Exploiting Loops in the Graph of Trifocal Tensors for Calibrating a Network of Cameras

Jérôme Courchay[1], Arnak Dalalyan[1], Renaud Keriven[1], and Peter Sturm[2]

[1] IMAGINE, LIGM, Université Paris-Est
[2] Laboratoire Jean Kuntzmann, INRIA Grenoble Rhône-Alpes

Abstract. A technique for calibrating a network of perspective cameras based on their graph of trifocal tensors is presented. After estimating a set of reliable epipolar geometries, a parameterization of the graph of trifocal tensors is proposed in which each trifocal tensor is encoded by a 4-vector. The strength of this parameterization is that the homographies relating two adjacent trifocal tensors, as well as the projection matrices depend linearly on the parameters. A method for estimating these parameters in a global way benefiting from loops in the graph is developed. Experiments carried out on several real datasets demonstrate the efficiency of the proposed approach in distributing errors over the whole set of cameras.

1 Introduction

Camera calibration from images of a 3-dimensional scene has always been a central issue in Computer Vision. The success of textbooks like [1,2] attests this interest. In recent years, many methods for calibration have been proposed. Most of these work either rely on known or partially known internal calibrations [3,4,5,6,7,8,9,10] or deal with an ordered sequence of cameras [11,12,13,14]. In many practical situations, however, the internal parameters of cameras are unavailable or available but very inaccurate. The absence of an order in the set of cameras is also very common when processing, for instance, Internet images.

In this paper, we deal with the problem of calibrating a network of cameras from a set of unordered images, the main emphasis being on the accuracy of the projective reconstruction of camera matrices. Traditionally, this situation is handled by factorizing the measurement matrix [15,16], which may be subject to missing data [17,18] because of occlusions. The methodology adopted in the present work is substantially different and is based on the notion of the graph of trifocal tensors rather than on the factorization. The experiments on real datasets show that our approach leads to highly competitive results that furnish a good initialization to the bundle adjustment (BA) algorithm [19].

Even in the case of calibrated cameras, most of the aforementioned methods are based on a graph of cameras (in which the edges are the epipolar geometries) which is made acyclic by discarding several edges. On the other hand, a number of recent studies, oriented toward city modeling from car or aerial sequences, point out the benefits of enforcing loop constraints. Considering loops in the graph of cameras has the advantage of reducing the drift due to errors induced while processing the trajectory sequentially

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part II, LNCS 6312, pp. 85–99, 2010.

(cf. Fig. 1). [20] merges partial reconstructions, [21] constrains coherent rotations for loops and planar motion. Adapted to their specific input, these papers often rely on trajectory regularization or dense matching [22,23]. [24] is a notable exception, where loop constraints are added to sparse Structure from Motion (SfM), yet taking as input an ordered omnidirectional sequence and assuming known internal parameters. The

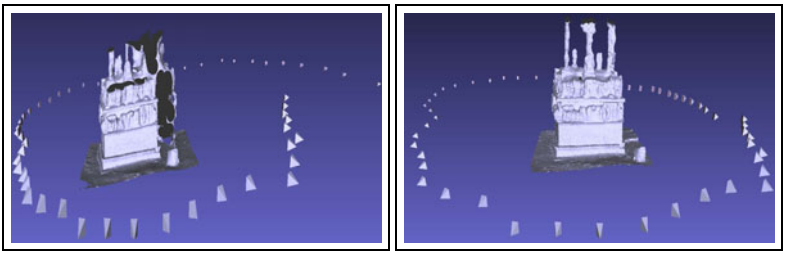

Fig. 1. Multi-view stereo reconstruction [25] using cameras calibrated without (left) and with (right) using the loop-constraint. When the loop constraint is not enforced, the accumulation of errors results in an extremely poor reconstruction.

method proposed in the present work consists of the following points:

- Our starting point is a set of unknown cameras linked by estimated epipolar geometries (EG). These are computed using a state-of-the-art version of RANSAC [26], followed by a maximum likelihood improvement described in [13]. We assume that along with the estimated fundamental matrices, reliable epipolar correspondences are known. These correspondences are made robust by simultaneously considering several camera pairs, like in [3]. This produces a set of three-view correspondences that will be used in the sequel.
- We group views into triplets. Three views (i, j, k) are considered as a valid triplet if (a) the EGs between i and j as well as between j and k have been successfully computed at the previous step and (b) there are at least 4 three-view correspondences in these images. To reduce the number of nodes, some of the estimated epipolar geometries are ignored, so that inside a triplet, only two of the three fundamental matrices are considered known. The advantage of this strategy is that we do not need to enforce the coherence of fundamental matrices. At first sight, this can be seen as a loss of information. However, this information is actually recovered via trifocal tensors.
- We define a graph having as nodes valid camera triplets. Therefore, there are two fundamental matrices available for each node. Two nodes are connected by an edge if they share a fundamental matrix. We demonstrate that for each node there exists a 4-vector such that all the entries of the three camera matrices are affine functions of this 4-vector with known coefficients. Moreover, the homographies that allow the registration of two adjacent nodes ν and ν' are also affine functions depending on 4 out of the 8 unknown parameters corresponding to ν and ν'. To speed-up the computations, for each node only 50 (or less) three-view correspondences that are the most compatible with the EGs are used.

– If the graph of triplets is acyclic, the equations of three-view correspondences for all nodes lead to a linear estimate of all the cameras. In case, the graph of triplets contains one or several loops, each loop is encoded as a (non-linear) constraint on the unknowns. Starting from an initial value computed as a solution to the unconstrained least squares, we sequentially linearize the loop constraints and solve the resulting problem by (sparse) linear programming. This can be efficiently done even for very large graphs. It converges very rapidly, but the loop constraints are fulfilled only approximately.
– In the case where the loop constraints are not satisfied exactly, we proceed by homography registration and estimation of camera matrices by linear least-squares under norm constraint. This is done exactly via a singular value decomposition producing as output all cameras in a projective space. To provide a qualitative evaluation, we recover the metric space using an implementation of [27], and a single Euclidean bundle adjustment that refines the metric space and camera positions.

Thus, we propose a method that accurately recovers geometries, without any sequential process, and attempts to enforce the compatibility of cameras within loops in the early stages of the procedure. An important advantage conferred by our approach is that the number of unknown parameters is kept fairly small, since we consider only the cameras (four unknowns for each triplet) and not the 3D points. Our reconstruction is further refined by bundle adjustment. Taking loops into account and avoiding error accumulation, the proposed solution is less prone to get stuck in local minima.

The remainder of the paper is organized as follows. Section 2 presents the background theory and terminology. Our algorithm is thoroughly described in Sections 3 and 4. The results of numerical experiments conducted on several real datasets as well as a comparison to state-of-the-art software is provided in Section 5. A discussion concludes the paper.

2 Background

In this work, we consider a network of N uncalibrated cameras and assume that for some pairs of cameras (i, j), where $i, j = 1, \ldots, N$, $i \neq j$, an estimation of the fundamental matrix, denoted by $\mathtt{F}^{ij}$, is available. Let us denote by $\mathbf{e}^{ij}$ the unit norm epipole in view j of camera center i. Recall that the fundamental matrix leads to a projective reconstruction of camera matrices $(\mathtt{P}^i, \mathtt{P}^j)$, which is unique up to a homography.

The geometry of three views i, j and k is described by the Trifocal Tensor, hereafter denoted by $\mathcal{T}^{ijk}$. It consists of three 3×3 matrices: $\mathtt{T}_1^{ijk}$,$\mathtt{T}_2^{ijk}$ and $\mathtt{T}_3^{ijk}$ and provides a particularly elegant description of point-line-line correspondences in terms of linear equations

$$\mathbf{p}_i^\mathsf{T} \begin{bmatrix} \mathbf{l}_j^\mathsf{T} \mathtt{T}_1^{ijk} \\ \mathbf{l}_j^\mathsf{T} \mathtt{T}_2^{ijk} \\ \mathbf{l}_j^\mathsf{T} \mathtt{T}_3^{ijk} \end{bmatrix} \mathbf{l}_k = 0, \tag{1}$$

where $\mathbf{p}_i$ is a point in image i (seen as a point in projective space $\mathbb{P}^2$) which is in correspondence with the line $\mathbf{l}_j$ in image j and with the line $\mathbf{l}_k$ in image k. Considering

the entries of $\mathcal{T}^{ijk}$ as unknowns, we get thus one linear equation for each point-line-line correspondence. Therefore, one point-point-point correspondence $\mathbf{p}_i \longleftrightarrow \mathbf{p}_j \longleftrightarrow \mathbf{p}_k$ leads to 4 independent linear equations by combining an independent pair of lines passing through $\mathbf{p}_j$ in image j with an independent pair of lines passing through $\mathbf{p}_k$ in image k.

Since a Trifocal Tensor has 27 entries, the previous argument shows that 7 point-point-point correspondences suffice for recovering the Trifocal Tensor as a solution of an overdetermined system of linear equations. Recall however that the Trifocal Tensor has only 18 degrees of freedom. Most algorithms estimating a Trifocal Tensor from noisy point-point-point correspondences compute an approximate solution to the linear system by a least squares estimator (LSE) and then perform a post-processing in order to get a valid Trifocal Tensor. An alternative approach consists in using a minimal solution that determines the three-view geometry from six points [28,29].

2.1 Main Ingredients of Our Approach

Let us describe now two elementary results that represent the building blocks of our approach, relying on the fact that when two out of three fundamental matrices are known, the Trifocal Tensor has exactly 4 degrees of freedom.

Proposition 1. *For three views i, j and k, given two fundamental matrices $\mathtt{F}^{ij}$ and $\mathtt{F}^{ik}$, there exists a 4-vector $\boldsymbol{\gamma} = [\gamma_0, \ldots, \gamma_3]$ such that $\mathcal{T}^{ijk}$ is given by:*

$$T_t^{ijk} = A_t^{ij} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \gamma_0 & \gamma_t \end{bmatrix} \left(A_t^{ik}\right)^{\mathsf{T}} \tag{2}$$

for every $t = 1, 2, 3$, where $A_t^{is} = \left[(F_{t,1:3}^{is})^{\mathsf{T}} \ , \ (F_{t,1:3}^{is})^{\mathsf{T}} \times \mathbf{e}^{is} \ , \ \mathbf{e}^{is}\right]$, for $s = j, k$. Moreover, $\mathcal{T}^{ijk}$ is geometrically valid, i.e. , there exist 3 camera matrices $\mathtt{P}^i$, $\mathtt{P}^j$ and $\mathtt{P}^k$ compatible with $\mathtt{F}^{ij}$ and $\mathtt{F}^{ik}$ and having $\mathcal{T}^{ijk}$ as the Trifocal Tensor.

The proof of this result is deferred to the supplemental material. It is noteworthy that the claims of Proposition 1 hold true under full generality, even if the centers of three cameras are collinear. In view of [1], the camera matrices parameterized by $\boldsymbol{\gamma}$ that are compatible with the fundamental matrices $\mathtt{F}^{ij}$ and $\mathtt{F}^{ik}$ as well as with the Trifocal Tensor defined by Eq. 2 are given by (up to a projective homography)

$$\begin{aligned} \mathtt{P}^i &= [\mathtt{I}_{3\times 3} \,|\, \mathbf{0}_{3\times 1}], \qquad \mathtt{P}^k = [\gamma_0 [\mathbf{e}^{ik}]_\times \mathtt{F}^{ki} \,|\, \mathbf{e}^{ik}], \\ \mathtt{P}^j &= \mathrm{kron}\big([\boldsymbol{\gamma}_{1:3}, 1]; \mathbf{e}^{ij}\big) - [[\mathbf{e}^{ij}]_\times \mathtt{F}^{ji} \,|\, \mathbf{0}_{3\times 1}], \end{aligned} \tag{3}$$

where $\mathrm{kron}(\cdot, \cdot)$ stands for the Kronecker product of two matrices.

In the noiseless setting, Proposition 1 offers a minimal way of computing the 4 remaining unknowns from point-point-point correspondences. One could think that one point-point-point correspondence leading to 4 equations is enough for retrieving the 4 unknowns. However, since two EGs are known, only one equation brings new information from one point-point-point correspondence. So we need at least 4 point-point-point

correspondences to compute the Trifocal Tensor compatible with the two given fundamental matrices. In the noisy case, if we use all 4 equations associated to point-point-point correspondences, the system is then overdetermined and one usually proceeds by computing the LSE.

The second ingredient in our approach is the parameterization of the homography that bridges two camera triplets having one fundamental matrix in common. Let i, j, k and ℓ be four views such that (a) for views i and k we have successfully estimated the fundamental matrix $\mathtt{F}^{ik}$ and (b) for each triplet (i,j,k) and (k,i,ℓ) the estimates of two fundamental matrices are available. Thus, the triplets (i,j,k) and (k,i,ℓ) share the same fundamental matrix $\mathtt{F}^{ik}$. Using equations (3), one obtains two projective reconstructions of camera matrices of views i and j based on two 4-vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$. Let us denote the reconstruction from the triplet (i,j,k) (resp. (k,i,ℓ)) by $\mathtt{P}^i_{\boldsymbol{\gamma}}$ and $\mathtt{P}^k_{\boldsymbol{\gamma}}$ (resp. $\mathtt{P}^i_{\boldsymbol{\gamma}'}$ and $\mathtt{P}^k_{\boldsymbol{\gamma}'}$). If the centers of cameras i and k differ, then there is a unique homography $\mathtt{H}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$ such that

$$\mathtt{P}^i_{\boldsymbol{\gamma}}\mathtt{H}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'} \cong \mathtt{P}^i_{\boldsymbol{\gamma}'}, \quad \mathtt{P}^k_{\boldsymbol{\gamma}}\mathtt{H}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'} \cong \mathtt{P}^k_{\boldsymbol{\gamma}'}, \tag{4}$$

where $\cong$ denotes equality up to a scale factor. Considering the camera matrices as known, one can solve (4) w.r.t. $\mathtt{H}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$. One readily checks that[1]

$$\mathtt{H}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'} = \left[\begin{array}{c|c} \mathrm{kron}(\boldsymbol{\gamma}'_{1:3}, \mathbf{e}^{ki}) - [\mathbf{e}^{ki}]_\times \mathtt{F}^{ik} & \mathbf{e}^{ki} \\ \hline -\frac{\gamma_0}{2}\,\mathrm{tr}([\mathbf{e}^{ik}]_\times \mathtt{F}^{ki}[\mathbf{e}^{ki}]_\times \mathtt{F}^{ik})(\mathbf{e}^{ik})^\mathsf{T} & 0 \end{array}\right]. \tag{5}$$

To sum up this section, let us stress that the main message to retain from all these formulas is that $\mathtt{H}_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$, as well as the camera matrices (3) are linear in $(\boldsymbol{\gamma},\boldsymbol{\gamma}')$.

3 Estimating Tensors by Sequential Linear Programming

This section contains the core of our contribution which is based on a graph-based representation of the triplets of cameras. This is closely related to the framework developed in [5], where the graph of camera pairs is considered. The advantage of operating with triplets instead of pairs is that there is no need to distinguish between feasible and infeasible paths.

3.1 Graph of Trifocal Tensors

The starting point for our algorithm is a set of estimated EGs that allow us to define a graph $\mathcal{G}_{\text{cam}}$ so that (a) $\mathcal{G}_{\text{cam}}$ has N nodes corresponding to the N cameras and (b) two nodes of $\mathcal{G}_{\text{cam}}$ are connected by an edge if a reliable estimation of the corresponding epipolar geometry is available. Then, a triplet of nodes i,j,k of $\mathcal{G}_{\text{cam}}$ is called valid if (a) there is a sufficient number of three view correspondences between i, j and k, and (b) at least two out of three pairs of nodes are adjacent in $\mathcal{G}_{\text{cam}}$.

If for some valid triplet all three EGs are available, we remove the least reliable one and define the graph $\mathcal{G}_{\text{triplet}} = (\mathcal{V}_{\text{triplet}}, \mathcal{E}_{\text{triplet}})$ having as nodes valid triplets of cameras

[1] See supplemental material for more details.

and as edges the pairs of triplets that have one fundamental matrix in common. In view of Proposition 1, the global calibration of the network is equivalent to the estimation of a 4-vector for each triplet of cameras. Thus, to each node v of the graph of triplets we associate a vector $\boldsymbol{\gamma}^v \in \mathbb{R}^4$. The large vector $\Gamma = (\boldsymbol{\gamma}^v : v \in \mathcal{V}_{\text{triplet}})$ is the parameter of interest in our framework.

If, by some chance, it turns out that the graph of triplets is acyclic, then the problem of estimating Γ reduces to estimating $N_V = \text{Card}(\mathcal{V}_{\text{triplet}})$ independent vectors $\boldsymbol{\gamma}^v$. This task can be effectively accomplished using point-point-point correspondences and the equation (1). As explained in Section 2, a few point-point-point correspondences suffice for computing an estimator of $\boldsymbol{\gamma}^v$ by least squares. In our implementation, we use RANSAC with a minimal configuration of four 3-view correspondences in order to perform robust estimation.

3.2 Calibration as Constrained Optimization

However, acyclic graphs are the exception rather than the rule. Even if the camera graph is acyclic, the resulting triplet graph may contain loops. To explain how the loops in the graph $\mathcal{G}_{\text{triplet}}$ are handled, let us remark that one can associate a homography (cf. (5)) to each adjacent pair (v, v') of nodes of $\mathcal{G}_{\text{triplet}}$. Using these homographies, each loop of the graph of triplets yields a constraint on the homographies and, therefore, on the parameter vector Γ. For instance, the 3-loop $v \longleftrightarrow v' \longleftrightarrow v'' \longleftrightarrow v$ gives rise to the constraint $\mathtt{H}_{\boldsymbol{\gamma}^v,\boldsymbol{\gamma}^{v'}}\mathtt{H}_{\boldsymbol{\gamma}^{v'},\boldsymbol{\gamma}^{v''}}\mathtt{H}_{\boldsymbol{\gamma}^{v''},\boldsymbol{\gamma}^{v}} \cong \mathtt{I}_{4\times 4}$. This equation defines a set of 15 polynomial constraints on the unknown vector Γ. If the triplet graph contains N_{loop} loops, then we end up with $15N_{\text{loop}}$ constraints. Our proposal—in the case of general graphs of triplets—is to estimate Γ by minimizing an energy derived from the equations (1) and point-point-point correspondences (similarly to the LSE proposed in the previous subsection) subject to $15N_{\text{loop}}$ constraints.

The main advantage of this approach is that if a solution to the proposed optimization problem is found, it is guaranteed to be consistent w.r.t. the loops, meaning that each camera matrix will be uniquely determined up to a scale factor and an overall homography ambiguity.

3.3 Sequential Linear Programming

Instead of solving the optimization problem that is obtained by combining the LSE with the loop-constraints, we propose here to replace it by a linear program. To give more details, let us remark that every loop-constraint can be rewritten as $f_j(\Gamma) = 0$, $j = 1, \ldots, 15$, for some polynomial functions f_j. Gathering these constraints for all N_{loop} loops, we get

$$f_j(\Gamma) = 0, \quad j = 1, \ldots, 15N_{\text{loop}}. \tag{6}$$

On the other hand, in view of (1) and (2), the point-point-point correspondences can be expressed as an inhomogeneous linear equation system in Γ

$$\mathtt{M}\Gamma = \mathbf{m}, \tag{7}$$

where $\mathtt{M}$ is a $4N_{\text{3-corr}} \times 4N$ matrix and $\mathbf{m}$ is a $4N_{\text{3-corr}}$ vector with $N_{\text{3-corr}}$ being the number of correspondences across three views. The matrix $\mathtt{M}$ and the vector $\mathbf{m}$ are computed using the known fundamental matrices. Since in practice these matrices are estimated from available data, the system (7) need not be satisfied exactly. Then, it is natural to estimate the parameter-vector $\mathit{\Gamma}$ by solving the problem

$$\min \|\mathtt{M}\mathit{\Gamma} - \mathbf{m}\|_q^q \qquad \text{subject to} \quad f_j(\mathit{\Gamma}) = 0,\ \forall j = 1, \ldots, 15N_{\text{loop}}, \tag{8}$$

for some $q \geq 1$. Unfortunately, there is no q for which this problem is convex and, therefore, it is very hard to solve. To cope with this issue, we propose a strategy based on local linearization.

We start by computing an initial estimator of $\mathit{\Gamma}$, *e.g.*, by solving the unconstrained (convex) problem with some $q \geq 1$. In our implementation, we use RANSAC with $q = 2$ for ensuring robustness to erroneous three-view correspondences. Then, given an initial estimator $\mathit{\Gamma}_0$, we define the sequence $\mathit{\Gamma}_k$ by the following recursive relation: $\mathit{\Gamma}_{k+1}$ is the solution to the linear program

$$\min \|\mathtt{M}\mathit{\Gamma} - \mathbf{m}\|_1 \qquad \text{subject to} \quad |f_j(\mathit{\Gamma}_k) + \nabla f_j(\mathit{\Gamma}_k)(\mathit{\Gamma} - \mathit{\Gamma}_k)| \leq \epsilon, \tag{9}$$

where ϵ is a small parameter (we use $\epsilon = 10^{-6}$). There are many softwares—such as GLPK, SeDuMi, SDP3—for solving problem (9) with highly attractive execution times even for thousands of constraints and variables. Furthermore, empirical experience shows that the sequence $\mathit{\Gamma}_k$ converges very rapidly. Typically, a solution with satisfactory accuracy is obtained after five to ten iterations.

3.4 Accounting for Heteroscedasticity

The goal now is to make the energy that we minimize in (9), which is purely algebraic, meaningful from a statistical viewpoint. Assume equations (7) are satisfied up to an additive random noise: $\mathtt{M}\mathit{\Gamma} = \mathbf{m} + \boldsymbol{\xi}$, where the random vector $\boldsymbol{\xi}$ has independent coordinates drawn from the centered Laplace distribution with constant scale. Then the energy in (9) is proportional to the negative log-likelihood. The constancy of the scale factor means that the errors are homoscedastic, which is a very strong hypothesis. We observed that all three view correspondences recorded by a fixed triplet have nearly the same scale for the errors, while the scales for different triplets are highly variable. To account for this heteroscedasticity of the noise, we use the initial estimator of $\mathit{\Gamma}$ to estimate one scale parameter σ_v per node $v \in \mathcal{V}_{\text{triplet}}$. This is done by computing the standard deviation of the estimated residuals. Using $\{\sigma_v\}$, the energy in problem (9) is replaced by $\sum_v \|\mathtt{M}_v\mathit{\Gamma} - \mathbf{m}_v\|_1/\sigma_v$. Here, $\mathtt{M}_v$ is the submatrix of $\mathtt{M}$ containing only those rows that are obtained from three-view correspondences recorded by v. The vector $\mathbf{m}_v$ is obtained from $\mathbf{m}$ in the same way.

4 Homography Registration and Estimation of Projection Matrices

Assume that we have a graph of trifocal tensors, $\mathcal{G}_{\text{triplet}}$, each node of which will be denoted by $v_1, v_2, \ldots, v_n$. In the previous step, we have determined parameters $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$,

such that $\boldsymbol{\gamma}_i$ characterizes the trifocal tensor represented by v_i. A naive strategy for estimating camera matrices is to set one of the cameras equal to $[\mathtt{I}_{3\times 3} \,|\, \mathbf{0}_{3\times 1}]$ and to recover the other cameras by successive applications of the homographies $\mathtt{H}_{\gamma,\gamma'}$ to the camera matrices reconstructed according to (3). However, in general situations, the vector Γ computed by sequential linear programming as described in the previous section satisfies the loop constraints up to a small error. Therefore, the aforementioned naive strategy has the drawback of increasing the error of estimation for cameras computed using many homographies $\mathtt{H}_{\gamma,\gamma'}$. In order to avoid this and to uniformly distribute the estimation error over the set of camera matrices, we propose a method based on homography registration by SVD. Thus, the input for the method described in this section is a vector Γ for which the loop constraints are satisfied up to a small estimation error.

4.1 The Case of a Single Loop

We assume in this subsection that $\mathcal{G}_{\text{triplet}}$ reduces to one loop, that is each node v_i has exactly two neighbors v_{i-1} and v_{i+1} with standard convention and $v_{n+i} = v_i$ for all i. (This applies to all the indices in this subsection.) For each node v_i representing three views, we have already computed a version of the projection matrices $\mathtt{P}^{1,\boldsymbol{\gamma}_i}, \mathtt{P}^{2,\boldsymbol{\gamma}_i}, \mathtt{P}^{3,\boldsymbol{\gamma}_i}$. Furthermore, for two neighboring nodes v_i and v_{i+1} we have computed a homography $\mathtt{H}^{i,i+1}$ so that $\mathtt{P}^{j,\boldsymbol{\gamma}_{i+1}} \cong \mathtt{P}^{j+1,\boldsymbol{\gamma}_i}\mathtt{H}^{i,i+1}$, $j \in \{1,2\}$. Based on the relative homographies $\{\mathtt{H}^{i,i+1}\}$ we want to recover absolute homographies $\mathtt{H}^{v_i}$ that allow to represent all the matrices $\mathtt{P}^{j,\boldsymbol{\gamma}_i}$ in a common projective frame. In other terms, in the ideal case where there is no estimation error, the matrices $\mathtt{H}^{v_i}$ should satisfy

$$\mathtt{P}^{j,\boldsymbol{\gamma}_i}\,\mathtt{H}^{v_i} \cong \mathtt{P}^{j+i-1,*}, \qquad j \in \{1,2,3\}. \tag{10}$$

Obviously, the set $\{\mathtt{H}^{v_i}\}$ can only be determined up to an overall projective homography.

Proposition 2. *If for some $i = 1,\ldots,n$, the cameras $\mathtt{P}^{i+1,*}$ and $\mathtt{P}^{i+2,*}$ have different centers, then $\mathtt{H}^{v_i} \cong \mathtt{H}^{i,i+1}\mathtt{H}^{v_{i+1}}$. Furthermore, if the centers of each pair of consecutive cameras are different, then one can find a projective coordinate frame so that*

i) $\mathtt{H}^{v_i} = \mathtt{H}^{i,i+1}\mathtt{H}^{v_{i+1}}, \qquad \forall i = 1,\ldots,n-1,$

ii) $\alpha\,\mathtt{H}^{v_n} = \mathtt{H}^{n,1}\mathtt{H}^{v_1}$, *where α can be determined by* $\alpha = \frac{1}{4}\text{Trace}(\prod_{i=1}^{n}\mathtt{H}^{i,i+1})$,

iii) Let $\bar{\mathtt{H}}$ be the $(4n)\times 4$ matrix resulting from the vertical concatenation of matrices $\mathtt{H}^{v_i}$. The four columns of $\bar{\mathtt{H}}$ are orthonormal.

This result, the proof of which is presented in the supplemental material, allows us to define the following algorithm for estimating the matrices $\{\mathtt{H}^{v_i}\}$. Given the relative homographies $\{\mathtt{H}^{i,i+1}\}$, we first compute α according to the formula in ii) and then minimize the cost function

$$\sum_{i=1}^{n-1}\frac{\|\mathtt{H}^{v_i} - \mathtt{H}^{i,i+1}\mathtt{H}^{v_{i+1}}\|_2^2}{\max(\sigma_{v_i}^2,\sigma_{v_{i+1}}^2)} + \frac{\|\alpha\,\mathtt{H}^{v_n} - \mathtt{H}^{n,1}\mathtt{H}^{v_1}\|_2^2}{\max(\sigma_{v_1}^2,\sigma_{v_n}^2)} \tag{11}$$

w.r.t. $\{\mathtt{H}^{v_i}\}$, subject to the orthonormality of the columns of $\bar{\mathtt{H}}$. Here, $\|\cdot\|_2$ is the Frobenius norm. The exact solution of this (non-convex) optimization problem can be computed using the singular value decomposition of a matrix of size $4n\times 4n$ constructed

First frame Second frame Last frame

Fig. 2. This figure illustrates the improvement achieved at each step of our algorithm. If the cameras are reconstructed without imposing loop constraints, the epipolar lines between the first and the last frames are extremely inaccurate (1st row). They become much more accurate when the constrained optimization is performed (2nd row). Finally, the result is almost perfect once the homography registration is done.

from α and $\{\mathtt{H}^{i,i+1}\}$. Since this is quite standard (based on the Courant-Fisher minimax theorem [30, Thm. 8.1.2]), we do not present more details here.

4.2 The Case of Several Loops

Assume now that we have identified several loops in the graph of trifocal tensors. Let N_{loop} be the number of these loops. We apply to each loop the method of the previous section and get a homography for every node of the loop. In general, one node of $\mathcal{G}_{\text{triplet}}$ may lie in several loops, in which case we will have several homographies for that node. It is then necessary to enforce the coherence of these homographies. To this end, we define the graph $\mathcal{G}_{\text{loop}}$ having N_{loop} nodes, each node representing a loop. Two nodes of $\mathcal{G}_{\text{loop}}$ are linked by an edge, if the corresponding loops have non-empty intersection. We will assume that the graph $\mathcal{G}_{\text{loop}}$ is connected, since otherwise it is impossible to simultaneously calibrate different connected components.

The next step consists in determining a minimal depth spanning tree $\mathcal{T}_{\text{loop}}$ of $\mathcal{G}_{\text{loop}}$. Since the number of loops is assumed small, this step will not be time consuming. Let $(\mathcal{L}, \mathcal{L}')$ be a pair of adjacent nodes of $\mathcal{T}_{\text{loop}}$. By an argument analogous to that of Proposition 2, one can show that there exists a 4×4 homography $\mathtt{H}^{\mathcal{L},\mathcal{L}'}$ such that $\mathtt{H}^{v,\mathcal{L}} \cong \mathtt{H}^{v,\mathcal{L}'}\mathtt{H}^{\mathcal{L}',\mathcal{L}}$ up to an estimation error, for every triplet of cameras $v \in \mathcal{L} \cap \mathcal{L}'$. Here, $\mathtt{H}^{v,\mathcal{L}}$ (resp. $\mathtt{H}^{v,\mathcal{L}'}$) stands for the homography assigned (*cf.* previous subsection) to the triplet v as a part of the loop $\mathcal{L}$ (resp. $\mathcal{L}'$). The homography $\mathtt{H}^{\mathcal{L}',\mathcal{L}}$ can be estimated by minimizing the objective function

$$\sum_{v \in \mathcal{L} \cap \mathcal{L}'} \|\alpha_v \mathtt{H}^{v,\mathcal{L}} - \mathtt{H}^{v,\mathcal{L}'} \mathtt{H}^{\mathcal{L}',\mathcal{L}}\|_2^2 / \sigma_v^2 \tag{12}$$

w.r.t. the matrix $\mathtt{H}^{\mathcal{L}',\mathcal{L}}$ and parameters $\{\alpha_v\}$ subject to $\|\mathtt{H}^{\mathcal{L}',\mathcal{L}}\|_2^2 + \sum_{v \in \mathcal{L} \cap \mathcal{L}'} \alpha_v^2 = 1$. Once again, this minimization can be carried out by computing the eigenvector corresponding to the smallest singular value of a suitably defined matrix.

Finally, to enforce the coherence of absolute homographies computed using different loops, we proceed as follows. We do not modify the homographies computed within the loop $\mathcal{L}_0$ constituting the root of the minimal depth spanning tree $\mathcal{T}_{\text{loop}}$. For any other loop $\mathcal{L}$, let $\mathcal{L}_0 \to \mathcal{L}_1 \to \ldots \to \mathcal{L}_k \to \mathcal{L}$ be the (unique) path joining $\mathcal{L}$ to the root. Then, every absolute homography $\mathtt{H}^{v,\mathcal{L}}$, $v \in \mathcal{L}$, computed within the loop $\mathcal{L}$ using the method of the previous subsection is replaced by $\mathtt{H}^{v,\mathcal{L}} \mathtt{H}^{\mathcal{L},\mathcal{L}_k} \cdots \mathtt{H}^{\mathcal{L}_1,\mathcal{L}_0}$. After this modification, the images by $\mathtt{H}^{v,\mathcal{L}}$ of the projection matrices $\mathtt{P}^{j,\gamma_v}$ $(j = 1, 2, 3)$ will all lie in nearly the same projective space. This makes it possible to recover the final projection matrices $\mathtt{P}^i$ by a simple computation presented in the next subsection.

4.3 Estimating Projection Matrices

Once the set of absolute homographies estimated, we turn to the estimation of camera matrices $\{\mathtt{P}^{j,*}\}$. Due to the estimates computed in previous steps, each projection matrix $\mathtt{P}^{j,*}$ can be estimated independently of the others. To ease notation and since there is no loss of generality, let us focus on the estimation of $\mathtt{P}^{1,*}$. We start by determining the nodes in $\mathcal{G}_{\text{triplet}}$ that contain the first view. Let $\mathcal{V}_1$ denote the set of these nodes. To each node $v \in \mathcal{V}_1$ corresponds one estimator of $\mathtt{P}^{1,*}$, denoted by $\mathtt{P}^{1,\gamma_v}$. Furthermore, we have a set of estimated homographies $\mathtt{H}^{v,\mathcal{L}}$ that satisfy, up to an estimation error, the relation $\mathtt{P}^{1,v}\mathtt{H}^{v,\mathcal{L}} \cong \mathtt{P}^{1,*}$. This is equivalent to $\alpha_{v,\mathcal{L}}\mathtt{P}^{1,v}\mathtt{H}^{v,\mathcal{L}} = \mathtt{P}^{1,*}$, $\forall v \in \mathcal{V}_1$, $\forall \mathcal{L} \supset \{v\}$ with some $\alpha_{v,\mathcal{L}} \in \mathbb{R}$. In these equations, the unknowns are the reals $\alpha_{v,\mathcal{L}}$ and the matrix $\mathtt{P}^{1,*}$. Since this matrix should be of rank 3, it has nonzero Frobenius norm. Therefore, we estimate $\mathtt{P}^{1,*}$ by $\mathtt{P}^1$ defined as a solution to

$$\arg\min_{\mathtt{P}} \min_{\{\alpha_{v,\mathcal{L}}\}:\|\mathtt{P}\|_2^2+\|\boldsymbol{\alpha}\|_2^2=1} \sum_{\mathcal{L}} \sum_{v \in \mathcal{L} \cap \mathcal{V}_1} \|\alpha_{v,\mathcal{L}}\mathtt{P}^{1,v}\mathtt{H}^{v,\mathcal{L}} - \mathtt{P}\|_2^2 / \sigma_v^2, \tag{13}$$

where $\boldsymbol{\alpha}$ stands for the vector having as coordinates the numbers $\alpha_{v,\mathcal{L}}$. Once again, the problem (13) can be explicitly solved using the SVD of an appropriate matrix.

5 Experiments

Implementation. In order to apply the methodology we have just described, we extract and match SIFT [31] descriptors from all the images. Then, epipolar geometries are estimated by DEGENSAC [32]. Note that some speed-up in this step can be achieved by using one of the recent versions of RANSAC [26,33]. Estimated EGs allow us to identify and remove wrong correspondences as well as to create feature tracks. Using these tracks and EGs as input for our algorithm, we compute as output the projection matrices of all the cameras. In order to be able to visually assess the reconstruction quality, all cameras and the 3D structure are upgraded to Euclidean [27].

Table 1. Characteristics of the datasets used for the experimental validation. From left to right: number of frames in each sequence, the resolution of each image, the number of 2D image points used for the final BA for our method and for bundler [5], the mean squared reprojection error.

Dataset	**#frames**	**resolution**	**# image points**		**MSRE (pxl)**	
			Our	Bundler	Our	Bundler
Dinosaur	36	576 × 720	45,250	37,860	0.27	0.25
Temple	45	480 × 500	26,535	23,761	0.08	0.11
Fountain P11	11	2048 × 3072	57,547	23,648	0.16	0.13
Herz-Jesu R23	23	2048 × 3072	129,803	–	0.41	–
Detenice	34	1536 × 2048	30,200	–	0.15	–
Calvary	52	2624 × 3972	54,798	–	0.51	–

Fig. 3. One frame of each dataset used to test our methodology. From left to right: dinosaur, temple, fountain P11, Herz-Jesu R23 [34], Calvary, Detenice fountain.

Datasets. We tested our methodology on six datasets: the *dinosaur* sequence (36 frames), the *temple* sequence (45 frames), the *fountain P11* sequence (11 frames), the *Herz-Jesu R23* sequence (23 frames), the *Detenice fountain* sequence (34 frames) and the *calvary* sequence (52 frames). For the first three datasets, the ground truth of camera matrices is available on the Web.

Quality measures. Since the main contribution of the present paper concerns the projective reconstruction, it is natural to assess the quality of the proposed approach using the distance:

$$d_{proj}(\{\mathtt{P}^j\},\{\mathtt{P}^{j,*}\}) = \inf_{\boldsymbol{\alpha},\mathtt{H}} \sum_{j=1}^{n} \|\alpha_j \mathtt{P}^j \mathtt{H} - \mathtt{P}^{j,*}\|_2^2, \tag{14}$$

where $\mathtt{P}^j$ and $\mathtt{P}^{j,*}$ are respectively the reconstructed and the true camera projection matrices, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is a vector of real numbers and $\mathtt{H}$ is a 3D-homography. Naturally, this measure can be used only on sequences for which the ground truth is available. Note also that the computation of the infimum in (14) is a non-convex optimization problem. We solve it by first computing the one-norm solution to the least squares problem $\min_{\boldsymbol{\alpha},\mathtt{H}} \sum_{j=1}^{n} \|\mathtt{P}^j\mathtt{H} - \alpha_j^{-1}\mathtt{P}^{j,*}\|_2^2$, and then use this solution as a starting point for an alternating minimization. For the examples considered here, this converges very rapidly and, since the results are good, we believe that the local minimum we find is in fact a global minimum, or at least not too far from it.

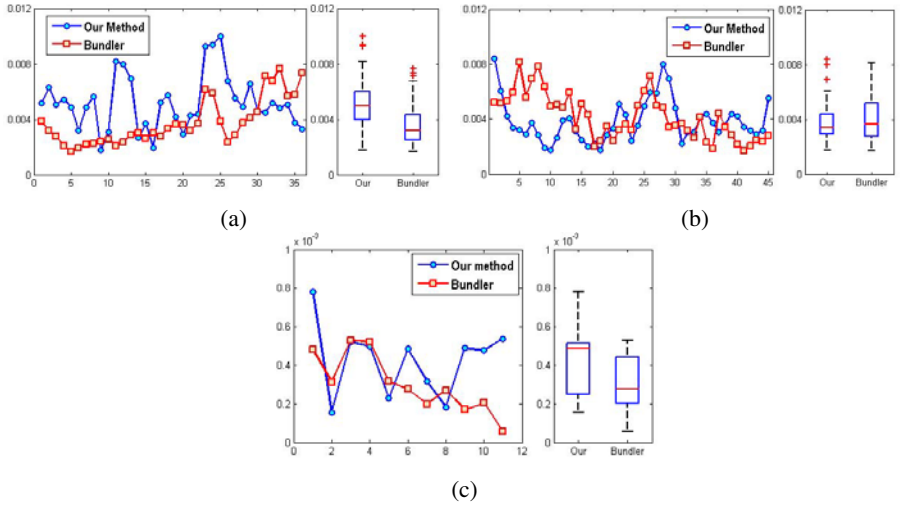

Fig. 4. This figure shows the errors in estimated camera matrices for our method and for bundler. The per-camera errors and their boxplots for the dinosaur sequence (a), the temple sequence (b) and for the fountain P11 sequence (c). One can remark that our method achieves the same level of accuracy as that of bundler, despite the fact that we do not use any information on the internal parameters, while bundler assumes that the skew is zero and the principal point is the center.

Results. For the dinosaur, temple and fountain P11 sequences, since ground truth exists, we compared our results with those of bundler [5], which is a state-of-the-art calibration software. The ground truth was normalized so that the Frobenius norm of all the cameras is one. For both reconstructions (ours and bundler), we computed numbers α_j and a homography $\mathtt{H}$ by minimizing (14). This allows us to define the per-camera error as $\|\alpha_j \mathtt{P}^j \mathtt{H} - \mathtt{P}^{j,*}\|_2^2$ for the jth camera. As shown in Fig. 4, not only these errors are small, but also our results are quite comparable to those of bundler despite the fact that our method does not perform intermediate BAs and does not assume that the principal point is in the center and the skew is zero. One can also note that the error is well distributed over the whole sequence of cameras due to the fact that both methods operate on the closed sequence. Furthermore, the results reported for fountain P11 are achieved without final BA, proving that the method we proposed furnishes a good starting point for the non-linear optimization.

As for the datasets where no ground truth is known, we have chosen to use as measure of evaluation the multiview stereo reconstruction of the scene based on the method of [25][2]. The results are shown in Fig. 1 (right) for the calvary sequence and in Fig. 5 for the Herz-Jesu R23 and the Detenice fountain sequences. In the aim of comparing our results with other approaches, let us recall that (as reported in [34]) on the Herz-Jesu R23 data the ARC3D software succeeded to calibrate four of the 23 cameras, while the method proposed in [4] calibrated all the cameras with a relatively large error for

[2] Since multiview stereo reconstruction is not the purpose of the paper and is only used for illustration, the results shown in Fig. 1 and 5 are obtained without the final mesh refinement.

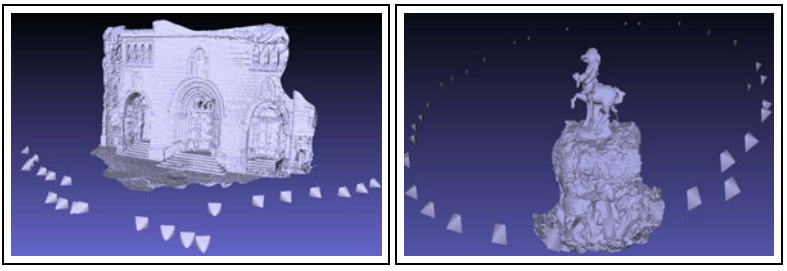

Fig. 5. Multi-view stereo reconstruction using the camera matrices estimated by our method for the Herz-Jesu R23 and Detenice fountain datasets. For these data, the ground truth is unavailable but the quality of the scene reconstruction demonstrates the accuracy of estimated cameras.

cameras 6-11. Although we are unable to quantitatively compare our reconstruction to that of [4], the accuracy of the 3D scene reconstruction makes us believe that the estimated cameras are very close to the true ones.

6 Conclusion

In this paper, we have proposed a new approach to the problem of autocalibration of a network of cameras. Our approach is based on a representation of the network of cameras by a graph of trifocal tensors and on a natural parameterization of camera matrices and relating homographies. We have proposed to estimate the unknown parameters by a constrained optimization that can be recast in a linear program. Thanks to the sparsity of the matrices involved in this linear program, the running times of the proposed algorithm are very attractive even for large scale datasets. The experiments reported in this paper show that our approach leads to state-of-the-art results without assuming any kind of information on the internal parameters.

Acknowledgments. We thank Vu Hiep for the results of multi-view stereo experiments, Daniel Martinec for the fountain dataset, Imagine for the calvary dataset and Christoph Strecha for fountain P11 and Herz-Jesu R23 datasets. This work was partially supported by ANR under grant Callisto.

References

1. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision, 2nd edn. Cambridge University Press, Cambridge (2003)
2. Faugeras, O., Luong, Q.T., Papadopoulou, T.: The Geometry of Multiple Images: The Laws That Govern The Formation of Images of A Scene and Some of Their Applications. MIT Press, Cambridge (2001)
3. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3D. ACM Press, New York (2006)
4. Martinec, D., Pajdla, T.: Robust rotation and translation estimation in multiview reconstruction. In: CVPR (2007)

5. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from Internet photo collections. Int. J. Comput. Vision 80, 189–210 (2008)
6. Furukawa, Y., Ponce, J.: Accurate camera calibration from multi-view stereo and bundle adjustment. International Journal of Computer Vision 84, 257–268 (2009)
7. Bujnak, M., Kukelova, Z., Pajdla, T.: 3D reconstruction from image collections with a single known focal length. In: ICCV, pp. 351–358 (2009)
8. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: ICCV (2009)
9. Havlena, M., Torii, A., Knopp, J., Pajdla, T.: Randomized structure from motion based on atomic 3d models from camera triplets. In: CVPR (2009)
10. Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P.: Generic and real-time sfm using local bundle adjustment. Image Vision Comput 27, 1178–1193 (2009)
11. Fitzgibbon, A.W., Zisserman, A.: Automatic camera recovery for closed or open image sequences. In: Burkhardt, H.-J., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1406, pp. 311–326. Springer, Heidelberg (1998)
12. Avidan, S., Shashua, A.: Threading fundamental matrices. IEEE Trans. Pattern Anal. Mach. Intell. 23, 73–77 (2001)
13. Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R.: Visual modeling with a hand-held camera. Int. J. Comput. Vision 59, 207–232 (2004)
14. Sinha, S.N., Pollefeys, M., McMillan, L.: Camera network calibration from dynamic silhouettes. In: CVPR (2004)
15. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. Int. J. Comput. Vision 9, 137–154 (1992)
16. Sturm, P., Triggs, B.: A factorization based algorithm for multi-image projective structure and motion. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1065, pp. 709–720. Springer, Heidelberg (1996)
17. Jacobs, D.: Linear fitting with missing data: Applications to structure-from-motion and to characterizing intensity images. In: CVPR, p. 206 (1997)
18. Martinec, D., Pajdla, T.: 3D reconstruction by fitting low-rank matrices with missing data. In: CVPR, pp. I: 198–205 (2005)
19. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment - a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) ICCV-WS 1999. LNCS, vol. 1883, pp. 298–372. Springer, Heidelberg (2000)
20. Klopschitz, M., Zach, C., Irschara, A., Schmalstieg, D.: Generalized detection and merging of loop closures for video sequences. In: 3DPVT (2008)
21. Scaramuzza, D., Fraundorfer, F., Pollefeys, M.: Closing the loop in appearance-guided omnidirectional visual odometry by using vocabulary trees. Robot. Auton. Syst. (to appear, 2010)
22. Cornelis, N., Cornelis, K., Van Gool, L.: Fast compact city modeling for navigation previsualization. In: CVPR (2006)
23. Tardif, J., Pavlidis, Y., Daniilidis, K.: Monocular visual odometry in urban environments using an omnidirectional camera. In: IROS, pp. 2531–2538 (2008)
24. Torii, A., Havlena, M., Pajdla, T.: From google street view to 3d city models. In: OMNIVIS (2009)
25. Vu, H., Keriven, R., Labatut, P., Pons, J.P.: Towards high-resolution large-scale multi-view stereo. In: CVPR (2009)
26. Chum, O., Matas, J.: Matching with PROSAC: Progressive sample consensus. In: CVPR, vol. I, pp. 220–226 (2005)
27. Ponce, J., McHenry, K., Papadopoulo, T., Teillaud, M., Triggs, B.: On the absolute quadratic complex and its application to autocalibration. In: CVPR, pp. 780–787 (2005)
28. Quan, L.: Invariants of six points and projective reconstruction from three uncalibrated images. IEEE Trans. Pattern Anal. Mach. Intell. 17, 34–46 (1995)

29. Schaffalitzky, F., Zisserman, A., Hartley, R.I., Torr, P.H.S.: A six point solution for structure and motion. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 632–648. Springer, Heidelberg (2000)
30. Golub, G.H., Van Loan, C.F.: Matrix computations, 3rd edn. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore (1996)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110 (2004)
32. Chum, O., Werner, T., Matas, J.: Two-view geometry estimation unaffected by a dominant plane. In: CVPR, vol. I, pp. 772–779 (2005)
33. Sattler, T., Leibe, B., Kobbelt, L.: Scramsac: Improving ransac‘s efficiency with a spatial consistency filter. In: ICCV (2009)
34. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In: CVPR (2008)