Line-based Robust SfM with Little Image Overlap — Supplementary Material —

Yohann Salaün[1,2], Renaud Marlet[1] and Pascal Monasse[1]
[1] LIGM, UMR 8049, École des Ponts, UPE (France) [2] CentraleSupélec (France)
{yohann.salaun,renaud.marlet,pascal.monasse}@enpc.fr

In this document, we supplement the paper (published in 3DV 2017) with complementary experimental data and technical details.

Section 1 provides statistics regarding the experiments on real data presented in the paper. It also details the reconstruction of planes, as shown on Fig. 1 of the paper, and displays additional reconstructions. Section 2 studies on synthetic data the behavior of coplanar constraints and compares the different kinds of features. Section 3 analyzes the degenerate cases evoked in Sections 4 and 5 of the paper. Section 4 discusses alternatives in our method and provides additionnal justifications for our choices. Finally, we describe in Section 5 a simple line match refinement that we use before bundle adjustment.

1. Experiments on real data

1.1. Statistics on feature detection, matching and calibration inliers

Table 1 below summarizes statistics concerning the detection and matching of features in the various real datasets that we used in our experiments, as well as the inliers we found after estimating the scale ratio in each image triplet.

For each kind of feature (point or line), we display the average number of detections per image, the average number of matched features per image pair, and the average number of features matched across image triplets. Note that, in the case of indoor datasets, there is a huge disparity across the images because tens of thousands of points are detected on the door and on the grids of the suspended ceiling, whereas some other views show textureless parts of the room with practically no point detections.

Each point and line triplet yields a (simplified) trifocal constraint. We also display the average number of constraints corresponding to the hypotheses of coplanar line pairs (per image). Additionally, for each kind of constraint (trifocal point, trifocal line, or coplanar line pair), we display the average number of inliers found after scale ratio estimation, and for readability, the corresponding percentage w.r.t. the average number of candidate inliers.

Finally, for point and line features, we record under "Jumps" the total number of image triplets with absolutely no trifocal feature match. These triplets absolutely need coplanarity constraints to be calibrated.

Scene \ Constraint type	Point triplet	Line triplet	Copl. line pair
Office P19	41 %	12 %	**47** %
Meeting P31	21 %	17 %	**62** %
Trapezoid P17	40 %	0 %	**60** %
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 %

Table 2. When selecting a scale ratio, % of best hypotheses originating from each constraint type, across all image triplets of each.

1.2. Statistics on the type of constraint used for scale ratio estimation

In our method, we use three kinds of constraints (coplanar line pair, trifocal point, trifocal line) and all of them impact the choice of a best scale ratio via their contribution to the NFA (see Section 7 of the paper). To compare their relative influence, we consider the scale ratio hypotheses that were retained because they were found with the lowest NFA, and we measure the percentage of constraint types that these (best) hypotheses originate from, across all image triplets of each dataset. These statistics are presented in Table 2. (Recall that the number of candidate constraints and inliers are shown in Table 1.)

In practice, the total number of features (hypothesized coplanar line pairs, trifocal points, trifocal lines) is low enough (typically less than 10,000) for all features to be tried rather than sampled. The results here are thus deterministic: our algorithm produces exactly the same result each time we run it. Yet, for cases where speed is a concern, we can randomly sample features up to a maximum number of features.

In Strecha *et al.*'s datasets [2], the number of point con-

Information / Scene	Points						Lines						Coplanar lines		
	Detections	Pairs	Triplets	Inliers	Inliers (%)	Jumps	Detections	Pairs	Triplets	Inliers	Inliers (%)	Jumps	Hypotheses	Inliers	Inliers (%)
Office P19	8592	156	44	42	95.0	1	199	20	6	5	80.2	2	165	34	20.6
Meeting P31	14110	290	76	70	91.7	0	335	45	82	14	17	0	403	63	15.7
Trapezoid P17	10420	265	34	31	90	0	216	27	50	5.5	11	0	219	22	10
Castle P19	16586	2440	1123	1078	96.0	0	682	148	41	38	91.5	0	1275	357	28.4
Castle P30	16535	3082	1635	1573	96.0	0	680	167	28	27	94.4	0	1327	383	28.9
Entry P10	17348	4523	2368	2291	96.8	0	1621	319	83	77	92.2	0	2216	892	40.3
Fountain P11	28192	9930	6083	5877	96.6	0	746	149	35	33	92.5	0	1080	252	23.3
Herz-Jesu P8	27073	6258	3289	3157	95.9	0	751	154	52	48	92.9	0	1092	401	36.7
Herz-Jesu P25	26175	4200	1709	1647	96.4	0	771	123	35	32	91.0	0	1059	291	27.5

Table 1. Statistics on the number of feature detection, matching and calibration inliers per image, for different kinds of feature constraints and for each dataset. "Jumps" counts the total number of image triplets with absolutely no trifocal feature match.

straints (# point triplets) is much larger than the number of line constraints (# line triplets). Considering only inliers, the number of point constraints is also much larger than the number of line coplanarity constraints. It explains why, in this textured dataset, so many scale ratios are decided by point constraints.

In the indoor datasets, the statistics is more balanced because the number of inlier constraints is comparable for trifocal points and coplanar line pairs, with a low but non negligible number of inliers for trifocal lines.

From these statistics, we can say that although coplanarity constraints and trifocal line constraints are generally less accurate than trifocal point constraints (see Section 2 above), they are useful to obtain more accurate results, especially in scenes with little texture. They are not "just" useful for robustness in case points are missing.

1.3. Plane clustering

Every inlier line pair of a triplet of cameras is composed of coplanar lines by hypothesis. To get an idea of the planes that are present in the scene, we may combine coplanarity constraints and define global planes.

To this end, we greedily cluster the coplanar pairs, first at the three-view stage, and then at a global stage. Concretely, for each triplet of cameras, we cluster coplanar pairs that share a common line and have a similar normal (we use a threshold of $15°$). Once clustered, we compute an average plane position and center. Then, at a global scale, we perform a similar clustering for planes that have close normal and close centers. The angular threshold is the same as previously, and the distance threshold is defined as a portion of the scene scale (we use 25% of the scene scale).

The planes obtained this way do not always correspond to "real" planes, such as walls, floors, windows, doors, but they correspond at least to virtual planes, e.g., accidental alignment of the door with the whiteboard edge, or aligned step edges of a stairway, as in Fig. 1, where they are represented by similar colors. Yet, these global planes provide information about the scene that could be used to help a dense reconstruction algorithm, especially in low texture scenes, as illustrated in Fig. 1 of the paper and in Figs. 2 and 3 of this supplementary material, where planes are represented by their bounding box (which is not forced here to be vertically aligned).

Figure 1. Cluster of segments belonging to the same 3D plane, represented with the same color. If a segment belongs to several clusters, although it appears in only one color, it is linked to its clusters through thin lines with the cluster color.

1.4. Visual results

An example for reconstructed structure for the Strecha outdoor dataset is given on Fig. 2.

For indoor datasets, which do not have a ground truth, we however provided a quantitative result that corresponds to the reprojection error of points (cf. Table 3 of the paper). Yet, if this error says a lot about the calibration accuracy, visual results are also worth of interest.

For each indoor dataset, we show in Fig. 3 the 3D reconstruction based on points, lines and planes. Planes correspond to the aggregation of coplanar constraints as ex-

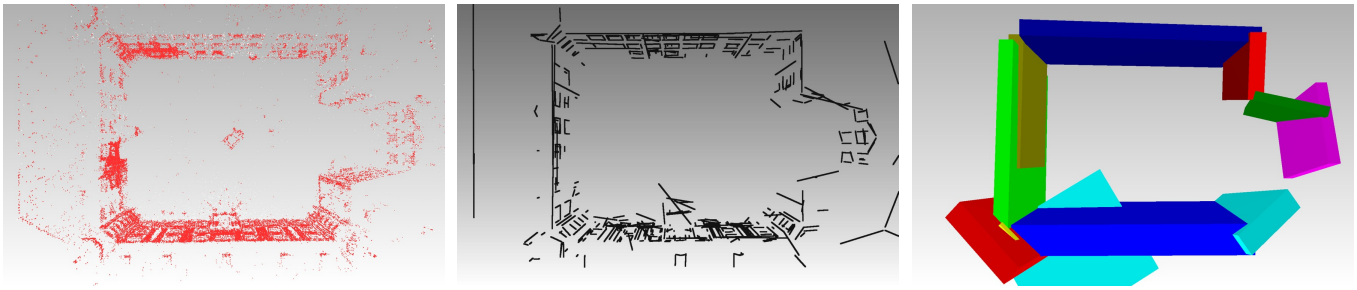

Figure 2. Reconstructed structure of Castle-P30: points (left), lines (middle), bounding boxes of coplanarity planes after BA (right).

plained in previous section.

2. 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 in the scene, number of detected lines, amount of noise in detections, degree of planarity. As shown below, the coplanarity constraints degrade smoothly when the setting becomes harder.

Synthetic setting. Although the dataset is synthetic, we used realistic camera and scene parameters, that are actually close to the configuration of the Office-P19 dataset (see below) and also provide a metric assessment. We consider the following setting:

- 3 cameras on a plane, pointing toward a scene, with camera i at $15° + \text{random}[0, 15°]$ from camera $i+1$,
- scene of size 2 m, at distance of 3 m from the cameras,
- cameras with focal length 20 mm and resolution of 4000×4000 pixels.

In each experiment we generate random planes, and random lines in each plane. Half of the lines are only seen in cameras 1 and 2, and the other half only in cameras 2 and 3, as in Fig. 2 of the paper. We consider 4 scene configurations, that generate a number of constraints that is similar to what we observe in a real dataset such as Office-P19:

- 2 planes with 10 lines in each plane, generating an average of 90 coplanar constraints,
- 2 planes with 20 lines in each plane, generating an average of 200 coplanar constraints,
- 3 planes with 10 lines in each plane, generating an average of 150 coplanar constraints,
- 3 planes with 20 lines in each plane, generating an average of 300 coplanar constraints.

Note that using at least 2 planes and 10 lines in the synthetic experiments is not a minimal requirement; it just corresponds to practical indoor situations with little data.

Note also that the number of constraints that are actually used varies: it depends on the number of line pairs that are discarded because they are below the angular threshold for near parallelism (cf. Sect. 6 of paper).

Last, we use a Gaussian noise to model inaccuracy in line detection as a "reprojection error" with a given standard deviation σ_{detect} (in pixels), perturbing both segment extremities, and also a Gaussian noise to model planarity error, to deviate lines from their plane with a given standard deviation σ_{planar} (expressed in mm).

For each scene configuration and each parameter value, we generate 500 random scenes and average the error measure $\epsilon(\tau) = |\tau - \tau_{GT}|/|\tau_{GT}|$, where τ is the estimated scale ratio, and τ_{gt} is the ground-truth scale ratio.

2.1. Coplanar constraints

We first study the sensitivity of coplanar constraints with respect to reprojection error and planarity error.

Sensitivity to detection inaccuracy. As seen from Fig. 4, when we vary the inaccuracy of line segment detection from $\sigma_{\text{detect}} = 0$ to 5 pixels, keeping the planarity noise at $\sigma_{\text{planar}} = 0$ and 20 mm, the error of scale ratio estimation degrades smoothly.

Sensitivity to the degree of planarity. As seen from Fig. 5, when we vary the planarity level from $\sigma_{\text{planar}} = 0$ to 50 mm, keeping the inaccuracy of line segment detection at $\sigma_{\text{detect}} = 0$ and 2 pixels, the error of scale ratio estimation also degrades smoothly.

Sensitivity to the number of lines. We can also observe from Fig. 4 and 5 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.

Sensitivity to the number of planes. The same figures show that 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.

2.2. Features comparison

We provide here synthetic experiments illustrating the sensitivity to detection inaccuracy for the different kinds of constraints, i.e., trifocal points, trifocal lines, and pairs of lines hypothesized as coplanar.

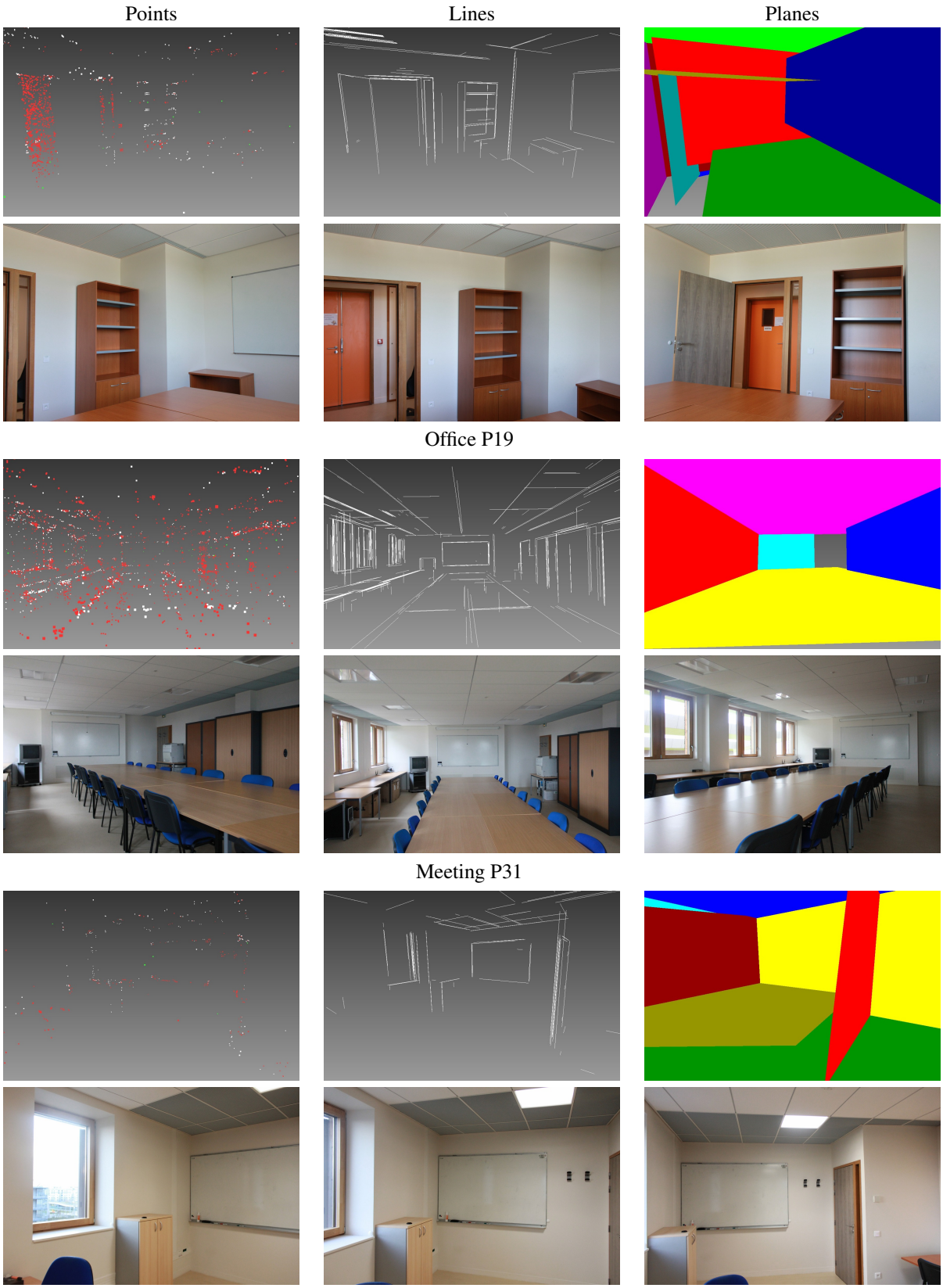

Figure 3. 3D reconstruction of indoor scenes, showing separately points (left), lines (middle) and coplanar features (right). Images of the same area in the dataset are shown in the following row.

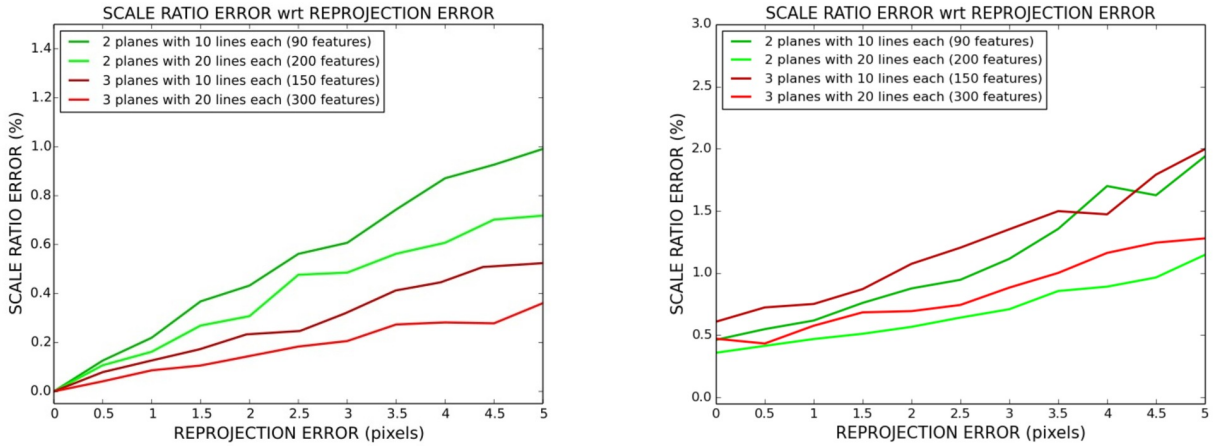

Figure 4. Sensitivity of scale ratio estimation when line segment detection inaccuracy increases. On the left with no planarity noise, on the right with a planarity noise of standard deviation $\sigma_{\text{planar}} = 20$mm.

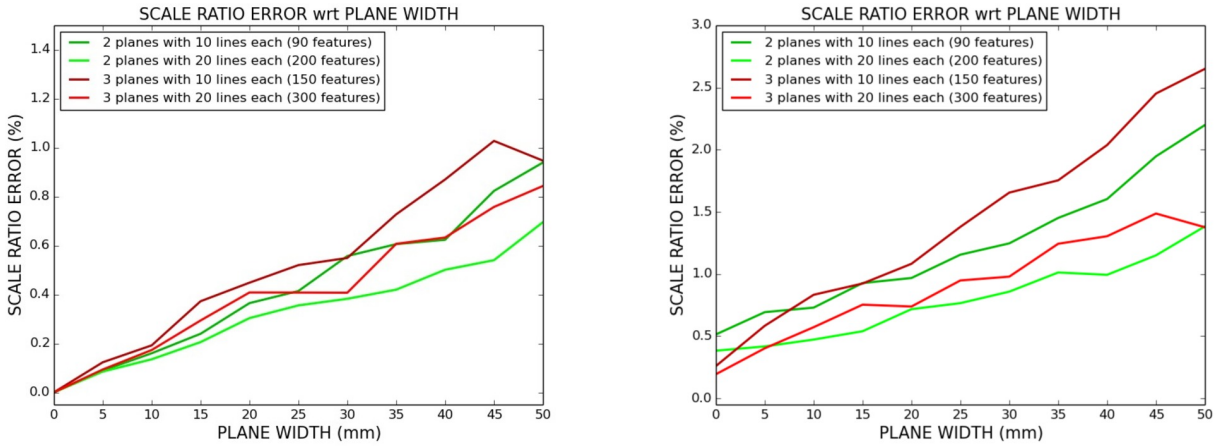

Figure 5. Sensitivity of scale ratio estimation as planarity degrades. On the left with no reprojection noise, on the right with a reprojection noise of standard deviation $\sigma_{\text{detect}} = 2$ pixels.

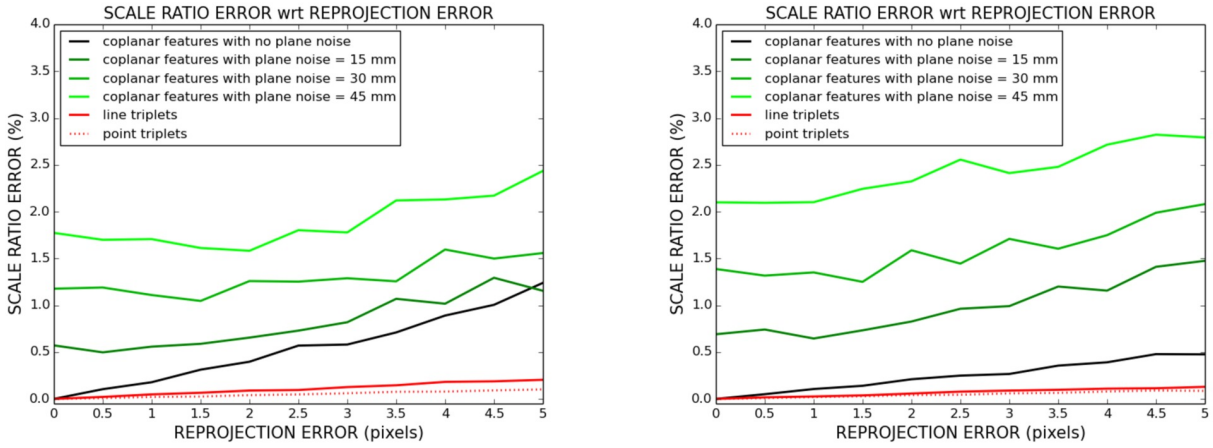

Figure 6. Sensitivity of scale ratio estimation when feature detection inaccuracy increases, for the different kinds of considered constraints: trifocal points, trifocal lines, and pairs of lines hypothesized as coplanar (for different values of coplanarity noise). On the left, configuration with 2 planes and on the right, configuration with 3 planes.

We use a 3-camera synthetic setting, and we consider the following scene configurations:

- 2 planes with 10 points and 10 lines in each plane, seen in all 3 cameras, resulting in $n_{\mathsf{pt}} = 20$ point triplets, $n_{\mathsf{seg}} = 20$ line triplets and an average of $n_{\mathsf{co}} = 200$ coplanar line pairs (considering all line pairs originating from matches in cameras 1-2 and cameras 2-3).
- 3 planes with 10 points and 10 lines in each plane, seen in all 3 cameras, resulting in $n_{\mathsf{pt}} = 30$ point triplets, $n_{\mathsf{seg}} = 30$ line triplets and an average of $n_{\mathsf{co}} = 300$ coplanar line pairs (considering all line pairs originating from matches in cameras 1-2 and cameras 2-3).

Points and lines are perfectly matched. It means that point and line triplets (tracks of length 3) do not contain any outlier. However, for coplanarity hypotheses, we consider all combinations of a line pair in cameras 1-2 with a line pair in cameras 2-3, which contain a large number of outliers (false hypotheses). For 2 planes, about $\frac{1}{2}$ of the hypotheses are outliers, and for 3 planes, about $\frac{2}{3}$ of the hypotheses are outliers.

We measure the scale ratio error as a function of the inaccuracy in feature detection, modeled as a Gaussian noise with a given standard deviation σ_{detect} (in pixels). In case of a line segment, we perturb both extremities in each camera. In case of a point feature, we perturb its reprojection location in each camera. Results are pictured in Fig. 6, where the noise σ_{detect} is labeled as a "reprojection error".

Note that, although we plot them on the same graph, we cannot really compare the resulting inaccuracy of line triplets vs point triplets. The reason is that neither the error model nor the noise levels for representing the inaccuracy of line and point detections are a priori comparable. Only the general behavior of both kinds of trifocal constraints is meaningful here.

Imperfection is also introduced in plane planarity, as a Gaussian noise deviating lines off their plane with a given standard deviation σ_{planar} (expressed in mm, given the setting in Section 2 above). We consider different noise levels, with $\sigma_{\mathsf{planar}} \in \{0\,\mathrm{mm}, 15\,\mathrm{mm}, 30\,\mathrm{mm}, 45\,\mathrm{mm}\}$. As trifocal lines and points are not sensitive to planarity noise, we only plot one curve for them.

3. Degenerate cases

Our constraints (based on coplanar lines or on trifocal features) have singular cases. We detail here these cases and show that a minimum angle between two vectors is enough to discard degenerated configurations.

3.1. Coplanar lines

Coplanar constraints generate an equation of the form:

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

Degenerate cases can then occur when either the denominator or the numerator vanishes, meaning one of the dot products is null. There are 6 different cases than can be split in 2 groups of symmetric cases. For clarity we only study the case that concerns line L_a:

- $l_a^1 \cdot (R_{21} p_a^2) = 0$ implies that the plane $(C_2 l_a^2)$ is parallel to the plane $(C_1 l_a^1)$. It means that line L_a cannot be determined by the observations on cameras 1 and 2.
- $n_{\mathcal{P}} \cdot (R_2^\top p_a^2) = 0$ implies that whatever the scale ratio, L_a is at a constant distance of the plane of normal $n_{\mathcal{P}}$ that contains L_b.
- $l_a^1 \cdot t_{21} = 0$ implies that the line L_a is constant whatever the scale ratio is.

These degenerate cases do not allow the computation of a scale ratio candidate. Moreover, they cannot measure if a given candidate is a good one. Thus, when we compute the coplanar pairs, we select all the possible pairs but ignore the ones that are degenerate. Note that these tests are performed without knowledge of the camera positions, only global rotations and translation directions are required.

3.2. Trifocal features

Trifocal features generate an equation of the form:

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

Degenerate cases occur when the vector family $(u, w, v \times w)$ is not a basis of $\mathbb{R}^3$. These cases correspond to specific geometric configurations that depends on the nature of the features.

For points, the equation becomes :

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

The degenerate cases correspond to linear combinations of the following two cases:

- C_2, C_3 and p_3 are aligned, meaning that whatever λ, the point cannot be triangulated from cameras 2 and 3.
- $C_3 p_3$ is orthogonal to the plane $(C_2, C_3, \tilde{P})$, meaning that whatever λ, it is not possible to intersect $C_3 P_3$ and $C_2 \tilde{P}$ and thus there is no constraint.

Note that these degenerate cases may also occur for cameras 1 and 2.

For lines, the equation (11) becomes:

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

The degenerate cases correspond to linear combinations of the following two cases:

- $l_3 \perp t_{23}$ and $l_3 \perp d_{\tilde{L}}$, which means that the plane $(C_3 l_3)$ is constant whatever λ.
- $l_3 \parallel (d_{\tilde{L}} \times (\tilde{P} - C_2)) \times (d_{\tilde{L}} \times t_{23})$, then the distance is constant whatever λ.

Note that all these degenerate cases correspond to specific geometric configuration that do not occur frequently. Moreover, they can be avoided by testing a triplet before using it to generate a candidate λ or check its validity.

4. Alternative choices

4.1. Coplanar line distance

We measure the coplanarity discrepancy of two lines based on the projection on the central image of their closest points in 3D (see Sect. 6 of the paper).

Considering $\tilde{p}^2_{ab}, \tilde{p}^2_{ba}$ rather than P_{ab}, P_{ba} removes one dimension of error, along the direction of C_2, possibly constraining λ less. But positioning by triangulation along this direction is generally less accurate and thus less meaningful, leading most methods to use reprojected distances (e.g., epipolar distance) and rely on other views to capture any error along this direction.

We tried a number of alternatives to measure coplanarity. The 3D distance is difficult to threshold or compare with as it is computed up to an unknown scale factor. We also tried different weight alternatives, e.g., to take into account uncertainty or based on how the cameras see the distance between the 2 lines, but results were not consistently better. In practice, the plain reprojected distance, which is common in many SfM approaches, works well enough, despite degenerate cases, which we discard.

4.2. Robust estimation

Our method to robustly estimate scale factors is based on an *a contrario* principle.

We tried a number of alternatives. An issue for finding the modes of a histogram was to find a proper bin discretization as the scale of the relative scale factors is itself unknown. It is also difficult to use a hypothesis test as the possible scale factors follow a distribution difficult to characterize, and its standard deviation is highly sensitive to the many outliers.

Our AC-RANSAC formulation, which is a form of hypothesis test, does not assume uniform scale ratios for the null hypothesis; it considers a distribution indirectly defined by a distribution of lines. Besides, AC-RANSAC allows to treat all features (points, lines, coplanar lines) in the same framework.

5. Match refinement

After two-view pose estimation, we actually refine line matches using the known transformation, in the same spirit as [1], but more simply. For each line segment l_1 matching with l_2 in the other image, we consider the strip S delimited by the epipolar lines of the extremities of l_2; we measure the ratio of the length of segment l_1 that overlaps S over the length of the (infinite) line l_1 that overlaps S. Segments below a threshold are discarded, and matches are recomputed as in LBD [3], but now ignoring LBD's geometric criterion, as we have more reliable information.

References

[1] M. Hofer, M. Donoser, and H. Bischof. Semi-global 3D line modeling for incremental Structure-from-Motion. In *British Machine Vision Conference (BMVC 2014)*, 2014. 7

[2] 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. 2

[3] 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. 7