Line-based Robust SfM with Little Image Overlap

Yohann Salaün[1,2], Renaud Marlet[1] and Pascal Monasse[1]

[1] LIGM, UMR 8049, École des Ponts ParisTech, UPE (France)   [2] CentraleSupélec (France)

{yohann.salaun,renaud.marlet,pascal.monasse}@enpc.fr

Abstract

Usual Structure-from-Motion (SfM) techniques require at least trifocal overlaps to calibrate cameras and reconstruct a scene. We consider here scenarios of reduced image sets with little overlap, possibly as low as two images at most seeing the same part of the scene. We propose a new method, based on line coplanarity hypotheses, for estimating the relative scale of two independent bifocal calibrations sharing a camera, without the need of any trifocal information or Manhattan-world assumption. We use it to compute SfM in a chain of up-to-scale relative motions. For accuracy, we however also make use of trifocal information for line and/or point features, when present, relaxing usual trifocal constraints. For robustness to wrong assumptions and mismatches, we embed all constraints in a parameterless RANSAC-like approach. Experiments show that we can calibrate datasets that previously could not, and that this wider applicability does not come at the cost of inaccuracy.

1. Introduction

Structure-from-Motion (SfM) has made spectacular improvements concerning scalability [8, 42, 14], accuracy [24, 5, 2, 39] and robustness [41, 26]. However, most approaches assume a significant amount of overlap between images. In this work, we consider the case where images have only little overlap, possibly as low as two images at most seeing the same part of the scene.

This situation commonly occurs when a scene is photographed by people with little or no knowledge in photogrammetry. Even when informed, they can make occasional mistakes, widening too much a baseline, or shooting a low-quality image that cannot be exploited and has to be skipped. Another example is when exploiting street views, e.g., taken from a vehicle, if the viewpoints are too distant one from another, or if the camera is too close to the facades because of narrow sidewalks. In such cases, SfM methods may yield partial or fragmented calibrations.

It also applies to situations where only a small number of pictures can be shot, because of physical or time constraints. It may be the case when digitizing a building which is still in use, not to disturb occupants. Reducing the number of images, to only a few per room for indoor scenes, is also a way to reduce the cost and time for acquiring and processing information, as long as a minimum level of accuracy can still be reached for the targeted application.

Another issue concerns the lack of texture in environments such as building interiors, as it greatly reduces the amount of feature points detected in images, also leading to uneven feature distributions. Besides, if the number of images is small, it is likely that the baselines are wide, as well as the view angles between overlapping images. Consequently, because of perspective distortion, point descriptors are harder to match; relaxing matching thresholds does not fix the problem, as it introduces outliers. Moreover, detected points may unknowingly lie on a single plane, possibly giving rise to degenerate configurations for camera registration. On the contrary, lines do not suffer from lack of texture and are prevalent in interiors, where they often occur at plane intersections and object boundaries, see Fig. 1. Furthermore, lines are robust to significant changes of viewpoints, although their matching eventually also degrades.

Our approach thus leverages on lines for pose estimation and reconstruction, although it also exploits points when available. Our main contributions are as follow:

- Given two successive bifocal calibrations, i.e., sharing a middle camera, we propose a novel method for computing their relative scale without the need for trifocal features. It is based on line coplanarity hypotheses, which is relevant in man-made scenes (cf. Sect. 4).

- To however exploit trifocality when present, we show how to relax usual trifocal constraints, and integrate them with coplanarity constraints in a unified framework. The relaxed constraint only needs 2 bifocal calibrations among a feature triplet (cf. Sect. 5).

- Robustness to outliers in this context (cf. Sect. 6) requires parameters that may be hard to set given the variety of scenes and constraint types. We propose a parameterless method that integrates all constraints and adapts automatically to the scene diversity (cf. Sect. 7).

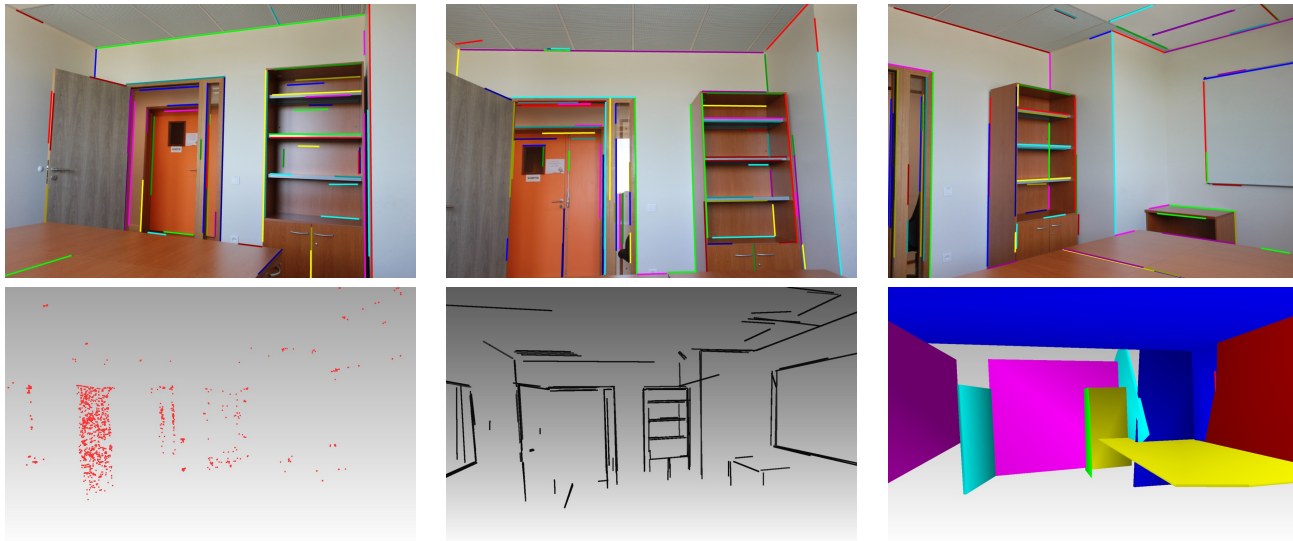- Using both standard SfM datasets with ground truth

Figure 1. Top row: 3 consecutive views of the Office-P19 dataset, with detected line segments. Bottom row: views of this area with reconstructed structures: points (left), lines (middle), bounding boxes of planes associated to clusters of coplanar lines after BA (right).

and difficult interior scenes with little overlap and little texture, we validate empirically our technical choices and we show (1) that we can calibrate images that cannot be calibrated with existing point- or line-based SfM methods, and (2) that our accuracy is on par with state-of-the-art SfM methods on less hard datasets.

Our method does *not* assume any Manhattan configuration.

2. Related Work

Incremental SfM methods [36, 23, 42] register a new image to a partial model already constructed, using 3D-2D correspondences. A number of Perspective-n-Points (PnP) algorithms [19, 10] have been proposed for solving this resection problem. Whatever the method, at least 3 correspondences common to 3 images are required: three 3D points visible in the same 3 images, and up to minimum of 6 [49]. Only [48] estimates the pose of a pinhole camera just from 2D matches, considering a generalized camera model.

Hierarchical SfM methods, that additionally merge partial models, also have similar constraints. In [13], the two models to merge are overlapping in that they share one or several images, and pairs of 3D points projecting to the same 2D features in both models are used to relate the models, implying tracks of length 3 or more. In [40], models are merged via feature matches between images separately associated to each model, requiring that the same four 3D points are reconstructed in both models, implying tracks of length 4 (connections between 2 tracks of length at least 2). Even with relaxed requirements where merging uses 4 points that are seen but not necessarily reconstructed in the other model [9], feature tracks of length 3 are required.

As for global SfM methods, their main objective is merg-ing relative motions between two cameras into a consistent graph of all cameras. Besides robustness concerns, to get rid of outlier edges, and various approaches to average rotations, one of the main issues is that the relative translations are only given up to an unknown scale factor; only their directions are known. Most methods to infer global translations rely on information redundancy assuming a densely-connected graph [11, 5], or on additional information from trifocal tensors [35, 24] (hence requiring 4 tracked points across 3 views). A number of other methods [17, 29, 1] compute the global translations, possibly along with the 3D points, by solving equations relating points visible in two images; however, they implicitly assume that enough points are visible in at least 3 images to cancel the degrees of freedom of the relative scale factors. Besides, they do not all address point match outliers.

The situation is similar with line-based SfM. In [46], an initial image triplet with common line matches is required. Then, given a partial model, Perspective-n-Line (PnL) methods estimate the pose of a camera in which three 3D lines reproject [21, 47], which implies that at least 3 lines are visible in 3 views. A minimum of 6 lines is some-times even desirable to prevent noise sensitivity [21], if not 9 for applicability and 25 for accuracy [27].

More generally, when associating both points and lines in a "Perspective-n-Features" framework, a minimum of 3 features visible in 3 views is still required [28, 43].

Our approach for relating the scale of two bifocal calibrations is based on coplanar line pairs. To our knowledge, coplanar lines have been used for pose estimation, but only in a two-view context and with a Manhattan-world assumption, to identify planar structures [18]. A related topic is plane-based SfM, but it has mostly been studied assuming

prior knowledge (user-given) about the scene planes [38, 3] or in tracking scenarios with videos [50]. In [30, 17], a reference plane is used to estimate both the global translations and 3D points, but it must be visible in all images.

A related work regarding the estimation of scale factors and the identification of planar structures concerns direct structure estimation (DSE) via homography estimations, with the computation of coplanar point clusters, but it does not estimate poses and it also relies on trifocal points [16].

Line triangulation and line bundle adjustment have been well studied given an initial global pose estimation [4], but not associated to coplanarity issues.

3. From Relative to Global Pose Estimation

In the epipolar graph, for every edge between camera i and camera j, we assume the relative pose (R_{ij}, t_{ij}) known (estimated), where R_{ij} is the relative rotation and t_{ij} is the unit norm translation direction. We want to estimate the scale factors λ_{ij} relating t_{ij} to the global relative translations $T_{ij} = \lambda_{ij} t_{ij}$, that is, up to a single global scale factor.

In the following, we only consider the case of a chain or cycle of cameras, for which we estimate global poses (R_j, T_j), where rotations R_j, translations T_j and camera centers C_j are defined in the same reference frame. Yet, our pose estimation method can be integrated in a general global SfM framework for arbitrary graphs, e.g., as described in [34] to evenly distribute errors over the whole graph in trying to satisfy the consistency constraints:

$$R_j = R_{ij} R_i, \quad T_j = R_{ij} T_i + T_{ij}, \quad C_j = -R_j^\top T_j. \quad (1)$$

This could be associated to a method to remove outlier edges and enforce cycle consistency [12, 44, 7]. In a chain of cameras, global motions are recursively defined as:

$$R_1 = I \quad \text{and} \quad R_{j+1} = R_{j,j+1} R_j \quad (2)$$
$$T_1 = 0 \quad \text{and} \quad T_{j+1} = R_{j,j+1} T_j + \lambda_{j,j+1} t_{j,j+1} \quad (3)$$

As the global pose remains defined up to an unknown scale factor, we additionally set $\lambda_{12} = 1$: distances are thus defined with unit length λ_{12}. (In case of a cycle, we could also include epipolar constraints to close the loop and distribute errors as in [34], but we do not in our implementation.) Finally, a bundle adjustment refines the initial pose estimation.

To simplify notations, we assume features are normalized: camera intrinsic matrices $K_j = I$. We denote a triplet of successive cameras $1, 2, 3$ rather than $j, j+1, j+2$.

4. Coplanarity Constraint

Let L_a and L_b be two non-parallel 3D line segments in a plane $\mathcal{P}$. (Coplanarity here is an hypothesis to be later validated, see Sect. 6.) Suppose L_a only appears on cameras

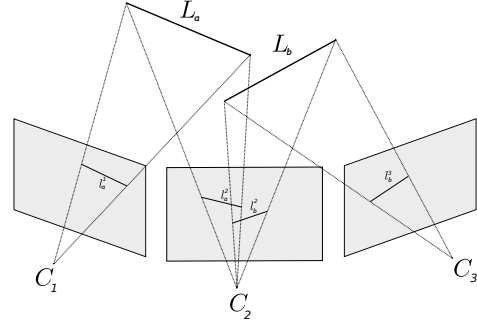Figure 2. If L_a and L_b, without trifocal overlap, are known to be coplanar, it is enough to place C_1, C_2, C_3 in the same reference frame knowing only bifocal calibrations for cameras 1-2 and 2-3.

1 and 2, and L_b only appears on cameras 2 and 3. Let l_i^j be the projection of L_i on camera j. (See Fig. 2.)

The coordinates of a 3D point P in the global reference frame relates to its projection p^j on camera j and depth z_p^j:

$$P = R_j^\top (z_p^j p^j - T_j) \quad (4)$$

Assuming $P \in L_i$, we also have for any camera k:

$$l_i^k \cdot (R_k P + T_k) = 0, \quad (5)$$

where "$\cdot$" denotes scalar product. Combining (4) and (5), and relating global to relative motions using (1), we can express the depth z_p^j in terms of the relative pose information:

$$z_p^j = -\frac{l_i^k \cdot T_{jk}}{l_i^k \cdot (R_{jk} p^j)}. \quad (6)$$

We also know that the normal to plane $\mathcal{P}$ is given by:

$$n_\mathcal{P} = d_{L_a} \times d_{L_b} = (R_1^\top l_a^1 \times R_2^\top l_a^2) \times (R_2^\top l_b^2 \times R_3^\top l_b^3) \quad (7)$$

where d_{L_i} is the 3D direction of line L_i. Let M be a point in $\mathcal{P}$. As any line in $\mathcal{P}$ is orthogonal to $n_\mathcal{P}$, we have:

$$P \in \mathcal{P} \Leftrightarrow n_\mathcal{P} \cdot (P - M) = 0 \quad (8)$$

From (8), since $P \in L_i \subset \mathcal{P}$, using (4) and (6) yields:

$$n_\mathcal{P} \cdot (C_j - M) = \frac{(l_i^k \cdot T_{jk})(n_\mathcal{P} \cdot (R_j^\top p^j))}{l_i^k \cdot (R_{jk} p^j)} \quad (9)$$

Now using (9) both for a point $P_a \in L_a$ with $j = 2$ and $k = 1$, and a point $P_b \in L_b$ with $j = 2$ and $k = 3$, we get:

$$\frac{\lambda_{23}}{\lambda_{21}} = \frac{(l_b^3 \cdot (R_{23} p_b^2))(n_\mathcal{P} \cdot (R_2^\top p_a^2))(l_a^1 \cdot t_{21})}{(l_a^1 \cdot (R_{21} p_a^2))(n_\mathcal{P} \cdot (R_2^\top p_b^2))(l_b^3 \cdot t_{23})} \quad (10)$$

Eq. (10) expresses the signed ratio of the distances between 3 successive camera centers in the same reference frame, in terms of only 2-view relative pose information, plus weak information about 2 coplanar lines that do not have to be visible in all 3 views. It has 3 degenerate configurations that

can be assessed and controlled with an angular threshold (please see the supplementary material for details).

Note that this criterion does not depend on the endpoints of line segments, that are notoriously inaccurate. Line segment matches can actually be grossly wrong because of over-segmentation, which is a common weakness of line segment detectors. On the contrary, our criterion only relies on (infinite) lines, which are much more stable. It can cope with line segments matches that are wrong w.r.t. endpoints but correct w.r.t. the supporting 3D line.

5. Simplified Trifocal Constraints, When Any

Absolutely no trifocal feature is required. Yet if a line or point is visible in three successive views 1-2-3, which is a rich information, we consider a corresponding simplified trifocal constraint, that also involves the scale factors $\lambda_{12}, \lambda_{23}$. For both kinds of features, we do as follows:

1. We set arbitrarily the first two camera centers given their relative pose: $C_1 = 0$, $\lambda_{12} = 1$, $C_2 = -R_2^T T_{12}$.

2. We compute the approximate 3D reconstruction of the feature from the first 2 camera poses, as well as its projection on the third camera.

3. We compute the scale factor λ_{23} such that the projected feature on camera 3 corresponds best to the detection.

As in step 1 we set $\lambda_{12} = 1$, the computed scale factor λ_{23} is actually also the ratio $\tau_{123} = \lambda_{23}/\lambda_{12}$. For a more robust and accurate estimation, we can symmetrize the setting: permuting the role of cameras and averaging the resulting scale factors. As we just consider here a chain of successive views, it makes sense to actually only take into account the case where cameras 1 and 3 are swapped, keeping camera 2 as central. It prevents camera configurations 1-3 or 3-1, which are likely to feature wider changes of viewpoints, i.e., wider baselines and view angles. Bifocal calibration 1-3 may even fail and not be available. (In our implementation, we thus only compute the ratios τ_{123} and τ_{321}, retaining the average $\tau = (\tau_{123} + 1/\tau_{321})/2$.) Note that, as opposed to usual trifocal constraints that require at least three trifocal features, one is enough here.

In step 3, as detailed below, the scale factor λ_{23} is expressed in the form:

$$\lambda_{23} = \arg\min_{\lambda \in \mathbb{R}} \frac{\|u \times (v + \lambda w)\|}{\|u\| \, \|v + \lambda w\|}, \quad (11)$$

where u, v, w are known vectors in $\mathbb{R}^3$. To find the minimum, we look for the values of λ that make the derivative vanish. It leads to a third-degree polynomial, that simplifies into a second-degree polynomial, whose roots can be tested to find the minimum of the original expression. As above, this formula has a degenerate configuration that can be checked and discarded using an angular threshold.

Trifocal point constraint. Given a triplet of matched points $\hat{p} = (p_1, p_2, p_3)$ in cameras 1-2-3, we can estimate the corresponding 3D point $\tilde{P}$ from p_1, p_2, assuming $\lambda_{12} = 1$, and reproject $\tilde{P}$ as $\tilde{p}_3$ on camera 3, for some given λ_{23}:

$$\tilde{p}_3 = R_3(\tilde{P} - C_3) = R_3(\tilde{P} - C_2) - \lambda_{23}t_{23}$$

We look for the scale factor λ_{23}^* that makes $\tilde{p}_3$ the closest to p_3. Rather than minimizing the point distance, we minimize the angle $\widehat{p_3 C_3 \tilde{p}_3}$, or instead $\|p_3 \times \tilde{p}_3\|/(\|p_3\| \, \|\tilde{p}_3\|)$:

$$\lambda_{23}^* = \arg\min_{\lambda_{23} \in \mathbb{R}} \frac{\|p_3 \times (R_3(\tilde{P} - C_2) - \lambda_{23}t_{23})\|}{\|p_3\| \, \|R_3(\tilde{P} - C_2) - \lambda_{23}t_{23}\|}, \quad (12)$$

which is in the form (11), and allows the exact minimum to be explicitly and rapidly computed.

Trifocal line constraint. Given a triplet of matched line segments $\hat{l} = (l_1, l_2, l_3)$ in cameras 1-2-3, we can estimate the corresponding 3D line $\tilde{L}$ from l_1, l_2, assuming $\lambda_{12} = 1$, and reproject $\tilde{L}$ as $\tilde{l}_3$ on camera 3, for some given λ_{23}. As $\tilde{l}_3$ also is the normal of the plane defined by C_3 and $\tilde{L}$, we have for any point $\tilde{P}$ on $\tilde{L}$:

$$\tilde{l}_3 \propto R_3[d_{\tilde{L}} \times (\tilde{P} - C_3)]$$

We look for the scale factor λ_{23}^* that makes $\tilde{l}_3$ the most colinear with observation l_3, minimizing their relative angle:

$$\lambda_{23}^* = \arg\min_{\lambda_{23} \in \mathbb{R}} \frac{\|l_3 \times (R_3[d_{\tilde{L}} \times (\tilde{P} - C_2)] - \lambda_{23}R_3[d_{\tilde{L}} \times t_{23}])\|}{\|l_3\| \, \|R_3[d_{\tilde{L}} \times (\tilde{P} - C_2)] - \lambda_{23}R_3[d_{\tilde{L}} \times t_{23}]\|} \quad (13)$$

which is in the form of (11). As for coplanarity (Sect. 4), this constraint does not depend on line segment endpoints.

6. Robust Estimation

A single pair of coplanar 3D lines as in Fig. 2 is enough to determine a scale factor (Eq. (10)), but confirming it calls for more lines. As there is no oracle to safely pick non-parallel coplanar lines (cf. Sect. 4), we adopt a RANSAC-like approach to sample candidate line pairs and select the associated scale λ_{23} with which the largest number of pairs agree. More generally, we also sample and check agreement w.r.t. trifocal points and lines when any (cf. Sect. 5), which provides robustness to wrong detections and matching too.

For this, we define a measure of residual error for the three different kinds of features (hypothesized coplanar line pairs, trifocal points, trifocal lines), given a presumed $\lambda_{23} = \lambda$ obtained from the sample (discarding degenerate cases).

Coplanar lines. For any quadruplet of line segments $\check{l} = (l_a^1, l_a^2, l_b^2, l_b^3)$ s.t. l_a^1, l_a^2 match in cameras 1-2 and l_b^2, l_b^3 match in cameras 2-3, we estimate the associated 3D lines L_a, L_b and consider the 3D point $P_{ab} \in L_a$ (resp. $P_{ba} \in L_b$) that is the closest to L_b (resp. L_a) in 3D, with $P_{ab} = P_{ba}$ if L_a, L_b are coplanar. We then consider their reprojection

$\tilde{p}^2_{ab}, \tilde{p}^2_{ba}$ on the shared camera 2. The residual error is their pixel distance: $d_{\mathsf{co}}(\check{l}) = d_{\mathsf{co}}(L_a, L_b) = d(\tilde{p}^2_{ab}, \tilde{p}^2_{ba})$.

To avoid degenerate cases with mostly parallel lines, we also discard line pairs whose 3D directions are similar. (In our experiments, we use a threshold of $15°$.) Note that these directions can be computed from the global rotations only, before global translations are estimated. As there can be many pairs candidate for coplanarity, we only consider, for each segment in camera 2, having a match in camera j, its N closest neighbors having a match in the other camera, with a distance defined as the minimum distance between segment endpoints. (In our experiments, $N=10$).

Trifocal point error. For each triplet of matched points $\hat{p} = (p_1, p_2, p_3)$ in cameras 1-2-3, we estimate the corresponding 3D point P from p_1, p_2, and reproject it as $\tilde{p}_3$ on camera 3 (assuming λ_{23}), as in Sect. 5. The residual error is the pixel distance of reprojection $\tilde{p}_3$ to the observed detection p_3. For robustness, we actually symmetrize this measure by swapping cameras 1 and 3, and averaging the distances: $d_{\mathsf{pt}}(\hat{p}) = (d(\tilde{p}_1, p_1) + d(\tilde{p}_3, p_3))/2$.

Trifocal line error. For each triplet of matched line segments $\hat{l} = (l_1, l_2, l_3)$ in cameras 1-2-3, we estimate the corresponding 3D line L from l_1, l_2, and reproject it as $\tilde{l}_3$ on camera 3 (assuming λ_{23}), as in Sect. 5. The residual error is the average of the pixelic distance between the reprojected (infinite) line $\tilde{l}_3$ and the two endpoints of the detected segment l_3, as in [4, 46]. For robustness, we also swap cameras 1 and 3, and define $d_{\mathsf{seg}}(\hat{l})$ as the average of both errors.

7. Parameter-Free Robust Estimation

Rather than depend on fixed arbitrary error thresholds to define feature agreement with a model, we actually resort to an *a contrario* (AC) approach [22, 23]: we compute the expectation of the number of false alarms (NFA), that measures the statistical meaningfulness (actually the converse, i.e., the insignificance) of λ_{23} w.r.t. to the features to test, and select the scale λ_{23} with the lowest NFA. It allows an automatic optimization of the inlier-outlier threshold, and thus more accurate results too. This is also consistent with the method we use for two-view pose estimation [33].

Coplanar line NFA. For coplanarity, we follow [33] and define a line error from the error for line pairs:

$$d_{\mathsf{co}}(L_i, \lambda) = \min_{L_j \text{ coplanar with } L_i} d_{\mathsf{co}}(L_i, L_j, \lambda) \qquad (14)$$

where $d_{\mathsf{co}}(L_i, L_j, \lambda)$ is the coplanarity distance defined in Sect. 6. The NFA for our sampling is then, following [23]:

$$NFA_{\mathsf{co}}(\lambda) = (n_2-2) \min_{k \in [3, n_{\mathsf{co}}]} n_2 N \binom{n_2}{k-2} \left[\frac{\pi\, d_{\mathsf{co}}(L_k, \lambda)^2}{\mathcal{A}} \right]^{k-2} \qquad (15)$$

where n_2 is the number of lines in camera 2 that have at least a match in camera 1 or 3, $\mathcal{A}$ is the image area, and L_k is the k-th best inlier (with lowest error).

Trifocal point NFA. For triplets of matched points, following [23], we have:

$$NFA_{\mathsf{pt}}(\lambda) = (n_{\mathsf{pt}}-1) \min_{k \in [2, n_{\mathsf{pt}}]} \binom{n_{\mathsf{pt}}}{k} k \left[\frac{\pi\, d_{\mathsf{pt}}(\hat{p}_k, \lambda)^2}{\mathcal{A}} \right]^{k-1} \qquad (16)$$

where n_{pt} is the total number of matched point triplets in cameras 1-2-3, $\mathcal{A}$ is the image area, $d_{\mathsf{pt}}(\hat{p}, \lambda)$ is the residual error of triplet $\hat{p}$ assuming scale λ (cf. Sect. 6), and $\hat{p}_k$ is the k-th lowest error.

Trifocal line NFA. For triplets of matched line segments, also following [23], we have:

$$NFA_{\mathsf{seg}}(\lambda) = (n_{\mathsf{seg}}-1) \min_{k \in [2, n_{\mathsf{seg}}]} \binom{n_{\mathsf{seg}}}{k} k \left[\frac{2\mathcal{D}\, d_{\mathsf{seg}}(\hat{l}_k, \lambda)}{\mathcal{A}} \right]^{k-1} \qquad (17)$$

where n_{seg} is the number of matched line triplets, $\mathcal{D}$ the image diagonal length, $d_{\mathsf{seg}}(\hat{l}, \lambda)$ the residual error of triplet $\hat{l}$ assuming scale λ (cf. Sect. 6), and $\hat{l}_k$ the k-th lowest error.

Global NFA. As these NFAs correspond to expectations of independent events, the global NFA is their product:

$$NFA(\lambda) = NFA_{\mathsf{co}}(\lambda).NFA_{\mathsf{pt}}(\lambda).NFA_{\mathsf{seg}}(\lambda). \qquad (18)$$

It estimates the insignificance of a candidate ratio λ based on coplanarity and trifocal constraints, when any. For robust estimation, we retain the λ with the overall lowest NFA. It defines a parameterless AC-RANSAC variant to Sect. 6.

8. Bundle Adjustment

The last step of our SfM method is a bundle adjustment (BA) that refines simultaneously the structure and the poses.

Reconstructed structure. In our context, we could consider as structures not only points and lines, but also planes corresponding to coplanar lines. Indeed, just as tracks of points or lines across images represent single structures, single planes could be associated to sets of lines sharing a coplanarity constraint. More precisely, only coplanar line pairs with similar plane orientations should be clustered together, as a line can belong to two different planes at edges, e.g., where a wall meets another wall, floor or ceiling.

Yet, we observed on real scenes that such a clustering of coplanar lines into single planes tends to degrade the accuracy of pose estimation. Our interpretation is that individual pairs of (real 3D) lines can be coplanar enough to robustly estimate sensible scale ratios but, when grouped in a single plane, their global coplanarity deteriorates.

In fact, contrary to 3D points, that are well determined although possibly misdetected or mismatched in images, there is no such thing *in the real world* as perfect 3D lines, perfect 3D planes, nor exact line parallelism, orthogonality or coplanarity. (Optical distortion and detection noise just

come on top of it.) This is especially true of edges and surfaces in a building, as tolerances of straightness and flatness in the construction industry range typically from 0.2 to 1%. Line-based calibration is thus prone to be less accurate in practice than point-based calibration, and even less when two lines are involved in a feature, as in line coplanarity.

Buildings also present many cases of near colinearity, hence near line coplanarity, due to close edges at the boundary of thin surfaces, e.g., baseboards, conduits, moldings, picture frames, whiteboards, window and door frames, etc. Furniture edges also tend to be almost but not exactly coplanar with wall edges. Due to small errors, such nearly coplanar lines are considered as inliers, but degrade accuracy.

This consideration is consistent with [33] where the authors observe that, in real data, lines that should "logically" be parallel turn out not to be as much as expected, leading to a number of close but different-enough vanishing points (VPs). They found however that treating parallel line pairs independently leads to more accurate calibrations than merging them into a single VP. Similarly, in our bundle adjustment, we do not consider planes as the support of many coplanar lines; line pairs that were determined as coplanar, i.e., RANSAC inliers, are treated individually.

Residuals and optimization. The parameters of our BA are tracked points and lines, as well as camera positions and orientations. The error to minimize is the sum of the square (pixelic) distance of the reprojected points and lines to their detection (cf. Sect. 6), in all the cameras that see (i.e., detect and match) them, plus the sum of the square coplanarity residuals $d_{co}(L_a, L_b)$ (cf. Sect. 6), for all line pairs L_a, L_b found as inliers, in all images that see both lines.

Concretely, we initialize the BA with the pose found by composing the scaled relative motions (cf. Sect. 3) and with triangulated features. As BA can be sensitive to initialization and given that rotations are often better estimated than translations, we first refine the structure and motion with fixed rotations, then refine all parameters, as [25, 24].

9. Experiments on synthetic data

We made synthetic experiments to study the behavior of our coplanarity-based method when the main scene parameters vary: number of planes, number of detected lines, amount of noise in detections, degree of line planarity. Although the data is synthetic, we used realistic camera and scene parameters, close to the configuration of the real dataset Office-P19 (see below). The setting is detailed in the supplementary material, as well the quantitave results.

We first varied the inaccuracy of line segment detection from $\sigma_{detect} = 0$ to 5 pixels, keeping the planarity noise at $\sigma_{planar} = 0$, then at 20 mm. We also varied the planarity level from $\sigma_{planar} = 0$ to 50 mm, keeping the inaccuracy of line segment detection at $\sigma_{detect} = 0$, then at 2 pixels. In both cases the error of scale ratio estimation degrades

Figure 3. Left: Meeting-P31. Right: Trapezoid-P17 (non Manhattanness can be seen on the ceiling at the room corner).

smoothly. These experiments also show that the more lines a plane contains, the lower the error. The reason is that the error is better averaged and smoothed with more features. Besides, a larger number of planes (and thus a larger number of associated lines) increases the number of constraints but also increases the resulting error, except in the absence of planarity noise. The reason is that non-perfect planes also introduce many outliers.

10. Experiments on real data

Feature detection and matching. Points are detected and matched with SIFT [20]. Lines are detected with MLSD [32] and matched with LBD [45]. Both kinds of features are tracked across consecutive pictures to identify match triplets (hence trifocal overlaps), when any.

Bifocal calibration. We implemented our SfM approach on top of the two-view relative pose estimation of [33]. Besides being parameterless, it has the advantage of robustly combining both line and point features, when any, providing state-of-the-art accuracy even in textureless environments and with wide baselines. When points are not available, it only assumes that both images contains at least two pairs of matched lines that are parallel in 3D; this constraint is most often met in indoor scenes. If not met, our assumption (seeing lines in views 1-2 coplanar with lines seen in views 2-3) is likely not to be met either.

Using method [33] is not intrinsic to our approach. We could have used just as well any other bifocal calibration method, e.g., based on points assuming enough (5 or 7) are available on all image pairs, or based on lines as in [6], although it assumes a Manhattan scene, contrary to [33].

Datasets. We consider both difficult interior scenes and a standard SfM dataset of outdoor scenes with ground truth.

- The indoor dataset pictures office rooms with little texture and little image overlap: Office-P19 (cf. Fig. 1), Meeting-P31 and Trapezoid-P17 (cf. Fig. 3), where Pn means n pictures. Trapezoid-P17 does not belong to a Manhattan world. Resolution is 5184×3456. As can be seen on Fig. 1, points are dense on some textured objects, like the door, but scarce in large other parts of the scene, e.g., white walls.
- Strecha *et al.*'s dataset [37] is a de facto standard for assessing the accuracy of SfM. It consists of 6 outdoor

Method / Scene	Our method	OpenMVG [24]	Olsson [25]	Cui [5]	Arie [1]	Jiang [15]	Bundler [36]	VSfM [42]
Castle P19	**22.2**	25.6	76.2	—	—	—	344	258
Castle P30	23.8	21.9	66.8	**21.2**	—	235	300	522
Entry P10	**5.6**	5.9	6.9	—	—	—	55.1	63.0
Fountain P11	2.4	2.5	**2.2**	2.5	4.8	14	7.0	7.6
Herz-Jesu P8	3.8	**3.5**	3.9	—	—	—	16.4	19.3
Herz-Jesu P25	5.4	5.3	5.7	**5.0**	7.8	64	21.5	22.4

Table 1. Average position error: comparison with global [24, 25, 5, 1, 15] and incremental [36, 42] SfM methods. Best results in **bold**.

Method / Scene	RANSAC threshold (pixels)					AC-RANSAC
	0.5	1	3	6	9	
Castle P19	146	90.5	107	**90**	196	101
Castle P30	91	83	73	76	144	**72**
Entry P10	**7.6**	7.9	8.4	9.3	10.6	7.8
Fountain P11	2.2	**2.1**	2.4	2.5	3.8	2.3
Herz-Jesu P8	4.0	4.1	**3.9**	7.1	6.4	4.2
Herz-Jesu P25	8.5	9.0	8.6	13.3	16.9	**7.9**
Average	43.2	32.8	33.9	33.0	63.0	**32.5**

Table 2. RANSAC with a single fixed threshold vs AC-RANSAC.

Feature / Scene	Points	Lines	Points +lines	Copl.	All	All +BA
Castle P19	97.9	**80**	98	129	101	22.2
Castle P30	80.3	111.2	78	134	**72**	23.8
Entry P10	7.9	12.1	8.0	8.9	**7.8**	5.6
Fountain P11	**2.3**	2.5	**2.3**	52	**2.3**	2.4
Herz-Jesu P8	**4.1**	4.3	**4.1**	5.4	4.2	3.8
Herz-Jesu P25	8.4	15.4	8.4	36	**7.9**	5.4
Average	33.5	37.6	33.1	60.9	**32.5**	10.5

Table 3. Feature comparison: resulting average position error of cameras (in mm, before bundle adjustment). Best results in **bold**. For coplanarity features alone, the residual includes the reprojection error of coplanar lines (which may or may not be trifocal).

scenes with ground truth for camera poses. It is easy to calibrate but useful to show that our robust method does not degrade accuracy in easy settings. We consider both the full dataset as well as subsets of images to reduce image overlap. Resolution is 3072×2048. All images have been corrected for radial distortion.

RANSAC with and without parameters. Table 2 compares RANSAC with a single fixed threshold for all three kinds of features (cf. Sect. 6) to the parameterless AC-RANSAC (cf. Sect. 7). Although AC-RANSAC does not always yield the best results, it works better on average. For a given scene, a better accuracy could be achieved by setting the different thresholds for each kind of features, but it would not be practical, hence the interest of AC-RANSAC. In the following, all experiments rely on AC-RANSAC.

Contribution of the different kinds of features. Tab. 3 reports the accuracy of the different kinds of features, alone or when used jointly. Trifocal line features provide a little less accuracy than trifocal points, which are prevalent in this textured datatset. But used together, they lead to a slightly better accuracy. Coplanarity features alone yield a bit less accuracy than trifocal lines; only Fountain-P11 has a poor calibration with coplanarity, probably because it contains fewer planes than other scenes. Yet, merged with other features, coplanarity does not degrade accuracy in general, and often even improves it. Sect. 8 explains the likely reason for the lesser accuracy of line-based constraints, besides detection issues. Indeed, while a 3D point exists "perfectly" in reality (although located with some inaccuracy), lines are never exactly parallel in real life. Line coplanarity has one more DOF, thus an even higher degree of inexactness. It is

Constraint type / Scene	Point triplet	Line triplet	Copl. line pair
Castle P19	**71** %	0 %	29 %
Castle P30	**86** %	0 %	14 %
Entry P10	**100** %	0 %	0 %
Fountain P11	**100** %	0 %	0 %
Herz-Jesu P8	**100** %	0 %	0 %
Herz-Jesu P25	**87** %	0 %	13 %
Office P19	41 %	12 %	**47** %
Meeting P31	21 %	17 %	**62** %
Trapezoid P17	40 %	0 %	**60** %

Table 4. When selecting a scale ratio, % of best hypotheses originating from each constraint type, across all image triplets of each.

why we excluded clusters of coplanar lines from our BA.

Tab. 4 gives statistics on scale ratio hypotheses that were retained because they were found with the lowest NFA.

State-of-the-art accuracy. Tab. 1 shows that our general approach outperforms incremental point-based SfM methods [36, 42] regarding accuracy, and is on par with state-of-the-art global SfM methods [24, 25, 5, 1, 15]. The wider applicability thus is not traded for accuracy, even in scenarios with dense features and much overlap.

It also shows that our relaxed trifocal constraints retain the most relevant part of trifocality contribution to calibration. Comparing to Tab. 3, we can see that coplanarity constraints alone provide in general comparable or better accuracy than incremental SfM methods [36, 42], except again

Method / Scene	Our method	OpenMVG [24]
Castle P19	0.30	0.21
Castle P30	0.29	0.21
Entry P10	0.006	0.90
Fountain P11	0.26	0.19
Herz-Jesu P8	0.28	0.20
Herz-Jesu P25	0.29	0.20
Office P19	1.07	6/19
Meeting P31	1.79	9/31
Trapezoid P17	3.60	3/17

Table 5. Mean reprojection error of all 3D points of all cameras (in pixels). In red, fraction of calibrated cameras in case of failure. Our residual error in indoor scenes is about 4 to 10 times larger than on Strecha's dataset, but picture resolution is 3 times as large.

on Fountain-P11. Residuals are shown in Tab. 5.

We could not compare with line-based SfM methods [46] as their code is unavailable and as they do not measure their accuracy against standard datasets with ground truth.

Succeeding when others fail. Our method can calibrate difficult datasets that other methods fail to calibrate:

In Office-P19, some triplets do not have overlapping features (neither points nor lines), which makes it impossible to calibrate with existing SfM methods. But our method succeeds. Fig. 1 illustrates some views, with line detections, and point, line and plane reconstructions. Apart from a few line outliers due to mismatches, line and plane reconstructions are qualitatively good. We can also calibrate Meeting-P31 and Trapezoid-P17 while OpenMVG fails; the residuals and number of calibrated cameras are given in Tab. 5.

We also study removing images from existing datasets as it is representative of our scenario where only few pictures are shot (or can be shot, or are good quality), with enough points or line intersections to define bifocal translation directions but not enough trifocal features to scale related bifocal calibrations — a failure case for existing methods.

Removing 3/19 images from Castle-P19 can be enough to cause other methods to fail to calibrate all cameras, thus also reducing reconstructed structures and ability to model the whole scene (cf. missing wall parts in Fig. 4, left). Our method calibrates all cameras, yielding an average location error of 13.2 cm. It is not as good as the state-of-the-art error of 2.3 cm when all 19 images are available (cf. Tab. 1), but still better than Bundler (34.4 cm) and VSfM (25.8 cm). We can actually remove up to 11/19 images and still estimate all camera poses with average error 18.4 cm, which remains better than incremental methods with all 19 images.

Likewise, keeping only 3 images (left-most, middle, right-most) from Herz-Jesu-P8 (Fig. 5) leads other methods to totally fail, while we calibrate these very-wide-baseline images with 6 mm accuracy, the same as full P8.
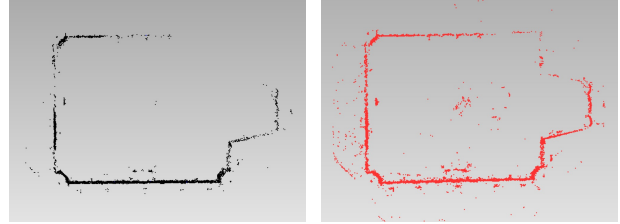

Figure 4. Castle-P(19−3): OpenMVG (left), our method (right).

Figure 5. Only our method can calibrate the very wide baseline and viewpoint change of Herz-Jesu-P3 (furthest 3 images in P8).

Features / Scene	Scale factor estim.				Bundle adjustment			
	Pts	Lin.	Cop.	All	Pts	Lin.	Cop.	All
Castle P19	5.6	15.8	15.8	42.4	2.7	2.4	2.2	5.5
Meeting P31	4.9	4.9	7.0	8.1	0.8	1.2	5.7	8.3

Table 6. Running time (s).

More generally, disconnected components in the trifocal graph make other SfM methods fail to calibrate all cameras, even if the bifocal graph is connected. (Other methods even actually require connections in the trifocal graph supported by at least 3 features.) But provided there are 3D lines configured as in Fig. 2, we can hope to calibrate all cameras.

Scalability. Given a chain of bifocal calibrations sharing a camera, our scale factor estimation method is linear in the number of cameras. Our BA is similar to BA in other SfM methods. Our code [31], in C++, is not particularly optimized, except for NFA computation, for which we precompute binomial coefficients. Running times are in Tab. 6.

More experimental data is in the supplementary material.

11. Conclusion

We have presented a novel SfM method that exploits lines and coplanarity constraints to estimate camera poses in difficult settings with little image overlap and possibly little or no texture, causing other methods to fail.

Our experiments show that our method combines the accuracy of usual trifocal constraints although we relax them, and the robustness (in the sense of wider applicability) of coplanar constraints. It can thus be blindly used to calibrate scenes that other methods fail to calibrate, and still to provide state-of-the-art results on less difficult scenes, with smaller baselines, more overlap or more texture.

Future work include the use of coplanarity constraints for dense surface reconstruction, even in textureless scenes.

References

[1] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *3DIMPVT*, 2012. 2, 7

[2] F. Arrigoni, B. Rossi, and A. Fusiello. On computing the translations norm in the epipolar graph. In *International Conference on 3D Vision (3DV 2015)*, 2015. 1

[3] A. Bartoli and P. Sturm. Constrained Structure and Motion from multiple uncalibrated views of a piecewise planar scene. *International Journal of Computer Vision (IJCV 2003)*, 52(1):45–64, 2003. 3

[4] A. Bartoli and P. F. Sturm. Structure-from-motion using lines: Representation, triangulation, and bundle adjustment. *Computer Vision and Image Understanding*, 100(3):416–441, 2005. 3, 5

[5] Z. Cui, N. Jiang, C. Tang, and P. Tan. Linear global translation estimation with feature tracks. In *British Machine Vision Conference (BMVC 2015)*, 2015. 1, 2, 7

[6] A. Elqursh and A. Elgammal. Line-based relative pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3049–3056, June 2011. 6

[7] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *ICCV Workshops*, pages 264–271, 2011. 3

[8] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a cloudless day. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *11th European Conference on Computer Vision (ECCV 2010)*, pages 368–381. Springer Berlin Heidelberg, 2010. 1

[9] A. Fusiello, F. Crosilla, and F. Malapelle. Procrustean point-line registration and the NPnP problem. In *International Conference on 3D Vision (3DV 2015)*, pages 250–255, Oct. 2015. 2

[10] V. Garro, F. Crosilla, and A. Fusiello. Solving the PnP problem with anisotropic orthogonal Procrustes analysis. In *International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission (3DIMPVT 2012)*, pages 262–269, Oct. 2012. 2

[11] V. M. Govindu. Combining two-view constraints for motion estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, 2001. 2

[12] V. M. Govindu. Robustness in motion averaging. In *ACCV*, 2006. 3

[13] M. Havlena, A. Torii, and T. Pajdla. Efficient Structure from Motion by graph optimization. In *11th European Conference on Computer Vision (ECCV 2010)*, pages 100–113, Berlin, Heidelberg, 2010. 2

[14] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the world* in six days *(as captured by the Yahoo 100 million image dataset). In *Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 2015. 1

[15] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *IEEE International Conference on Computer Vision (ICCV 2013)*, pages 481–488, 2013. 7

[16] N. Jiang, W. Lin, M. N. Do, and J. Lu. Direct structure estimation for 3D reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 2655–2663, 2015. 3

[17] F. Kahl and R. I. Hartley. Multiple-view geometry under the l_∞-norm. *IEEE Trans. PAMI*, 30(9):1603–1617, 2008. 2, 3

[18] C. Kim and R. Manduchi. Planar structures from line correspondences in a Manhattan world. In *12th Asian Conference on Computer Vision (ACCV 2014)*, pages 509–524, 2014. 2

[19] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision (IJCV 2008)*, 81(2):155–166, 2008. 2

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 6

[21] F. M. Mirzaei and S. I. Roumeliotis. Globally optimal pose estimation from line correspondences. In *IEEE International Conference on Robotics and Automation (ICRA 2011)*, pages 5581–5588. IEEE, 2011. 2

[22] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *Int. J. Comput. Vision*, 57(3):201–218, May 2004. 5

[23] P. Moulon, P. Monasse, and R. Marlet. Adaptive Structure from Motion with a contrario model estimation. In *11th Asian Conference on Computer Vision (ACCV 2012) - Volume Part IV*, ACCV'12, pages 257–270, Berlin, Heidelberg, 2013. Springer-Verlag. 2, 5

[24] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3248–3255, 2013. 1, 2, 6, 7, 8

[25] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *SCIA*, pages 524–535, 2011. LNCS 6688. 6, 7

[26] O. Özyeşil and A. Singer. Robust camera location estimation by convex programming. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, June 2015. 1

[27] B. Přibyl, P. Zemčík, and M. Čadik. Camera pose estimation from lines using Plcker coordinates. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *British Machine Vision Conference (BMVC 2015)*, pages 45.1–45.12. BMVA Press, Sept. 2015. 2

[28] S. Ramalingam, S. Bouaziz, and P. F. Sturm. Pose estimation using both points and lines for geo-localization. In *IEEE International Conference on Robotics and Automation (ICRA 2011)*, pages 4716–4723, 2011. 2

[29] A. L. Rodríguez, P. E. López-de Teruel, and A. Ruiz. Reduced epipolar cost for accelerated incremental SfM. In *Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011. 2

[30] C. Rother. Linear multi-view reconstruction of points, lines, planes and cameras using a reference plane. In *9th IEEE International Conference on Computer Vision (ICCV 2003) - Volume 2*, ICCV '03, pages 1210–1217, Washington, DC, USA, 2003. IEEE Computer Society. 3

[31] Y. Salaün. LineSfM, 2017. Code available at https://github.com/ySalaun/LineSfM. 8

[32] Y. Salaün, R. Marlet, and P. Monasse. Multiscale line segment detector for robust and accurate SfM. In *23rd International Conference on Pattern Recognition (ICPR)*, 2016. 6

[33] Y. Salaün, R. Marlet, and P. Monasse. Robust and accurate line- and/or point-based pose estimation without Manhattan assumptions. In *European Conference on Computer Vision (ECCV)*, 2016. 5, 6

[34] G. C. Sharp, S. W. Lee, and D. K. Wehe. Toward multiview registration in frame space. In *IEEE International Conference on Robotics and Automation (ICRA 2001)*, 2001. 3

[35] K. Sim and R. Hartley. Recovering camera motion using l_∞ minimization. In *Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, volume 1, 2006. 2

[36] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transations on Graphics (TOG 2006)*, 25(3):835–846, July 2006. 2, 7, 8

[37] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, June 2008. 6

[38] P. F. Sturm. Algorithms for plane-based pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 1706–1711, 2000. 3

[39] C. Sweeney, T. Sattler, T. Hollerer, M. Turk, and M. Pollefeys. Optimizing the viewing graph for Structure-from-Motion. In *IEEE International Conference on Computer Vision (ICCV 2015)*, pages 801–809, Washington, DC, USA, 2015. 1

[40] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding (CVIU 2015)*, 140(C):127–143, Nov. 2015. 2

[41] K. Wilson and N. Snavely. Robust global translations with 1DSfM. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *13th European Conference on Computer Vision (ECCV 2014)*, pages 61–75, 2014. 1

[42] C. Wu. Towards linear-time incremental structure from motion. In *International Conference on 3D Vision (3DV 2013)*, pages 127–134, 2013. 1, 2, 7, 8

[43] C. Xu, L. Zhang, L. Cheng, and R. Koch. Pose estimation from line correspondences: A complete analysis and a series of solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2016)*, June 2016. 2

[44] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010. 3

[45] L. Zhang and R. Koch. An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency. *J. Vis. Comun. Image Represent.*, 24(7):794–805, Oct. 2013. 6

[46] L. Zhang and R. Koch. Structure and motion from line correspondences: Representation, projection, initialization and sparse bundle adjustment. *Journal of Visual Communication and Image Representation*, 25(5):904–915, 2014. 2, 5, 8

[47] L. Zhang, C. Xu, K.-M. Lee, and R. Koch. Robust and efficient pose estimation from line correspondences. In K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, editors, *11th Asian Conference on Computer Vision (ACCV 2012)*, pages 217–230. Springer, 2012. 2

[48] E. Zheng and C. Wu. Structure from motion using structure-less resection. In *IEEE International Conference on Computer Vision (ICCV 2015)*, pages 2075–2083, Dec 2015. 2

[49] Y. Zheng, Y. Kuang, S. Sugimoto, K. Åström, and M. Okutomi. Revisiting the PnP problem: A fast, general and optimal solution. In *IEEE International Conference on Computer Vision (ICCV 2013)*, pages 2344–2351, Dec. 2013. 2

[50] Z. Zhou, H. Jin, and Y. Ma. Robust plane-based Structure from Motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pages 1482–1489, Washington, DC, USA, 2012. 3