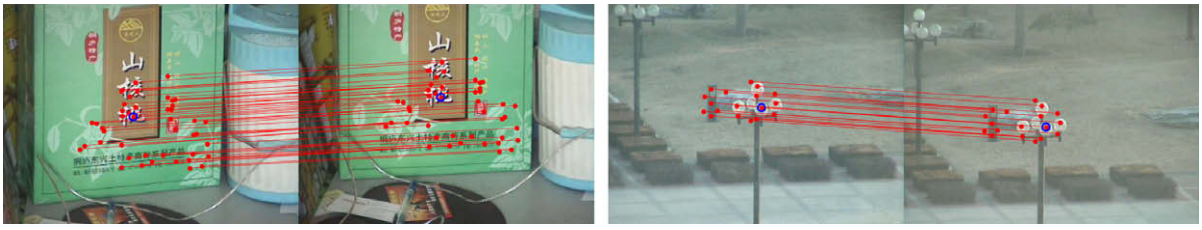

Fig. 7. Two samples: high-resolution image pairs with point matches. The point pair with blue circle indicates the collected sample.

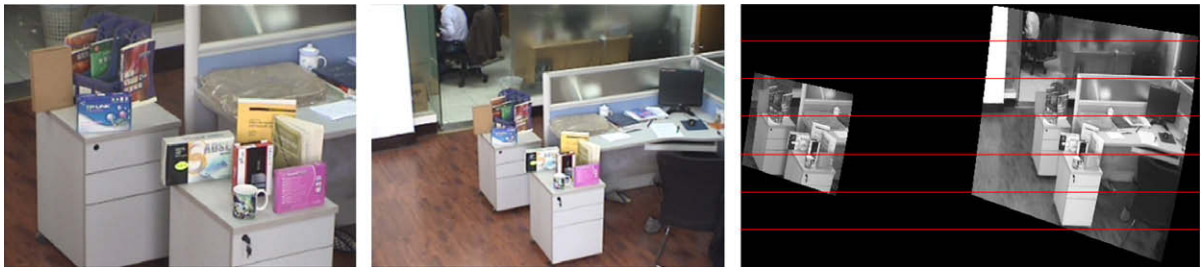

Fig. 8. A rectification result using the calibrated parameters.

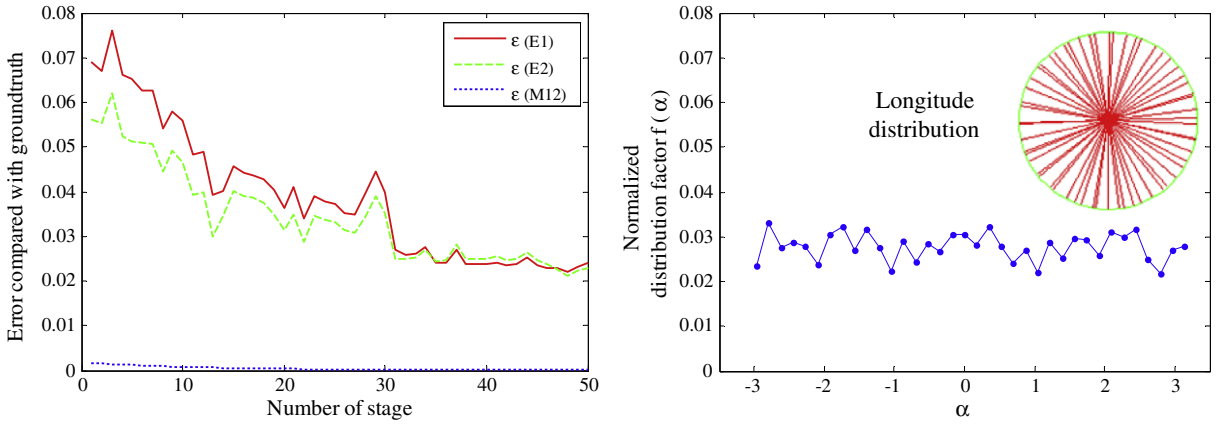

Fig. 9. Experiment on simulated data. The left image shows the difference compared with groundtruth, and the right image shows the final sample distribution.

lated data. This experiment is not perform stage by stage. We divide the α space into two parts: $[-\pi, 0)$ (part I) and $[0, \pi)$ (part II). We uniformly generate N_{I} and N_{II} samples from two parts, respectively; then, use all samples to estimate the calibration parameter:

(1) *Initialization:* we use the initialization method mentioned in Section 3.1 to generate θ^0. In this experiment, we assume all samples are inlier.
(2) *Optimization:* we use the iterative optimization algorithm mentioned in Section 3.2 to estimate the final θ.

This procedure is performed for 500 times independently, and we compare the mean and standard variance of the three kind of error mentioned before with given proportion $(N_{\mathrm{I}} : N_{\mathrm{II}})$ of samples from the two parts. Table 1 shows the result.

From this experiment, we can see that the distribution of samples could affect the accuracy of calibration. If the distribution is more uniform, the accuracy will be higher in general.

4.4. Anti-local-noise ability

In order to verify the performance of the integrated-parameters optimization method used in our study (see Section 3.2) is prior to that of the ordinal-parameters optimization method based on fundamental or essential matrix estimation, which is used in [10](i.e., use the fundamental or essential matrix to estimate E_i first, and then M_i) when all calibration samples are given, we compare these two methods on the simulated data with different σ (the standard variance of Gaussian noise added on X_i). This experiment could also reveal the anti-local-noise ability of both methods.

Table 1
Calibration error with different samples' distribution.

Sample distribution $(N_{\mathrm{I}} : N_{\mathrm{II}})$	$\epsilon(E_1)$ (Mean [SD])	$\epsilon(E_2)$ (Mean [SD])	$\epsilon(M_{12})$ (Mean [SD])
40:10	0.0225 [0.0120]	0.0223 [0.0119]	2.07 [1.58] $\times 10^{-4}$
10:40	0.0220 [0.0116]	0.0217 [0.0113]	1.97 [1.34] $\times 10^{-4}$
25:25	0.0200 [0.0108]	0.0199 [0.0104]	1.64 [1.25] $\times 10^{-4}$

In [10], E_i is computed from the fundamental matrix which is estimated by eight-point method [18]. In order to improve the performance, we use the five-point algorithm within a RANSAC scheme [16,17] to generate an initial (H_{21}^0) and also collect inliers. Since the samples, X_1^k and X_2^k, are normalized camera coordinates, we use the algebraic error criteria. Then the iterative algorithm [13] is applied to estimate H_{21} by minimizing the algebraic error subject to $\mathrm{rank}(H_{21}) = 2$ among all inliers.

We randomly generate 50 samples whose longitude values are uniformly distributed in $[-\pi, \pi)$. After the two methods finish estimating, we compute the three kinds of errors with respect to the groundtruth. For each σ, we independently run the estimation for 200 times, and record the mean and standard variance of errors in Table 2 ('method 1' indicates the integrated-parameters optimization, and 'method 2' indicates the ordinal-parameters optimization).

From this experiment, we can see that the integrated-parameters optimization have better performance than the ordinal-parameters optimization, especially for larger σ. When the local noise is very small, these two methods have similar accuracy. That is the reason that we use the ordinal-parameters optimization method in each stage of the proposed framework.

5. Conclusion

In this paper, we have proposed a novel self-calibration framework for spherical rectification model by using dual-PTZ-camera system. This framework works in an iterative way. In each stage,

Table 2
Anti-local-noise abilities.

σ	Optimization method	$\epsilon(E_1)$ (Mean [SD])	$\epsilon(E_2)$ (Mean [SD])	$\epsilon(M_{12})$ (Mean [SD])
0.0001	Method 1	$2.01\ [1.05] \times 10^{-3}$	$2.00\ [1.04] \times 10^{-3}$	$1.74\ [1.23] \times 10^{-5}$
	Method 2	$2.08\ [1.10] \times 10^{-3}$	$2.09\ [1.07] \times 10^{-3}$	$1.79\ [1.33] \times 10^{-5}$
0.001	Method 1	$2.03\ [1.09] \times 10^{-2}$	$1.96\ [1.07] \times 10^{-2}$	$1.67\ [1.35] \times 10^{-4}$
	Method 2	$2.11\ [1.15] \times 10^{-2}$	$2.01\ [1.10] \times 10^{-2}$	$1.73\ [1.37] \times 10^{-4}$
0.004	Method 1	$8.14\ [4.02] \times 10^{-2}$	$8.07\ [3.99] \times 10^{-2}$	$8.00\ [6.01] \times 10^{-4}$
	Method 2	$9.87\ [5.45] \times 10^{-2}$	$9.82\ [5.36] \times 10^{-2}$	$8.49\ [6.75] \times 10^{-4}$
0.007	Method 1	0.1281 [0.0716]	0.1232 [0.0725]	0.0015 [0.0020]
	Method 2	0.2878 [0.1800]	0.2843 [0.1817]	0.0631 [0.2727]
0.01	Method 1	0.1732 [0.1278]	0.1707 [0.1286]	0.0030 [0.0057]
	Method 2	0.8462 [0.7122]	0.8432 [0.7313]	0.3437 [0.4083]

the system evaluates a distribution of current calibration samples according to current estimated parameters. In order to make this distribution to be more uniform, we use the designed camera control strategy to collect suitable samples, so that the target parameters are likely to be refined. We also use a sample removing mechanism to remove samples which are thought to be outliers. So the accuracy of estimation can be improved stage by stage.

Since this framework can be performed automatically, it will be convenient for real application. Furthermore, the sample management mechanism and camera control strategy could ensure all calibration samples are well distributed and have high accuracy, so the calibration accuracy can be guaranteed. Experimental results also testify this conclusion.

References

[1] D. Wan, J. Zhou, Stereo vision using two PTZ cameras, Comput. Vision Image Understand. 112 (2) (2008) 184–194.
[2] D. Wan, J. Zhou, Multi-resolution and wide-scope depth estimation using a dual-PTZ-camera system, IEEE Trans. Image Process. 18 (3) (2009) 677–682.
[3] M.Z. Brown, D. Burschka, G.D. Hager, Advances in computational stereo, IEEE Trans. Pattern Anal. Mach. Intell. 25 (8) (2003) 993–1008.
[4] D. Papadimitriou, T. Dennis, Epipolar line estimation and rectification for stereo images pairs, IEEE Trans. Image Process. 3 (4) (1996) 672–676.
[5] M. Pollefeys, S.N. Sinha, Iso-disparity surfaces for general stereo configurations, in: ECCV, vol. 3, 2004, pp. 509–520.
[6] C.T. Loop, Z. Zhang, Computing rectifying homographies for stereo vision, in: CVPR, 1999, pp. 1125–1131.
[7] R.I. Hartley, Theory and practice of projective rectification, Int. J. Comput. Vision 35 (2) (1999) 115–127.
[8] S. Roy, J. Meunier, I.J. Cox, Cylindrical rectification to minimize epipolar distortion, in: CVPR, 1997, pp. 393–399.
[9] M. Pollefeys, R. Koch, L.J.V. Gool, A simple and efficient rectification method for general motion, in: ICCV, 1999, pp. 496–501.
[10] D. Wan, J. Zhou, D. Zhang, A spherical rectification for dual-PTZ-camera system, in: Proceedings of the ICASSP, vol. 1, 2007, pp. 777–780.
[11] S. Sinha, M. Pollefeys, Towards calibrating a pan–tilt–zoom cameras network, in: OMNIVIS 2004, ECCV Conference Workshop CD-ROM Proceedings, 2004.
[12] A. Senior, A. Hampapur, M. Lu, Acquiring multi-scale images by pan–tilt–zoom control and automatic multi-camera calibration, in: WACV, 2005, pp. 433–438.
[13] R.I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, second ed., Cambridge University Press, Cambridge, ISBN: 0521540518, 2004.
[14] J. Fujiki, A. Torii, S. Akaho, Epipolar geometry via rectification of spherical images, in: MIRAGE, 2007, pp. 461–471.
[15] D.G. Lowe, Distinctive image features from scale-invariant key points, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[16] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[17] D. Nistér, An efficient solution to the five-point relative pose problem, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 756–777.
[18] R.I. Hartley, In defense of the eight-point algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 19 (6) (1997) 580–593.