Automatic camera calibration using multiple sets of pairwise correspondences

Francisco Vasconcelos, João P. Barreto, and Edmond Boyer

Abstract—We propose a new method to add an uncalibrated node into a network of calibrated cameras using only pairwise point correspondences. While previous methods perform this task using triple correspondences, these are often difficult to establish when there is limited overlap between different views. In such challenging cases we must rely on pairwise correspondences and our solution becomes more advantageous. Our method includes an 11-point minimal solution for the intrinsic and extrinsic calibration of a camera from pairwise correspondences with other two calibrated cameras, and a new inlier selection framework that extends traditional RANSAC to sampling across multiple datasets. Our method is validated on different application scenarios where a lack of triple correspondences might occur: addition of a new node to a camera network; calibration and motion estimation of a moving camera inside a camera network; and addition of views with limited overlap to a Structure-from-Motion reconstruction.

Index Terms—Camera Calibration, Camera Networks, Minimal Algorithms, RANSAC

1 Introduction

A camera network, in the context of this article, is a set of cameras with synchronous image acquisition and partial overlap in the field-of-views (FOVs). These camera networks are popular in application domains that are concerned with the capture, the record, and the analysis of dynamic scenes, such as surveillance, gait analysis, human-motion capture, or 3D modelling for the movie industry [1]. Such applications invariably require the camera network to be calibrated, meaning that both intrinsic and extrinsic parameters must be known for all camera nodes in order to fuse the multiple-view information.

The problem of camera network calibration has been broadly addressed in the past. One possibility is to use a known calibration object, such as a checkerboard pattern, that is simultaneously observed by all nodes [2], [3], [4]. Some authors have recently proposed to observe the object through planar mirror reflections in order to handle situations of little or no overlap in the FOVs [5], [6]. Another option is to freely move a Light-Emitting Diodes (LED) in a dark room for obtaining accurate image correspondences that are used as inputs into factorization step [7], [8], [9]. All these calibration procedures are explicit, in the sense that they require substantial human intervention, and are meant to be carried as an initial off-line step before starting operating the network.

In spite of the many explicit methods for accomplishing camera network calibration, there are situations for which an automatic, unsupervised scheme is highly advantageous. Let us imagine that, while in operation, it is necessary to add or adjust the position of a camera, or that a node is inadvertently touched such that its 3D pose changes. Repeating the initial off-line procedure, not only is tiresome and requires interrupting operation, but it is also an overkill because all remaining nodes are correctly calibrated. Some efforts have been made to accomplish this task in real time in the context of sports broadcasting [10], [11]. These approaches take advantage from the fact that a sports field provides easily detectable features from a planar region. When no assumptions are made about the viewed scene a more suitable alternative is to estimate the camera parameters from natural image point correspondences with neighbouring calibrated views. It is well known that the camera projection matrix can be estimated in a DLT like manner from 6 or more triple correspondences [12], [13], with triple correspondence standing for an image point that is viewed in two other calibrated nodes such that it is possible to find its 3D coordinates. Unfortunately, and as shown in Fig. 1, triple correspondences are often difficult to establish in practice, either because cameras are separated by a wide-baseline and present very different perspectives, or because the dynamic scene creates relative occlusions that preclude matching. Thus, we propose to relax the requirements in the input data and carry the calibration from independent pairwise correspondences. The problem has 11 unknown parameters (5 intrinsics and 6 extrinsics), which means that in theory the solution can be fully constrained from a total of 11 pairwise correspondences between the uncalibrated camera and two distinct calibrated views. To the best of our knowledge the calibration of a camera from independent pairwise

- *F. Vasconcelos and J. P. Barreto are with the Institute for Systems and Robotics, University of Coimbra, Portugal*
- *E. Boyer is with INRIA Grenoble Rhône-Alpes, France*

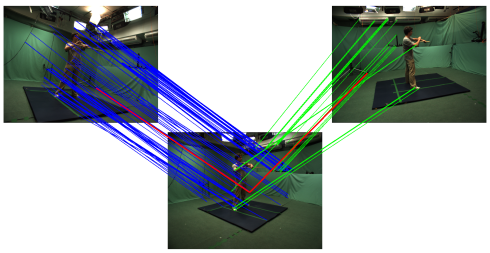


Fig. 1. Pairwise and triple correspondences extracted from SIFT features in a camera network. Given the wide baseline between the different views there is a single reliable triple correspondence (red) while there are many reliable pairwise correspondences (blue and green).

correspondence has never been solved. We propose the first minimal solution for the problem (the 11-pt algorithm) that requires 7 matches with the first view and 4 matches with the second view.

Such minimal algorithm can be used in a Random Sample Consensus (RANSAC) step for robust, accurate calibration of a camera in a network [14]. The standard RANSAC formulation assumes a single set of correspondences that is iteratively sampled to compute candidate models that are used to split data into inliers and outliers. In our case the search requires sampling not one but two datasets that might present different inlier-outlier statistics. It is shown that overlooking this fact and applying the standard RANSAC formulation leads to poor results. Thus, we propose a modified RANSAC version specifically designed to simultaneously sample multiple datasets. The usefulness of such RANSAC extension goes beyond the calibration problem at hands and can benefit other algorithms such as the Structure-from-Motion approaches from Clipp et al. [15], that use point correspondences across two and four cameras, and from Raposo et al. [16], [17], that mixture point and plane correspondences.

It is important to refer that the present article builds on our previous conference publication [18] that discloses the 11-pt algorithm. We extend this prior work by showing how to use the minimal solution in practice for accomplishing robust, accurate camera node calibration in a fully automatic manner. Thus, the contributions can be summarised as follows:

- A minimal algorithm for estimating the intrinsic and extrinsic parameters of a camera from 11 independent pairwise correspondences with two other calibrated cameras.
- Extensions of the well known RANSAC [14], MLESAC [19], and MAPSAC [20] formulations for sampling not one but multiple different datasets in simultaneous.
- A simple and efficient implementation of the complete solution that is tested in calibrating stationary camera nodes in a network, as well as in finding the parameters of a hand-held camera that freely moves in the network space to acquire close-ups of foreground dynamic scenes. The experiments confirm the superiority of the described algorithms with respect to the state-of-the-art.

2 Related Work

Since a camera array or network can be understood as a generalised camera [23], the extrinsic calibration of a camera from independent pairwise correspondences with multiple views relates with the problem of relative pose estimation between non-central cameras. It is well known that the rotation and translation between two generalised cameras can be solved linearly from 17 correspondences [21] and solved minimally from 6 pairwise correspondences [22]. However, these methods degenerate in many particular configurations, namely when one of the generalised views is a pin-hole as it happens in our case. In the case of the camera network having just two nodes the extrinsic calibration problem from pairwise correspondences can be potentially solved using methods developed for visual odometry using stereo cameras. There is a minimal solution for the relative pose between stereo pairs using 6 pairwise correspondences [15] that estimates an up-to scale relative pose solution using 5 correspondences with one camera [24] and solves the scale factor with an additional correspondence from another camera. A non-minimal solution using 10 correspondences was also proposed for the case of any arbitrary combination of correspondences between the 4 views of two stereo rigs [25]. Since in this paper we focus on both intrinsic and extrinsic calibration from pairwise correspondences, the above mentioned works relate but do not directly apply.

Whenever triple correspondences are available the calibration objective can be accomplished using standard techniques described in text books [12], [13]. These approaches typically rely on reconstructing 3D points from the the calibrated stereo views via triangulation [26], and using these points as known reference to calibrate the third view [8], [13]. Unfortunately, and as discussed in the introduction, triple correspondences are not always available. A possible alternative is to build a measurement matrix with the image correspondences, and perform projective factorization using the Sturm-Triggs algorithm [27] with a suitable extension for handling missing data [7]. However, this class of methods is meant for problems with multiple cameras and large number of correspondences, and it is unlikely that the approach will converge to a solution using only pairwise correspondences. Levi and Werman propose to build a viewing graph where pairs of camera nodes are linked by their fundamental matrices [28]. They show that, given a subset of

known fundamental matrices, it is possible to determine the remaining edges in the graph as far as each camera node is connected with at least two other camera nodes. This condition is not verified whenever we aim to calibrate a camera that has been just added to an existing camera network. In [29] Josephson et al. investigate the problem of calibrating a camera node from mixtures of triple and pairwise correspondences. However, and to the best of our knowledge, the calibration of a camera using exclusively independent pairwise correspondences with two other views has never been addressed in the literature before.

Additionally we are interested in robust estimation with RANSAC when candidate solutions are generated by sampling multiple datasets. This problem is only briefly addressed in [15] when estimating the relative pose between stereo rigs. In that problem RANSAC must independently select one sample from 3 datasets containing 2-view, 3-view, and 4-view correspondences. It is shown that the number of RANSAC iterations must be computed in a different way when the different datasets have different inlier ratios, and also that different types of correspondences should be weighted differently on the cost function. However, the observations made in [15] can only be directly extended to problems where we know exactly how many samples are selected in each dataset. In our calibration problem this might not be the case, since there are different combinations of pairwise correspondences among different cameras that can generate a solution. Adapting other variants from the RANSAC-family (e. g. MLESAC [19]) to a multiple dataset framework have also never been addressed before.

2.1 Notation

Scalars are represented by plain letters, e.g. λ, vectors are indicated by bold symbols, e.g. $\mathbf{t}$, and matrices are denoted by letters in sans serif font, e.g. T. 3D lines are expressed in homogeneous Plucker coordinates, e.g. the 6×1 vector $\mathbf{L}$. The equality up to scale is denoted by $\sim$ in order to be distinguished from the strict equality $=$, and the operator $[\mathbf{v}]_\times$ designates the 3×3 skew symmetric matrix of a 3×1 vector $\mathbf{v}$. We also use matrix superscripts, e. g. $\mathsf{T}^{\{n\}}$, to denote its nth column.

3 Problem Statement

Let us consider two calibrated cameras $\mathbf{C}_A$ and $\mathbf{C}_B$, such that the matrices of intrinsic parameters are K_A and K_B, and the absolute poses are expressed in a world coordinate system $\mathbf{O}_w$ by the rotation matrices R_A and R_B, and the translation vectors $\mathbf{t}_A$ and $\mathbf{t}_B$. Consider an additional camera $\mathbf{C}$ for which both the intrinsic calibration K, and the extrinsic calibration R, $\mathbf{t}$ are unknown. Our article addresses the problem of calibrating this third camera using as input data a set

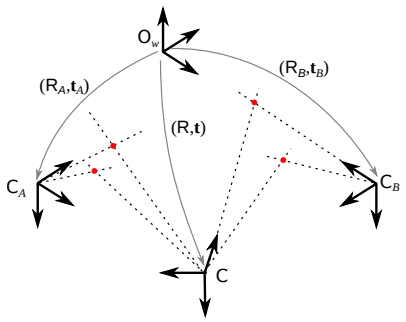


Fig. 2. We consider the problem of fully calibrating the camera $\mathbf{C}$, given pairwise correspondences with two calibrated cameras $\mathbf{C}_A$ and $\mathbf{C}_B$.

of a image correspondences $(\mathbf{x}^{(i)}, \mathbf{x}_A^{(i)})$ between $\mathbf{C}$ and $\mathbf{C}_A$, and set of b image correspondences $(\mathbf{x}^{(a+j)}, \mathbf{x}_B^{(j)})$ between $\mathbf{C}$ and $\mathbf{C}_B$ (Fig. 2). We assume that the two sets of pairwise matches are independent, meaning that

$$\mathbf{x}^i \neq \mathbf{x}^{a+j}, \forall_{i=1\ldots a,\, j=1\ldots b}\,. \tag{1}$$

4 Linear Constraints

In this section we derive a system of linear equations that has a minimum number of unknowns and fully constrains the camera calibration. The problem is formulated in the context of epipolar geometry between general camera models [23], with one side being the uncalibrated pin-hole camera $\mathbf{C}$, and the other side being the pair of calibrated cameras $\mathbf{C}_A$ and $\mathbf{C}_B$ that can be understood as a particular instance of a non-central imaging device denoted by $\mathbf{C}_A \cup \mathbf{C}_B$. It is shown below that under such configuration the corresponding back-projection lines must satisfy a bilinear relation expressed by a 3×5 matrix, and that the estimation of the epipolar geometry using a DLT-like approach cannot be achieved with less than 14 pairwise matches.

Note that when the intrinsics are known, this problem is a particular case of the pose estimation between calibrated general camera models [23] that has already been solved both linearly [21] and using the minimal number of 6 pairwise correspondences [22].

4.1 Line Incidence Relations

Let $\mathbf{x}_A$ and $\mathbf{x}_B$ be image points in $\mathbf{C}_A$ and $\mathbf{C}_B$. Since the cameras are fully calibrated, the corresponding back-projection lines $\mathbf{L}_A$ and $\mathbf{L}_B$ can be expressed in the common world reference frame $\mathbf{O}_w$ by a homogeneous Plücker vector [30]

$$\mathbf{L}_{A/B} \sim \begin{pmatrix} \mathbf{d}_{A/B} \\ \mathbf{m}_{A/B} \end{pmatrix}, \tag{2}$$

with the 3-vectors $\mathbf{d}_{A/B}$ and $\mathbf{m}_{A/B}$ being respectively the direction and the momentum of the line. In a similar manner, an image point $\mathbf{x}$ in $\mathbf{C}$ gives rise to a

back-projection line $\mathbf{L}$ that is represented in the local camera reference frame by

$$\mathbf{L} \sim \begin{pmatrix} \mathbf{d} \\ \mathbf{0} \end{pmatrix}, \tag{3}$$

with the direction depending on the matrix of intrinsic parameters K

$$\mathbf{d} \sim \mathsf{K}^{-1}\mathbf{x}. \tag{4}$$

If $\mathbf{x}$ and $\mathbf{x}_{A/B}$ are image correspondences, then the back-projection lines $\mathbf{L}$ and $\mathbf{L}_{A/B}$ must be incident. Given the rigid displacement between the reference frames $\mathbf{O}_w$ and $\mathbf{C}$, and the condition for two lines in Plücker coordinates to intersect, it comes that the following condition must hold [30]

$$\mathbf{L}^{\mathsf{T}} \begin{pmatrix} \mathsf{0} & \mathsf{I} \\ \mathsf{I} & \mathsf{0} \end{pmatrix} \begin{pmatrix} \mathsf{R} & \mathsf{0} \\ [\mathbf{t}]_{\times}\mathsf{R} & \mathsf{R} \end{pmatrix} \mathbf{L}_{A/B} = 0. \tag{5}$$

Since the momentum of $\mathbf{L}$ is always zero, then the above equation can be re-written as

$$\mathbf{d}^{\mathsf{T}} \begin{pmatrix} [\mathbf{t}]_{\times}\mathsf{R} & \mathsf{R} \end{pmatrix} \mathbf{L}_{A/B} = 0. \tag{6}$$

Equation 6 is the particular case of the generalized epipolar constraint proposed in [23] when one of the cameras is a conventional pin-hole. However, and similarly to the general case, the bilinear relation between back-projection lines is expressed by a 3×6 matrix that encodes the calibration parameters. In a first glance it might seem that linearly estimating the 18 entries of this matrix up to a global scale factor can be carried with 17 or more image correspondences between $\mathbf{C}$ and the camera pair $\mathbf{C}_A \cup \mathbf{C}_B$. However, and as discussed below, these correspondences only provide 15 independent linear constraints in the matrix parameters.

In our case the parametrization of equation 6 leads to a linear estimation problem that is sub-determined. This is a situation similar to the degenerate configurations recently reported in [21] in the context of motion estimation using a calibrated multi-camera rig.

4.2 Compact linear formulation

The image rays belonging to two pinhole cameras $\mathbf{C}_A$ and $\mathbf{C}_B$ define a subset of lines that intersect a common axis $\mathbf{h}$ (Fig 4). This subset is called a *linear line congruent* [31], and all its elements can be defined as a linear combination of five lines $\mathbf{G}_1$, $\mathbf{G}_2$, $\mathbf{G}_3$, $\mathbf{G}_4$, $\mathbf{G}_5$ that intersect $\mathbf{h}$. In our calibration problem every possible back-projection line $\mathbf{L}_{A/B}$ must intersect the line going through $\mathbf{C}_A$ and $\mathbf{C}_B$ (the baseline). Thus, the lines $\mathbf{L}_{A/B}$ can be represented in a unique manner as the linear combination of any 5 lines $\mathbf{G}_i$ that intersect the baseline

$$\mathbf{L}_{A/B} \sim \underbrace{\begin{pmatrix} \mathbf{G}_1 & \mathbf{G}_2 & \mathbf{G}_3 & \mathbf{G}_4 & \mathbf{G}_5 \end{pmatrix}}_{\mathsf{G}} \boldsymbol{\lambda}_{A/B}, \tag{7}$$

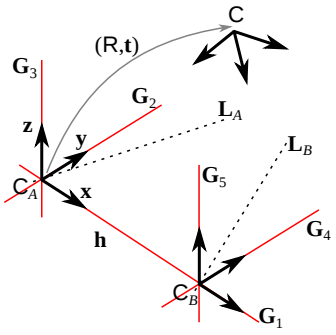


Fig. 3. The space generated by two bundles of lines (the rays of 2 pinhole cameras) can be fully represented as the linear span of $\{\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3, \mathbf{G}_4, \mathbf{G}_5\}$.

where G is a 6×5 matrix with full rank, and $\boldsymbol{\lambda}_{A/B}$ is a 5-vector defined up to scale. Replacing in equation 6 yields

$$\mathbf{d}^{\mathsf{T}} \begin{pmatrix} [\mathbf{t}]_{\times}\mathsf{R} & \mathsf{R} \end{pmatrix} \mathsf{G}\boldsymbol{\lambda}_{A/B} = 0. \tag{8}$$

We have just re-written the epipolar constraint of equation 6 as a bilinear relation between the direction $\mathbf{d}$ of the line $\mathbf{L}$ in camera $\mathbf{C}$, and the representation $\boldsymbol{\lambda}_{A/B}$ of the back-projection line $\mathbf{L}_{A/B}$ in the generalized camera $\mathbf{C}_A \cup \mathbf{C}_B$. Since the bilinear relation is now encoded by a 3×5 matrix with 15 entries, then 14 image point correspondences are sufficient for estimating the epipolar geometry in a DLT-like manner.

Given the two arbitrary calibrated cameras, it is always possible to perform a change of reference frames for achieving the configuration exhibited in Fig. 4. We consider, without any loss of generality, that the world reference frame is aligned with the coordinate system of camera $\mathbf{C}_A$, and that the X-axis is coincident with the baseline defined by the projection centers of the two pin-holes. The local reference frame of the second camera is assumed to have origin in $\mathbf{C}_B$ and to be parallel to the coordinate system of $\mathbf{C}_A$. Under such circumstances the rigid transformation that maps point coordinates from $\mathbf{C}_B$ to $\mathbf{C}_A$ is given by

$$\mathsf{T}_{B\to A} = \begin{pmatrix} \mathsf{I} & \mathbf{h} \\ \mathbf{0} & 1 \end{pmatrix} \tag{9}$$

with I being the 3×3 identity matrix and $\mathbf{h} = \begin{pmatrix} h & 0 & 0 \end{pmatrix}^{\mathsf{T}}$. Since the axes X, Y, Z of the system of coordinates of $\mathbf{C}_A$, and the axes Y, Z of the reference frame of $\mathbf{C}_B$ are linearly independent lines, then they can be used to establish a basis G for the LLC defined by the baseline. It comes that

$$\mathsf{G} \sim \begin{pmatrix} \mathsf{I} & \mathsf{I}^{\{2,3\}} \\ 0 & [\mathbf{h}]_{\times}^{\{2,3\}} \end{pmatrix} \tag{10}$$

with the upper script $\{2,3\}$ denoting the second and third columns of the matrix.

Let us now consider an image correspondence $(\mathbf{x}, \mathbf{x}_A)$ between $\mathbf{C}$ and $\mathbf{C}_A$. The back-projection of

$\mathbf{x}_A$ is a line $\mathbf{L}_A$ with direction $\mathbf{d}_A$ expressed in the reference frame of $\mathbf{C}_A$. Given the basis G above, it comes that $\mathbf{L}_A \sim \mathsf{G}\boldsymbol{\lambda}_A$ with

$$\boldsymbol{\lambda}_A \sim \begin{pmatrix}\mathbf{d}_A^\mathsf{T} & 0 & 0\end{pmatrix}^\mathsf{T}. \tag{11}$$

Replacing in equation 8, and making $\mathbf{d} \sim \mathsf{K}^{-1}\mathbf{x}$, yields

$$\mathbf{x}^\mathsf{T}\,\mathsf{F}_A\,\mathbf{d}_A = 0 \tag{12}$$

with F_A being the standard fundamental matrix between the uncalibrated camera $\mathbf{C}$ and the calibrated view $\mathbf{C}_A$

$$\mathsf{F}_A = \mathsf{K}^{-\mathsf{T}}\,[\mathbf{t}]_\times\,\mathsf{R}\,. \tag{13}$$

Repeating the reasoning for the case of an image correspondence $(\mathbf{x}, \mathbf{x}_B)$ between $\mathbf{C}$ and $\mathbf{C}_B$, it comes that

$$\mathbf{x}^\mathsf{T}\,\mathsf{F}_B\,\mathbf{d}_B = 0 \tag{14}$$

with F_B being the fundamental matrix between $\mathbf{C}$ and $\mathbf{C}_B$ that can be written as

$$\mathsf{F}_B = \mathsf{F}_A + \mathsf{K}^{-1}\mathsf{R}[\mathbf{h}]_\times. \tag{15}$$

It follows from the equation above that the first columns of matrices F_A and F_B are always equal ($\mathsf{F}_A^{\{1\}} = \mathsf{F}_B^{\{1\}}$).

Given the image correspondences $(\mathbf{x}^{(i)}, \mathbf{x}_A^{(i)})$, with $i = 1, \dots a$, and $(\mathbf{x}^{(a+j)}, \mathbf{x}_B^{(j)})$ with $j = 1, \dots b$, we can determine the line directions $\mathbf{d}_A^{(i)} \sim \mathsf{K}_A^{-1}\mathbf{x}_A^{(i)}$ and $\mathbf{d}_B^{(j)} \sim \mathsf{K}_B^{-1}\mathbf{x}_B^{(j)}$, and establish a system of linear equations (equation 16) based on the bilinear constraints of equations 12 and 14.

If $a + b \geq 14$ then the fundamental matrices F_A and F_B can be determined up to a common scale factor using a standard DLT approach.

5 Minimal Solution

We have shown that the two fundamental matrices, F_A and F_B, that encode the calibration information K, R, and $\mathbf{t}$, can be determined from a minimum of 14 independent image correspondences. However, the total number of independent unknowns is 11 (5 intrinsic parameters and 6 extrinsic parameters) meaning that the estimation problem can be further constrained. Two of these constrains are rather obvious:

$$\det(\mathsf{F}_A) = 0\,, \tag{17}$$

$$\det(\mathsf{F}_B) = 0\,. \tag{18}$$

For the third constraint it must be observed that the sum of F_A and F_B is still a fundamental matrix. From equations 13 and 15 it comes after algebraic manipulation that

$$\mathsf{F}_A + \mathsf{F}_B = \mathsf{K}^{-1}[2\mathbf{t} + \mathsf{R}\mathbf{h}]_\times\mathsf{R}\,, \tag{19}$$

which means that the following condition must hold

$$\det(\mathsf{F}_A + \mathsf{F}_B) = 0\,. \tag{20}$$

The equation above basically enforces the condition that F_A and F_B must be two fundamental matrices encoding the same rotation R.

5.1 Outline of the estimation algorithm

F_A and F_B can be estimated from a minimum number of $a + b = 11$ pairwise correspondences. Note, however, that a single fundamental matrix can be estimated from 7 pairwise correspondences with a single camera, and therefore if $a > 7$ or $b > 7$ some equations are redundant. There are only two solvable minimal configurations in this problem: $(a = 7, b = 4)$ and $(a = 6, b = 5)$. We consider only the case $(a = 7, b = 4)$:

1) Build the linear system of equation 16 from the 11 pairwise correspondences.
2) Use the top 7 equations of this system determine a 2-dimensional solution space for the 9 parameters of F_A using SVD. This enables to write $\mathsf{F}_A(\alpha) = \mathsf{A}' + \alpha\mathsf{A}$ with α being a free parameter.
3) Compute α by solving the cubic constraint of equation 17 and determine F_A.
4) Substitute the up-to scale solution of F_A in the linear system of equation 16. This system has now only 7 unknowns: the 6 parameters of F_B and the scale factor of F_A. The bottom 4 equations of this system can be used to determine a 3-dimensional solution space for F_B. This enables to write $\mathsf{F}_B(\beta_1, \beta_2) = \mathsf{B}'' + \beta_1\mathsf{B}' + \beta_2\mathsf{B}$.
5) Substitute F_A and $\mathsf{F}_B(\beta_1, \beta_2)$ in equations 18 and 20. This leads to a bivariate system of 2 quadratic equations. Compute β_1 and β_2 by solving the bivariate system [32], and determine the fundamental matrix F_B.

Since the cubic equation of step 3 gives up to 3 discrete solutions, and the bivariate system of quadric equations has at most 4 distinct solutions, then there is a maximum of 12 possible solutions for the pair of fundamental matrices $(\mathsf{F}_A, \mathsf{F}_B)$.

5.2 Degenerate Configurations

The 11-point solution degenerates in two cases. If the 7 pairwise correspondences that are established with the same camera belong to a single plane the linear system of equation 16 is rank deficient. This is a known degenerate configuration of the 7-point algorithm that applies to our case as well. The second degeneracy happens when there is no translation between the two calibrated cameras C_A and C_B. In this case all calibrated image raysbelong to the same bundle and the problem becomes equivalent to the estimation of a fundamental matrix between a calibrated and an uncalibrated pinhole views.

6 Factorization of F_A and F_B

In order to solve the calibration problem, F_A and F_B must be factorized into the intrinsic parameters K

$$\begin{pmatrix} x_1^{(1)}\mathbf{d}_A^{(1)\mathsf{T}} & x_2^{(1)}\mathbf{d}_A^{(1)\mathsf{T}} & x_3^{(1)}\mathbf{d}_A^{(1)\mathsf{T}} & 0^{\mathsf{T}} & 0^{\mathsf{T}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{(a)}\mathbf{d}_A^{(a)\mathsf{T}} & x_2^{(a)}\mathbf{d}_A^{(a)\mathsf{T}} & x_3^{(a)}\mathbf{d}_A^{(a)\mathsf{T}} & 0^{\mathsf{T}} & 0^{\mathsf{T}} \\ x_1^{(a+1)}\mathbf{d}_B^{(1)\mathsf{T}} & 0^{\mathsf{T}} & 0^{\mathsf{T}} & x_2^{(a+1)}\mathbf{d}_B^{(1)\mathsf{T}} & x_3^{(a+1)}\mathbf{d}_B^{(1)\mathsf{T}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{(a+b)}\mathbf{d}_B^{(b)\mathsf{T}} & 0^{\mathsf{T}} & 0^{\mathsf{T}} & x_2^{(a+b)}\mathbf{d}_B^{(b)\mathsf{T}} & x_3^{(a+b)}\mathbf{d}_B^{(b)\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathsf{F}_A^{\{1\}} \\ \mathsf{F}_A^{\{2\}} \\ \mathsf{F}_A^{\{3\}} \\ \mathsf{F}_B^{\{2\}} \\ \mathsf{F}_B^{\{3\}} \end{pmatrix} = 0 \tag{16}$$

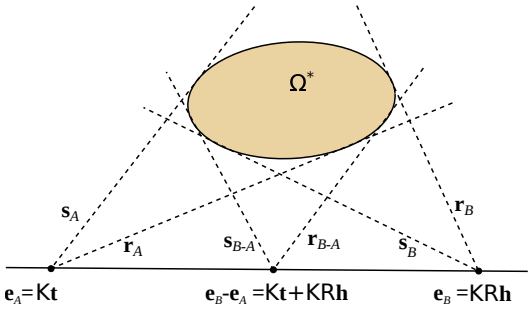


Fig. 4. Conic envelope Ω establishes linear relations $\mathbf{s}^{\mathsf{T}}\mathsf{K}\mathsf{K}^{\mathsf{T}}\mathbf{s} = 0$ and $\mathbf{r}^{\mathsf{T}}\mathsf{K}\mathsf{K}^{\mathsf{T}}\mathbf{r} = 0$.

and the relative pose R, $\mathbf{t}$. Let us first discuss the extraction of the matrix K. Consider the fundamental matrix F_A that is given in equation 13. After some algebraic manipulations we obtain that

$$\mathsf{F}_A\,\mathsf{F}_A^{\mathsf{T}} \sim [\mathbf{e}_A]_\times\,\mathsf{K}\mathsf{K}^{\mathsf{T}}\,[\mathbf{e}_A]_\times \tag{21}$$

with $\mathbf{e}_A = \mathsf{K}\mathbf{t}$ denoting the left side epipole of F_A (the image on C of the principal point of C_A). From the result above it follows that, if $\mathbf{y}$ is a point in the projective plane that satisfies

$$\mathbf{y}^{\mathsf{T}}\,\mathsf{F}_A\,\mathsf{F}_A^{\mathsf{T}}\,\mathbf{y} = 0\,, \tag{22}$$

then the line defined by $\mathbf{y}$ and $\mathbf{e}_A$ lies in the conic envelope $\Omega^* = \mathsf{K}\mathsf{K}^{\mathsf{T}}$ that is the dual of the image of the absolute conic (DIAC) [12], [13]. $\mathsf{F}_A\,\mathsf{F}_A^{\mathsf{T}}$ is a rank 2 symmetric matrix that can be understood as a degenerate conic locus comprising the points lying in two lines $\mathbf{s}_A$, $\mathbf{r}_A$ that intersect $\mathbf{e}_A$. From the two observations above it is easy to conclude that $\mathbf{s}_A$, $\mathbf{r}_A$ must belong to the DIAC, as shown in Fig. 5. The same reasoning can be applied to the fundamental matrix F_B of equation 15

$$\mathsf{F}_B\,\mathsf{F}_B{}^{\mathsf{T}} \sim [\mathbf{e}_B]_\times \mathsf{K}\mathsf{K}^{\mathsf{T}}[\mathbf{e}_B]_\times\,, \tag{23}$$

and to the matrix $\mathsf{F}_B - \mathsf{F}_A$ that is still rank deficient because the first columns of the two fundamental matrices are equal

$$(\mathsf{F}_B - \mathsf{F}_A)(\mathsf{F}_B - \mathsf{F}_A)^{\mathsf{T}} \sim [\mathbf{e}_B - \mathbf{e}_A]_\times \mathsf{K}\mathsf{K}^{\mathsf{T}}[\mathbf{e}_B - \mathbf{e}_A]_\times \tag{24}$$

Summarizing, and as shown in Fig. 5, the DIAC is fully constrained by the line pairs arising from the rank 2 degenerate conics $\mathsf{F}_A\mathsf{F}_A^{\mathsf{T}}$, $\mathsf{F}_B\mathsf{F}_B^{\mathsf{T}}$, and $(\mathsf{F}_B - \mathsf{F}_A)(\mathsf{F}_B - \mathsf{F}_A)^{\mathsf{T}}$. It is important to refer that, although

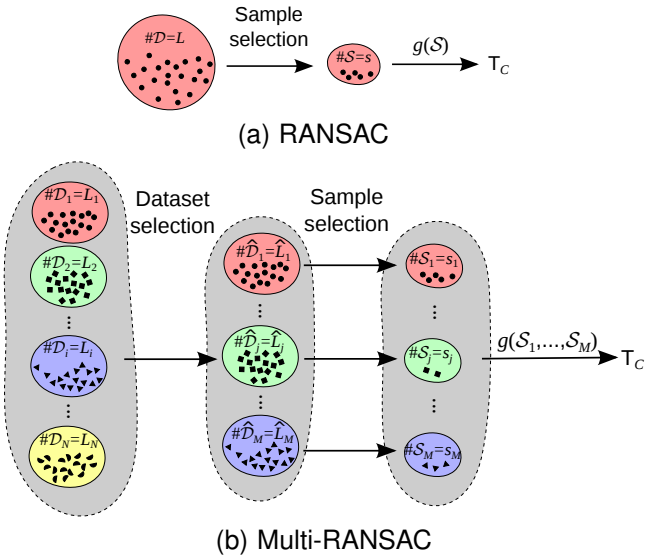


Fig. 5. Generation of candidate models T_C. (a) In each RANSAC iteration a subset $\mathcal{S}$ with s samples is selected from dataset $\mathcal{D}$. (b) In in each multi-RANSAC iteration there are two sampling steps: the first step randomly selects M of the N datasets $\mathcal{D}_1$, ..., $\mathcal{D}_N$; in the second step s_j samples are selected from each of the M selected datasets $\hat{\mathcal{D}}_j$.

we have six lines, they only give rise to five independent constraints on the parameters of the DIAC. This is explained by the fact that their pairwise intersections are collinear.

After knowing K, we can compute the essential matrix E_A and apply standard techniques for determining the rotation R and the translation $\mathbf{t}$ up to scale factor [12], [13]. The scale factor can be easily found using the known baseline between $\mathbf{C}_A$ and $\mathbf{C}_B$.

7 RANSAC WITH MULTIPLE DATASETS

RANSAC [14] is the most widely used method to eliminate outlier correspondences when a minimal solution is available. This method attempts to fit a model T to a single dataset $\mathcal{D}$ with L correspondences which are either inliers or outliers. RANSAC iteratively generates candidate models $\mathsf{T}_C = g(\mathcal{S})$ by randomly selecting a subset $\mathcal{S}$ with s random samples from $\mathcal{D}$ (Fig. 6(a)). In each iteration a candidate model T_C is evaluated using some cost metric and whenever a model with a lower cost is found it is

stored as the current best candidate. After a certain number of iterations n RANSAC stops and outputs the best candidate. Different versions of RANSAC have different cost metrics: original RANSAC [14] minimizes the number of outliers for a pre-defined threshold t; MLESAC [19] chooses the model with maximum likelihood, assuming that the residue of inliers follows a gaussian distribution, while the residue of outliers follow a uniform distribution; MAPSAC [20] maximizes the posterior probability of a model and its latent parameters. Despite these differences, all versions of RANSAC work under the assumption that samples are selected from a single dataset $\mathcal{D}$ with a certain inlier ratio γ whose value is updated according to the current best candidate. An accurate estimation of the inlier ratio γ is important to know the required number of RANSAC iterations and also to compute the cost metrics of MLESAC and MAPSAC.

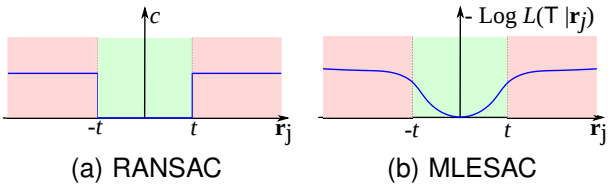
Our problem, however, does not fit into the standard assumptions of RANSAC. A model generator for our problem involves selecting 7 correspondences from one dataset and 4 from another. These two datasets might have different inlier ratios and thus all RANSAC assumptions that depend on a single value γ must be revised. Additionally we might think of a scenario where there are correspondences with $N > 2$ cameras and thus to use RANSAC with our algorithm we must first select 2 cameras and only then sample 7 and 4 correspondences from them. With these issues in mind we propose a new framework, called multi-RANSAC, that takes into account the sampling of different datasets.

For the sake of generalization we assume an arbitrary problem with N datasets $\mathcal{D}_1, ..., \mathcal{D}_N$ and a model generator $\mathsf{T}_C = g(\mathcal{S}_1, ..., \mathcal{S}_M)$ that requires M subsets $\mathcal{S}_j$, each of them containing s_j samples from one of the datasets $\mathcal{D}_1, ..., \mathcal{D}_N$. As displayed in Fig. 6(b), the sampling process is done in two steps: it first selects M datasets $\hat{\mathcal{D}}_1, ..., \hat{\mathcal{D}}_M$ from the N datasets $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N$, with $M \leq N$; then it selects a subset $\mathcal{S}_j$ with s_j samples from each selected dataset $\hat{\mathcal{D}}_j$.

In this section we discuss the necessary adaptations to RANSAC, MLESAC, and MAPSAC when dealing with multiple datasets, which we designate by multi-RANSAC, multi-MLESAC, and multi-MAPSAC respectively.

7.1 Multi-RANSAC

In standard RANSAC it is assumed that inlier samples have a uniform error distribution over some bounded interval. All samples with an error greater than a threshold t are considered outliers. In this case the evaluation cost of each candidate model is simply the total number of outliers (Fig. 7(a)). For the multi-RANSAC approach we can use the same evaluation metric by summing up the outliers in all datasets. We might also consider tuning different thresholds



(a) RANSAC (b) MLESAC

Fig. 6. Cost evaluation metrics for a model T, given a sample $\mathbf{d}_j$ with residue r_j. The threshold t separates inliers (green) from outliers (red). In RANSAC the cost function c is zero if $\mathbf{d}_j$ is an inlier or a constant value if d_j is an outlier. In MLESAC the cost function approximates a squared residue if $\mathbf{d}_j$ is an inlier and approximates a constant cost if $\mathbf{d}_j$ is an outlier.

for each dataset, when it makes sense in a particular problem.

In standard RANSAC the number of iterations n is determined by guaranteeing that at least in one of the iterations a model is generated only from inlier samples with a probability p, set to a value close to 1. The sampling process is approximated by a succession of s Bernoulli trials, i. e., a succession of s independent sample selections with a constant probability γ of selecting an inlier in each of them. Therefore, the probability p_{ins} of selecting s consecutive inliers is

$$p_{ins} = \gamma^s \tag{25}$$

Note that the probability γ also represents the inlier ratio in dataset $\mathcal{D}$. We want to guarantee that the probability of never selecting s inliers after n iterations is lower than $1 - p$, .i e.

$$(1 - \gamma^s)^n < 1 - p\,. \tag{26}$$

Therefore, the number of RANSAC iterations is

$$n = \frac{\log(1-p)}{\log(1-\gamma^s)}\,. \tag{27}$$

Whenever a new best model is found, the values γ and consequently n are updated.

In the multi-RANSAC case n must be computed differently since the sample selection process has two steps and each dataset $\mathcal{D}_i$ might have a different inlier ratio γ_i. The probability of obtaining an inlier by first selecting a random dataset $\mathcal{D}_i$ and then selecting a random sample from it is

$$p_{in} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i\,. \tag{28}$$

In an analogous manner to the standard RANSAC formulation we assume that the probability of selecting an inlier from dataset $\mathcal{D}_i$ is a constant value γ_i for successive selections (i. e. the second selection step is a succession of Bernoulli trials). The probability of selecting s_j inliers by first selecting a random dataset

$\mathcal{D}_i$ and then selecting s_j random samples from it is

$$p_{ins} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i^{s_j}\,. \tag{29}$$

We now further approximate the first selection step (dataset selection) by a succession of Bernoulli trials. The complete multi-RANSAC sampling process is thus simplified to selecting a random dataset M times and for each of them successively selecting s_1, s_2, ..., s_M samples. The probability of selecting only inliers in this process is

$$p_{inall} = \prod_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\gamma_i^{s_j}\,. \tag{30}$$

Using the same reasoning behind equation 27, the number of multi-RANSAC iterations is now

$$n = \frac{\log(1-p)}{\log(1-p_{inall})}. \tag{31}$$

Note however that approximating the dataset selection step by Bernoulli trials is only valid when M is much smaller than N because the same dataset cannot be selected more than once. Specifically in the case that $M = N$ the above equation might grossly misestimate the number of required iterations. In this case the dataset selection step outputs a random permutation of all datasets, choosing for each dataset $\mathcal{D}_i$ which number of samples s_j is selected. There are $N!$ possible dataset selections, and to obtain p_{inall} we must weigh in the probability of selecting only inliers for all possible permutations.

To illustrate this case consider our calibration problem when there are only correspondences with two cameras. In this case we have two datasets with pairwise correspondences $\mathcal{D}_A$, $\mathcal{D}_B$ with inlier ratios γ_A, γ_B and we want to select 7 correspondences from one of them and 4 from the other. In this case $N = M = 2$ and equation 30 becomes

$$p_{inall} = \frac{1}{4}\left(\gamma_A^7\gamma_B^4 + \gamma_B^7\gamma_A^4 + \gamma_A^7\gamma_A^4 + \gamma_B^7\gamma_B^4\right) \tag{32}$$

However, in a more careful analysis, we can observe that the dataset selection step has only two possible outcomes: either 7 correspondences are selected from $\mathcal{D}_A$ and 4 from $\mathcal{D}_B$, or 4 correspondences are selected from $\mathcal{D}_A$ and 7 from $\mathcal{D}_B$. The probability of selecting only inliers in this process is

$$p_{inall} = \frac{1}{2}(\gamma_A^7\gamma_B^4 + \gamma_B^7\gamma_A^4)\,. \tag{33}$$

All the results derived for computing the number of multi-RANSAC iterations also extend to the multi-MLESAC and multi-MAPSAC formulations discussed next.

7.2 Multi-MLESAC

MLESAC [19] aims at finding the model T with minimum negative log-likelihood, given a set of measurements $\mathcal{D}$. Each sample $\mathbf{d}_k$ in $\mathcal{D}$ can be put into one of two subsets: the inliers $\mathcal{I}$ or the outliers $\mathcal{O}$.

The residue of samples in $\mathcal{I}$ is assumed to follow a gaussian distribution $N(0,\sigma)$. A model T, given an inlier sample $\mathbf{d}_k$ with residue $r_k^{\mathcal{I}}$, has a likelihood

$$L(\mathsf{T}|r_k^{\mathcal{I}}) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-|r_k|^2}{2\sigma^2}}\,. \tag{34}$$

The samples from $\mathcal{O}$ are observations independent from the model, and their residue is assumed to follow a uniform distribution over an interval $[-\frac{v}{2},\frac{v}{2}]$. A model T, given an outlier sample $\mathbf{d}_k$ with residue $r_k^{\mathcal{O}}$ has a constant likelihood

$$L(\mathsf{T}|r_k^{\mathcal{O}}) = \frac{1}{v}\,. \tag{35}$$

The samples from dataset $\mathcal{D}$ follow a mixed distribution of inliers and outliers (Fig. 7(b)) and therefore the likelihood $L(\mathsf{T}|r_k)$ of a model T, given a random sample $\mathbf{d}_k$ from $\mathcal{D}$ with residue r_k is

$$L(\mathsf{T}|r_k) = \left(\gamma\left(\frac{1}{\sqrt{2\pi}\sigma}\right)e^{\frac{-|r_k|^2}{2\sigma^2}} + \frac{1-\gamma}{v}\right), \tag{36}$$

where γ is the probability of d_k being an inlier.

The MLESAC problem can now be formulated by considering the negative log-likelihood of T given all L samples in $\mathcal{D}$

$$\min_{\mathsf{T}}\left(-\sum_{k=1}^{L}\log L(\mathsf{T}|r_k)\right). \tag{37}$$

Note that the inlier ratio γ is updated in each iteration using expectation maximization with the following constraint:

$$\gamma = \frac{1}{L}\sum_{k=1}^{L}Pr(r_k^{\mathcal{I}}|\gamma) \tag{38}$$

where γ is initialized to 0.5 on the left side of the equation and is iteratively updated until convergence.

We now consider the multi-MLESAC problem. When sampling from N different datasets we aim at maximizing the likelihood of model T given datasets $\mathcal{D}_1$, ..., $\mathcal{D}_N$, each of them with a number of samples L_i, an inlier standard deviation σ_i, an outlier range v_i, and an inlier ratio γ_i. In this case the likelihood of a model T, given a sample $\mathbf{d}_{i,k}$ from dataset $\mathcal{D}_i$ with a residue $r_{i,k}$ is

$$L(\mathsf{T}|r_{i,k}) = \left(\gamma_i\left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)e^{\frac{-|r_{i,k}|^2}{2\sigma_i^2}} + \frac{1-\gamma_i}{v_i}\right). \tag{39}$$

The multi-MLESAC problem for N datasets can now be formulated as

$$\min_{\mathsf{T}}\left(-\sum_{i=1}^{N}\sum_{k=1}^{L_i}\log L(\mathsf{T}|r_{i,k})\right). \tag{40}$$

Note that to compute γ_1, ..., γ_N in each iteration we have to solve N expectation maximization problems with the form of equation 38.

After multi-MLESAC is finished, the inliers of the best candidate model can be found by checking for each sample if its probability of being an inlier is higher than of being an outlier

$$\gamma_i L(\mathsf{T}|r_{i,k}^{\mathcal{I}}) > (1-\gamma_i) L(\mathsf{T}|r_{i,k}^{\mathcal{O}}) \tag{41}$$

which, by observation of equations 34 and 35, can be rewritten as

$$|r_{i,k}|^2 < -2\sigma_i^2 \ln \frac{\sqrt{2\pi}\sigma_i(1-\gamma_i)}{\gamma_i v_i} \,. \tag{42}$$

The most notable difference when we step from a standard MLESAC formulation to multi-MLESAC is that different datasets might have different inlier ratios γ_i. This reflects a practical scenario where some datasets are consistently more reliable than others. Multi-MLESAC is able to capture those differences by estimating separate values γ_i for each dataset, which in turn results in a different cost function and inlier threshold for each dataset.

7.3 Multi-MAPSAC

The MLESAC formulation can be further generalized to a maximum a posteriori problem (MAPSAC [20]). While [20] does a very exhaustive bayesian analysis of random sampling for geometric problems, we are only interested in its key observation that an algorithm from the RANSAC family does not only estimate the parameters of model T but also an additional set of latent parameters, namely by deciding whether each sample is an inlier or an outlier through the expectation maximization of the inlier ratio γ. Taking this into account we formulate the multi-MAPSAC problem as

$$\max_{\mathsf{T},\gamma_1,...,\gamma_N} Pr(\mathsf{T},\gamma_1,...,\gamma_N|\mathbf{R}_1,...,\mathbf{R}_N) \,, \tag{43}$$

where $\mathbf{R}_1,...,\mathbf{R}_N$ represent the residues of all samples from $\mathcal{D}_1,...,\mathcal{D}_N$ respectively, which follow the mixed inlier-outlier distribution of multi-MLESAC. This formulation can be re-written as

$$\max_{\mathsf{T},\gamma_1,...,\gamma_N} Pr(\gamma_1,...,\gamma_N,\mathsf{T}) Pr(\mathbf{R}_1,...,\mathbf{R}_N|\mathsf{T},\gamma_1,...,\gamma_N) \tag{44}$$

Note that although this is a MAP formulation, it is not a step-by-step generalization of the MAPSAC method described in [20], which deals with the marginalization of parameter γ and the effect of additional latent parameters, e. g. reconstructed 3D points. In this paper we do not take these issues into account.

When compared to multi-MLESAC, equation 44 has an additional prior on the model and the latent parameters $Pr(\gamma_1,...,\gamma_N,\mathsf{T})$. Prior knowledge about the model T is a very specific issue in each application scenario and we ignore it in the context of this paper. Our main motivation behind this formulation is to account for prior knowledge about the inlier ratios γ_i. While multi-MLESAC assumes that parameters γ_i are independent from each other, with multi-MAPSAC we want to account for the possibility that this is not the case.

Using prior knowledge on the relative distribution of inlier ratios γ_i is important in the context of our calibration problem. For simplicity, consider the case where there are pairwise correspondences with only two cameras ($N = M = 2$). Correspondences with one camera just give us a fundamental matrix, while the pairwise correspondences with two cameras give us both the extrinsic and intrinsic camera calibration. This means that a candidate model with many inliers in one dataset and very few on the other is a poor solution that is over-fitting to a particular fundamental matrix. To tackle this issue we use the multi-MAPSAC approach and define a prior probability function $Pr(\gamma_A,\gamma_B)$ that penalizes significantly uneven distributions of inliers

$$Pr(\gamma_A,\gamma_B) = (\alpha+1)^2(\gamma_A\gamma_B)^\alpha \,, \tag{45}$$

where the parameter α is set to a value with the same order of magnitude as the number of correspondences in each dataset. Note that the constant factor $(\alpha+1)^2$ is just to guarantee that $Pr(\gamma_A,\gamma_B)$ is a probability density function for γ_A,γ_B between 0 and 1. In the context of maximum a posteriori estimation it can be ignored. These observations also extend to a scenario where there are correspondences with N cameras C_1,...,C_N. Note, however, that the over-fitting case discussed above can only happen when a candidate solution fits well into just one dataset. Thus, the prior term $Pr(\gamma_A,\gamma_B)$ should be computed using only the two highest values from γ_1,...,γ_N.

8 Camera Calibration with N views

The multi-MAPSAC formulation can be used to together with the minimal solution described in section ? to automatically calibrate a new node into a network with N views. In each multi-MAPSAC iteration, we start by randomly sampling two out of N calibrated views and then sample 7 and 4 correspondences from each of those views respectively. Although each candidate solution is generated from two views, we can use it to compute the inlier correspondences across all views and incorporate them into the cost function.

8.1 Algorithm Outline

8.2 Bundle adjustment

A final refinement with bundle adjustment must be used to achieve an optimal solution. Usually bundle adjustment minimizes the re-projection of reconstructed 3D points onto the cameras. However, since

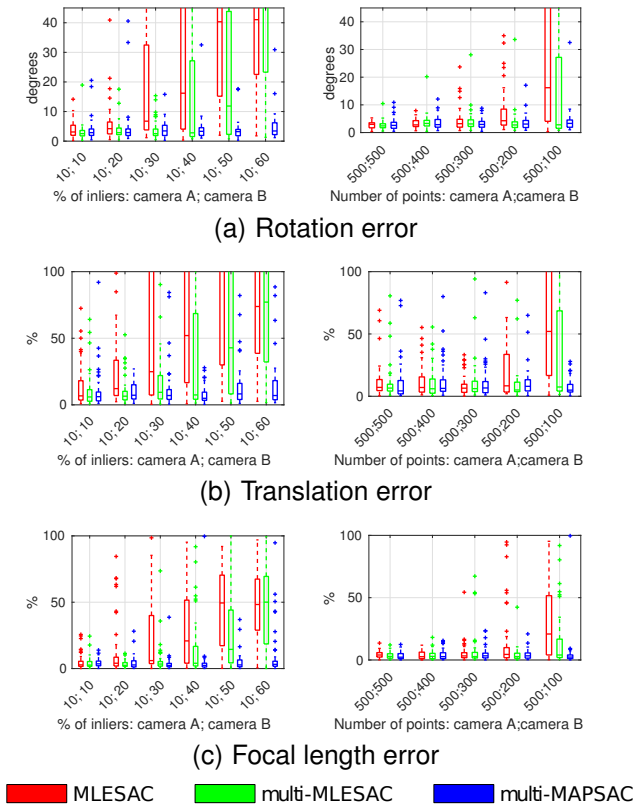


Fig. 7. Comparison between multi-MAPSAC (red) and MLESAC (green) with synthetic data. Error distributions over 50 calibration trials for different levels of injected outliers.

our formulation only uses pairwise correspondences, the introduction of unknown 3D points is an unnecessary burden. As described in [13], an explicit representation of 3D points can be avoided by minimizing the perpendicular distances between point correspondences and their epipolar lines.

Given a pairwise point correspondence $(\mathbf{x}, \hat{\mathbf{x}})$ between two cameras related by a fundamental matrix F, the epipolar error r can measured by the distance in pixels between point $\mathbf{x}$ and the epipolar line of $\hat{\mathbf{x}}$

$$r = \frac{\mathbf{x}^\mathsf{T}\mathsf{F}^\mathsf{T}\hat{\mathbf{x}}}{||\mathtt{I}_{2\times 3}\mathsf{F}^\mathsf{T}\hat{\mathbf{x}}||}. \tag{46}$$

with

$$\mathtt{I}_{2\times 3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{47}$$

Analogously, the distance between point $\hat{\mathbf{x}}$ and the epipolar line of $\mathbf{x}$ is

$$\hat{r} = \frac{\hat{\mathbf{x}}^\mathsf{T}\mathsf{F}\mathbf{x}}{||\mathtt{I}_{2\times 3}\mathsf{F}\mathbf{x}||}. \tag{48}$$

We now consider a network with N calibrated cameras with rotations $\{\mathsf{R}_1, \mathsf{R}_2, ..., \mathsf{R}_N\}$, translations $\{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$, and intrinsics $\{\mathsf{K}_1, \mathsf{K}_2, ..., \mathsf{K}_N\}$ in a common reference frame, and a new camera with unknown parameters R, $\mathbf{t}$, K. The new camera has a set of L_i pairwise correspondences $\{(\mathbf{x}_{i,1}, \hat{\mathbf{x}}_{i,1}), (\mathbf{x}_{i,2}, \hat{\mathbf{x}}_{i,2}), ..., (\mathbf{x}_{i,L_i}, \hat{\mathbf{x}}_{i,L_i})\}$ with each calibrated camera $i = 1, 2, ..., N$. Therefore, the bundle adjustment problem becomes

$$\min_{\mathsf{R},\mathbf{t},\mathsf{K}} \sum_{i=1}^{N}\sum_{k=1}^{L_i} r_{i,k}^2 + \hat{r}_{i,k}^2 \tag{49}$$

with

$$r_{i,k} = \left(\frac{\mathbf{x}_{i,k}{}^\mathsf{T}\mathsf{F}_i{}^\mathsf{T}\hat{\mathbf{x}}_{i,k}}{||\mathtt{I}_{2\times 3}\mathsf{F}_i{}^\mathsf{T}\hat{\mathbf{x}}_{i,k}||}\right) \tag{50}$$

$$\hat{r}_{i,k} = \left(\frac{\hat{\mathbf{x}_{i,k}}{}^\mathsf{T}\mathsf{F}_i\mathbf{x}_{i,k}}{||\mathtt{I}_{2\times 3}\mathsf{F}_i\mathbf{x}_{i,k}||}\right) \tag{51}$$

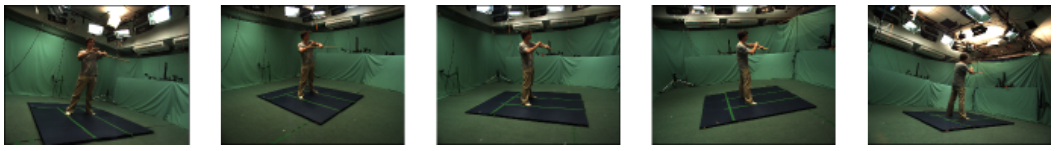
$$\mathsf{F}_i = \mathsf{K}_i[\mathsf{R}_i{}^\mathsf{T}\mathbf{t} + \mathbf{t}_i]_\times \mathsf{R}_i{}^\mathsf{T}\mathsf{R}\mathsf{K}\,. \tag{52}$$

9 Experiments

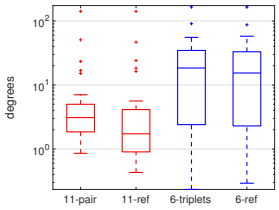
In this section we validate our calibration method using both synthetic data and real imagery of dynamic scenes acquired by a synchronized camera network. Real data was acquired with the Grimage platform [33] that comprises a set of camera nodes in a room. The cameras are calibrated both intrinsically and extrinsically with the method described in [9].

In a first set of experiments we use synthetic data to demonstrate that in challenging scenarios our multi-MAPSAC formulation is essential to obtain accurate calibrations. We then demonstrate the usefulness of our calibration method in practice with two distinct experiments. The first experiment concerns adding, or re-calibrating, a camera node during network operation, using image point correspondences at a certain frame time instant. The second experiment refers to recovering the trajectory and intrinsics of a free moving camera whose acquisition is synchronized with the network. This camera can be a camcorder used to obtain close-ups of an object/person of interest in the scene. In the experiments with real data we compare our 11-point calibration algorithm against the standard 6-point approach [12]. The former uses independent pairwise correspondences with two calibrated views, while the latter requires triple correspondences such that each point in the uncalibrated images is seen by at minimum of two calibrated cameras in order to enable 3D reconstruction.

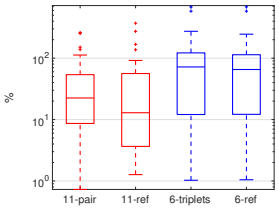
We use SIFT features to establish point correspondences between the images. For both our method and the 6-point approach we perform a pre-filtering step with 7-point fundamental matrix estimation. For our method we use multi-MAPSAC and the bundle adjustment described in section 8.2. On the other hand, the 6-point approach is a single dataset formulation and relies on 3D point estimation, therefore we use the standard versions of both MLESAC and bundle adjustment.



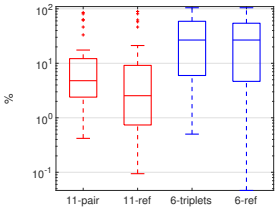
(a) Input images



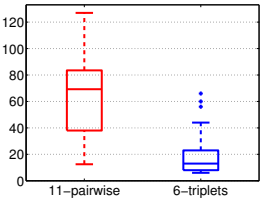
(b) Rotation error



(c) Translation error



(d) Focal length error



(e) Number of inliers

Fig. 8. Addition of a new node to a camera network. In each trial we try to calibrate one of the cameras in (a) assuming that the remaining four are calibrated. (b), (c), (d), (e) show the comparative performance between 11-point (pairwise) and 6-point (triple) for 250 calibration trials.

9.1 Validation of multi-MAPSAC

We built a simulated environment in order to show that in some conditions, our multi-MAPSAC formulation clearly outperforms a standard MLESAC approach that assumes all correspondences belong to the same dataset. Note that comparing multi-MAPSAC against standard MAPSAC instead of MLESAC would not make any difference in the context of this problem since our defined prior probabilities depend exclusively on the assumption that different datasets have different inlier ratios. We generate calibrated cameras C_A, C_B, and an uncalibrated camera C in random poses such that they share a common field of view. Then we generate 500 points that are viewed by cameras C_A and C and 100 points that are viewed by cameras C_B and C. All these correspondences are injected with gaussian noise with 1 pixel standard deviation, and also a predefined ratio of outliers. We tried to calibrate camera C using the 11-point algorithm with both multi-MAPSAC and MLESAC. We performed 50 calibration trials for each of six different levels of injected outliers in cameras C_A and C_B. In Fig 8 we show the error distributions for rotation, translation, and focal length when compared against groundtruth values. It is clear that multi-MAPSAC is able to perform better in situations where the inlier ratios are significantly different in cameras C_A and C_B.

9.2 Addition of a new node to a calibrated network

In this experiment we aim at fully calibrating a camera using pairwise correspondences with a set of frames acquired at the same time instant. We want to compare our 11-point approach using correspondences with only two views against the 6-point approach using all available correspondences across the 5 views. For this purpose we selected a particular moment of the *stick* dataset from the 4drepository [34]. We chose five views that are shown in Fig. 9(a) and tried to calibrate each of them assuming the remaining four were calibrated. The selected camera nodes in this experiment have significant changes in viewpoint, making it very difficult to establish triple correspondences. We want to show that in many situations there are enough pairwise correspondences for our 11-point algorithm to provide accurate results but not enough triple correspondences for the 6-point algorithm to work.

For our 11-point approach we perform the pre-filtering step with the four calibrated cameras and select the two with the highest number of inliers. Since there is a wide baseline between the five cameras the two closest cameras typically produce the majority (if not all) the reliable correspondences. In the case of the 6-point algorithm all pre-filtered triple correspondences from the four cameras are used.

After the pre-filtering step, and for each of the five cameras, the calibration with the two methods is carried 50 times, summing up to 250 calibration tests for each approach. The error distributions for all calibration attempts are provided in Figs. 9(b),9(c),9(d). Note that the errors are displayed in logarithmic scale and our algorithm provides extremely more accurate results than the 6-triplets approach, which completely fails to provide a reasonable calibration in most cases. This can be explained by the fact that it is possible to establish a much higher number of pairwise correspondences than triple correspondences (Fig. 9(e)), despite the fact that triple correspondences are established across the four calibrated cameras, while for our algorithm we only use the pairwise correspondences from two cameras.

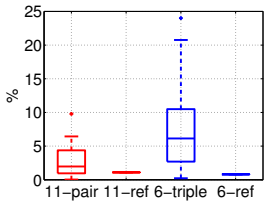
9.3 Calibration of a hand-held camera

We acquired a set of synchronized video sequences using two nodes of the calibrated network and a hand-held moving camera. Each sequence is composed of 30 frames in which the hand-held camera shares its field of view with two other calibrated cameras (Fig. 10). In comparison with the previous experiment the calibrated camera nodes have a smaller baseline and the viewed scene is richer in features. This benefits the standard 6-point approach, as it is easier to establish triple correspondences. However, the viewed scene is highly dynamic and contains significant occlusions in some frames, making it difficult to establish triple correspondences. The intrinsic parameters of the hand-held camera were previously determined using the method described in [9] and we use these values as groundtruth for comparison with our estimates.

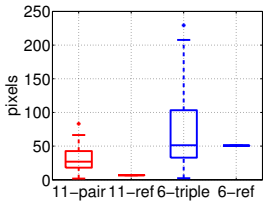
Both the intrinsic parameters and the trajectory of the hand-held camera are recovered with both our 11-point method and the 6-point method. In a first step, we calibrate each frame independently using pairwise correspondences with the synchronized frames from the calibrated cameras. This is convenient for the case of a hand-held camera with motorized lenses for which the zoom varies while moving. However, since in this experiment we know that the camera intrinsics are stationary, a final estimation with bundle adjustment is made assuming a single set of intrinsics for all frames.

The error distribution for the intrinsic parameters before and after global refinement is presented in Fig. 10(b) and 10(c). Note that although the results for the standard 6-point approach are worse than our 11-point approach, they are significantly better than in the previous experiment. As explained earlier this is to be expected, since it is easier to establish correspondences in this set of acquisitions. Although the initialization results are sufficient for the focal length to converge to similarly accurate values with both algorithms, our algorithm is able to provide a much better estimation of the principal point. Our estimated camera trajectory is also significantly smoother and in line with a reasonable hand-held trajectory (Fig. 10(e)), specially when significant occlusions occur (e. g. the leftmost frames in Fig. 10(d)). Since we do not have groundtruth values for the camera trajectory, in Fig. **??** both trajectories are projected onto the image plane of a third calibrated camera in which the person handling the free camera is visible. For the selected frames it is quite clear that our estimations with the 11-point approach (red) are significantly more accurate than with the 6-point approach (blue). This confirms the intuition from Fig. 10(e) that our algorithm provides accurate trajectory estimations.

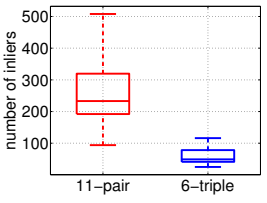
(a) Sample Frames



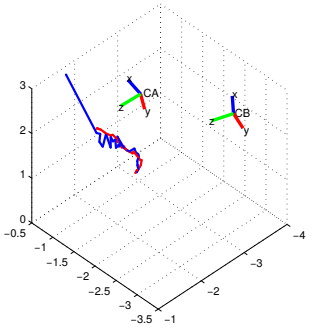
(b) Focal length



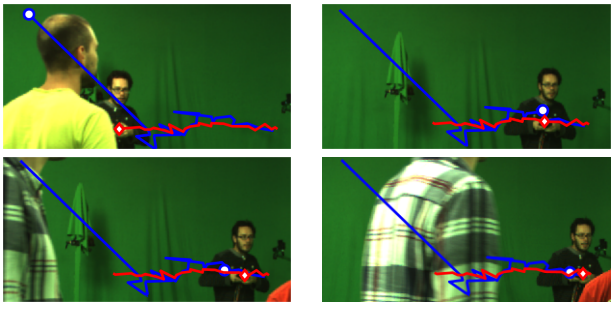
(c) Principal point



(d) Number of inliers



(e) Camera trajectory



(f) Projected Trajectory

Fig. 9. Online calibration of a hand-held camera.

9.4 Addition of new nodes to an SfM reconstruction

10 Conclusion

We presented a new minimal solution for the intrinsic and extrinsic calibration of a camera from pairwise correspondences with other two calibrated cameras. We observed that our algorithm requires certain modifications to the standard RANSAC formulation and provided a random sampling framework for extracting a model from multiple datasets. We have

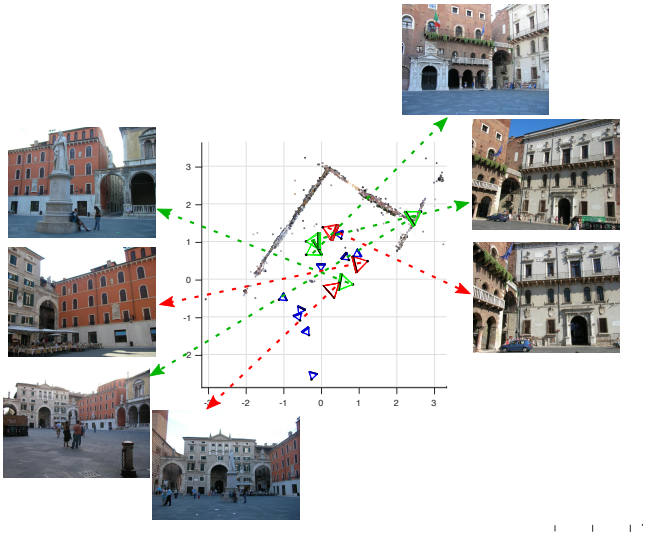


Fig. 10.

shown that within a camera network scenario there are cases in which pairwise correspondences must be used to calibrate a new camera since there are very few correspondences simultaneously seen across more than two cameras. This makes our minimal solution outperform previous calibration methods that rely on triple correspondences.

REFERENCES

[1] J. Starck and A. Hilton, "Surface capture for performance-based animation," *Computer Graphics and Applications, IEEE*, vol. 27, no. 3, pp. 21 –31, 2007.

[2] Z. Zhao and Y. Liu, "Practical multi-camera calibration algorithm with 1d objects for virtual environments," in *Multimedia and Expo, 2008 IEEE International Conference on*, 2008, pp. 1197 –1200.

[3] E. Shen and R. Hornsey, "Multi-camera network calibration with a non-planar target," *Sensors Journal, IEEE*, vol. 11, no. 10, pp. 2356 –2364, 2011.

[4] J. Courchay, A. Dalalyan, R. Keriven, and P. Sturm, "A global camera network calibration method with linear programming," in *Proceedings of the International Symposium on 3D Data Processing, Visualization and Transmission*, 2010.

[5] R. Kumar, A. Ilie, J.-M. Frahm, and M. Pollefeys, "Simple calibration of non-overlapping cameras with a mirror," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–7.

[6] R. Rodrigues, J. a. P. Barreto, and U. Nunes, "Camera pose estimation using images of planar mirror reflections," in *Computer Vision ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6314, pp. 382–395.

[7] T. Svoboda, D. Martinec, and T. Pajdla, "A convenient multicamera self-calibration for virtual environments," *Presence: Teleoper. Virtual Environ.*, vol. 14, no. 4, pp. 407–422, 2005.

[8] J. Barreto and K. Daniilidis, "Wide area multiple camera calibration and estimation of radial distortion," in *OMNIVIS'2004 - Int. Workshop in Omnidirectional vision, camera networks, and non-conventional cameras*, 2004.

[9] A. Zaharescu, R. Horaud, R. Ronfard, and L. Lefort, "Multiple camera calibration using robust perspective factorization," in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, 2006, pp. 504 –511.

[10] G. Thomas, "Real-time camera tracking using sports pitch markings," *Journal of Real-Time Image Processing*, vol. 2, no. 2-3, pp. 117–132, 2007.

[11] J. Puwein, R. Ziegler, J. Vogel, and M. Pollefeys, "Robust multi-view camera calibration for wide-baseline camera networks," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, Jan 2011, pp. 321–328.

[12] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge Academic Press, 2003.

[13] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An invitation to 3-D vision: from images to geometric models*. Springer, 2004.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[15] B. Clipp, C. Zach, J.-M. Frahm, and M. Pollefeys, "A new minimal solution to the relative pose of a calibrated stereo camera with small field of view overlap," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1725–1732.

[16] C. Raposo, M. Lourenço, J. P. Barreto, and M. Antunes, "Plane-based odometry using an rgb-d camera," in *BMVC*, 2013.

[17] C. Raposo, M. Antunes, and J. P. Barreto, "Piecewise-planar stereoscan:structure and motion from plane primitives," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8690, pp. 48–63.

[18] F. Vasconcelos, J. P. Barreto, and E. Boyer, "A minimal solution for camera calibration using independent pairwise correspondences," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 724–737.

[19] P. H. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[20] P. H. S. Torr, "Bayesian model estimation and selection for epipolar geometry and generic manifold fitting," *International Journal of Computer Vision*, vol. 50, no. 1, pp. 35–61, 2002.

[21] J. Kim, L. Hodong, and R. Hartley, "Motion Estimation for Nonoverlapping Multicamera Rigs: Linear Algebraic and Linf Geometric Solutions," *IEEE Trans. in Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1044–1058, 2010.

[22] H. Stewénius, D. Nistér, M. Oskarsson, and K. Åström, "Solutions to minimal generalized relative pose problems," in *Workshop on Omnidirectional Vision*, Beijing China, 2005.

[23] R. Pless, "Using many cameras as one," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2003.

[24] D. Nister, "An efficient solution to the five-point relative pose problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 756–770, June 2004.

[25] F. Vasconcelos and J. P. Barreto, "Towards a minimal solution for the relative pose between axial cameras," in *BMVC*, 2013.

[26] R. Hartley and P. Sturm, "Triangulation," *Computer Vision and Image Understanding*, 1997.

[27] P. Sturm and B. Triggs, "A factorization based algorithm for multi image projective structure and motion," in *European Conference in Computer Vision*, 1996.

[28] N. Levi and M. Werman, "The viewing graph," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, June 2003, pp. I–518–I–522 vol.1.

[29] K. Josephson, M. Byrod, F. Kahl, and K. Åström, "Image-based localization using hybrid feature correspondences," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1 –8.

[30] P. Sturm, S. Ramalingam, J.-P. Tardif, S. Gasparini, and J. a. Barreto, "Camera models and fundamental concepts used in geometric computer vision," *Found. Trends. Comput. Graph. Vis.*, vol. 6, no. 1–2, pp. 1–183, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1561/0600000023

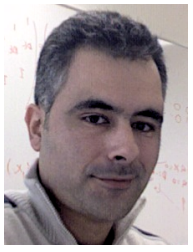
[31] H. Pottmann and J. Wallner, *Computational line geometry*, 1st ed. Berlin: Springer Verlag, 2001.

[32] J. P. Barreto, "General central projection systems: Modeling, calibration and visual servoing," Ph.D. dissertation, PhD Thesis, University of Coimbra, Coimbra, Portugal, 2004.

[33] J. Allard, J.-S. Franco, C. Merrier, E. Boyer, and B. Raffin, "The grimage platform: A mixed reality environment for

interactions," in *Computer Vision Systems, 2006 ICVS'06. IEEE International Conference on*. IEEE, 2006, pp. 46–46.
[34] "4d repository," http://4drepository.inrialpes.fr/pages/home.

Francisco Vasconcelos received the PhD degree from the University of Coimbra, Portugal, in 2016. He is a Postdoctoral Associate Researcher at the University College London, United Kingdom, as part of the Centre for Medical Image Computing and the Surgical Robot Vision group. His current research interests focus on geometry problems in computer vision, medical imaging, and robotics.



João P. Barreto received the Licenciatura and Ph.D. degrees from the University of Coimbra, Coimbra, Portugal, in 1997 and 2004, respectively. From 2003 to 2004, he was a Postdoctoral Researcher with the University of Pennsylvania, Philadelphia. Since 2004, he has been an Assistant Professor with the University of Coimbra, where he is also a Senior Researcher with the Institute for Systems and Robotics. He is the author of more than 30 peer reviewed publications. His current research interests include different topics in computer vision, with a special emphasis in geometry problems and applications in robotics and medicine. Dr. Barreto is a regular Reviewer for several conferences and journals, having received 4 Outstanding Reviewer Awards in the last few years.

Edmond Boyer received the PhD degree from the Institut National Polytechnique de Lorraine, France, in 1996. He is an associate professor at Grenoble Universities, France. He started his professional career as a research assistant in the Department of Engineering, University of Cambridge, United Kingdom. He joined INRIA Grenoble in 1998. His fields of competence cover computer vision, computational geometry, and virtual reality. He is a cofounder of the 4D View Solution Company in the domain of spatiotemporal modeling. His current research interests are in 3D dynamic modeling from images and videos, motion capture and recognition from videos, and immersive and interactive environments.