Towards structureless bundle adjustment with two- and three-view structure approximation

Ewelina Rupnik, Marc Pierrot Deseilligny

TOWARDS STRUCTURELESS BUNDLE ADJUSTMENT WITH TWO- AND THREE-VIEW STRUCTURE APPROXIMATION

Ewelina Rupnik [1] and Marc Pierrot Deseilligny[1]

[1] LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mande, France

KEY WORDS: Global SfM, Bundle adjustement, Structure approximation

ABSTRACT:

The global approaches solve SfM problems by independently inferring relative motions, followed be a sequential estimation of global rotations and translations. It is a fast approach but not optimal because it relies only on pairs and triplets of images and it is not a joint optimisation. In this publication we present a methodology that increases the quality of global solutions without the usual computational burden tied to the bundle adjustment. We propose an efficient structure approximation approach that relies on relative motions known upfront. Using the approximated structure, we are capable of refining the initial poses at very low computational cost. Compared to different benchmark datasets and software solutions, our approach improves the processing times while maintaining good accuracy.

1. INTRODUCTION

Photogrammetry and computer vision Structure from Motion (SfM) algorithms underwent a remarkable evolution during the past two decades. Much of its evolution was driven by the advances in sensor technology, growing computer capabilities and democratisation of the fields through different software solutions [Snavely et al., 2006], [Pierrot Deseilligny, Cléry, 2011], [Moulon et al., 2016], [Schonberger, Frahm, 2016a].

Existing SfM approaches for pose estimation become computationally inefficient as the number of images increases. Two solutions exist: sequential and global methods. Sequential methods [Schonberger, Frahm, 2016a, Snavely et al., 2006] generate one or multiple (e.g. hierarchical methods [Klopschitz et al., 2010, Toldo et al., 2015]) seed images and sequentially concatenate overlapping images. Non-linear least squares bundle adjustment is then used to systematically eliminate accumulated errors [Triggs et al., 1999]. On the other hand, global methods [Govindu, 2001, Moulon et al., 2013, Wilson, Snavely, 2014] simultaneously and independently find relationships between pairs or triplets of images. These relationships are encoded in an epipolar graph which forms the basis of inferring the global rotations and translations. Global methods require the knowledge of image correspondences only in the first phase when computing the two-view or three-view geometries. This makes them faster and more computationally efficient. However, identifying outliers becomes challenging because the computation is bound to a pair or a triplet.

Global solutions to SfM problems provide only initialisation of camera poses. Ordinarily, a final bundle adjustment (BA) including poses and point correspondences is run to calculate the optimal solution [Triggs et al., 1999], which is a costly processing. The objective of this work is to go towards truly structureless bundle adjustment by refining the initial poses with very few points. We propose a structure (i.e. 3D points) approximation algorithm that relies on pairs and triplets of images. We firstly introduce a robust method for the relative motion calculation. Then, we redefine the per-relative-motion structure through the intermediary of an ellipsoid. The goal here is to generate a minimal set of 3D points that are sufficient to infer the local structure. This set is later used to refine the global solution in a bundle adjustment routine. Please note that we do not present a global motion approach in this work. Instead, to produce an initialization to the BA, we use the global approach of SfmInit [Wilson, Snavely, 2014] or MicMac [Pierrot Deseilligny, Cléry, 2011, Rupnik et al., 2017]. We evaluate our approach on the Strecha benchmark [Strecha et al., 2008], as well as sequential acquisitions by a mobile mapping platform and a drone. Compared to other approaches [Wilson, Snavely, 2014], [Pierrot-Deseilligny et al., 2016], [Schonberger, Frahm, 2016b] we obtain accurate results in less time. The running time reduction factor is proportional to the ratio of the input point correspondences to the fictitious observations. All methods are implemented and available from: $github.com/\ xxx$. The sequential datasets and their respective ground truth are also made available.

2. RELATED WORK

On global motion methods Determination of global motions is performed separately for the rotational and translation parts [Govindu, 2001]. To calculate the rotations, the existing solutions: ignore the orthogonality constraint and employ linear methods [Martinec, Pajdla, 2007, Arie-Nachimson et al., 2012, Moulon et al., 2013]; or optimise with robust ℓ_1 as Hartley *et al.* [Hartley et al., 2011] or as Chatterjee *et al.* [Chatterjee, Govindu, 2013] who base their algorithm on the Lie Group. A review of rotation averaging algorithms was published by Hartley *et al.* [Hartley et al., 2013]. The translation component is always computed in the second place. Unlike for the rotational component, there is no direct way to compute scale-consistent translations from a set of available pairwise constraints. Some methods exploit only pairwise constraints [Govindu, 2001], [Martinec, Pajdla, 2007], [Arie-Nachimson et al., 2012], [Wilson, Snavely, 2014] but are susceptible to a degenerate solution when all cameras are aligned. Other methods formulate the problem using trifocal dependencies [Courchay et al., 2010, Zach et al., 2010, Jiang et al., 2013, Moulon et al., 2013]. Moreover, approaches based on linear solvers [Arie-Nachimson et al., 2012], [Govindu, 2001] and ℓ_∞ norm optimisation [Sim, Hartley, 2006, Kahl, Hartley, 2008, Moulon et al., 2013] can be distinguished.

Methods formulating the BA estimation problem in terms of epipolar constraints [Steffen et al., 2010], [Rodriguez et al., 2011], [Indelman et al., 2012], [Cefalu et al., 2016] are not discussed herewith.

On outliers in relative motions Relative motions are the main ingredient in global pose estimation methods. This optimization scheme is particularly sensitive to outliers due to a limited number of motion observations. Since the motions are derived from automated algorithms for sparse correspondence search (e.g. SIFT [Lowe, 2004]), they inherently contain outliers, especially across ambiguous scenes with repetitive patterns, low texture, moving objects, etc. Existing approaches cope with the outliers by either eliminating them prior to optimising for global rotations and translations (i.e. filtering), or by adopting robust optimisation techniques. Examples of the latter are robust ℓ_1 optimization used for rotation averaging [Hartley et al., 2011, Chatterjee, Govindu, 2013] and translations solving [Dalalyan, Keriven, 2009, Olsson et al., 2010, Zhu et al., 2017]. Some recent works have also addressed the evident complementarity of sequential and global methods. In Zhou *et al.* [Zhu et al., 2017, Zhu et al., 2018] many local incremental SfMs help to discard mismatches and invalid motions.
Among the solutions based on filters, the pioneering works [Govindu, 2006], [Sim, Hartley, 2006], [Martinec, Pajdla, 2007] focus on pairwise constraints and remove the outliers by: RANSAC epipolar sampling [Govindu, 2006]; using iterative removal techniques [Sim, Hartley, 2006]; or identifying false matches via partial reconstructions [Martinec, Pajdla, 2007]. Later works exploit the redundancy in the epipolar graph by looking at triplets of images. Courchay *et al.* [Courchay et al., 2010] introduce a trifocal graph parametrization and estimate robust global poses using linear programming with imposed loop constraints and RANSAC. Similarly to this work, Zach *et al.* [Zach et al., 2010] uses loopy belief propagation to identify erroneous geometries in loops. Jiang *et al.* [Jiang et al., 2013] establishes a trifocal coplanarity constraint that minimizes a geometric error. Work in Moulon *et al.* [Moulon et al., 2013] fuses the Bayesian inference of Zach *et al.* [Zach et al., 2010] and weighted graph filtering of Enqvist *et al.* [Enqvist et al., 2011] to remove outliers. Sweeney *et al.* [Sweeney et al., 2015] updates the fundamental matrices by optimizing a cost function defined over the trifocal point transfer. Finally, Wilson *et al.* [Wilson, Snavely, 2014] finds a purely pairwise method adapted to very large image collection where pairwise translations are projected to 1D subspace and a global ordering fitting the pairwise constraints is sought.

CONTRIBUTION AND OVERVIEW

To increase the robustness of the relative motions, our method combines several hypotheses, from most to least conservative, and tests each with a suitable direct algorithm (Section 3.1). To infer the triplets we take a similar perspective as [Jiang et al., 2013] and [Sweeney et al., 2015]. We present a linear algorithm that rapidly calculates the triplet motions (*Quick triplet algorithm*). The robust triplet algorithm (Section 3.2) tests two hypotheses to see whether the camera in question has a short, normal, or a long focal length. As the long focal lengths approach the orthogonal projection, we replace the perspective camera model with the one proposed by Tomasi *et al.* [Tomasi, Kanade, 1992]. Otherwise, the pose of the "third" image of the triplet is found with the spatial resection [Grunert, 1841]. We intentionally avoided the trifocal tensor parametrization knowing its unstable behavior on flat surfaces [Julià, Monasse, 2017]. Finally, our structure reduction approach (Section 4) is accurate, thus effectively propagates the information from original points. It also accelerates the BA without having to resort to PCG-based optimizations [Agarwal et al., 2010, Jeong et al., 2011, Kushal, Agarwal, 2012]. Therefore, we keep the access to the reduced camera system's covariance matrix. This is advantageous for many metrological applications that necessitate a degree of confidence associated with their measurements.

3. BUILDING THE RELATIVE MOTIONS

The computation of relative motion consistent within triplets of images is divided into three stages. Firstly, given a set of image correspondences, we create the epipolar hypergraph $\mathcal{H}(\mathcal{V}, \mathcal{E})$, where each vertex $\mathcal{V}_i \in \mathcal{V}$ represents an image, and the edge $\mathcal{E}_{ij}$ connects a pair of images $(\mathcal{V}_i, \mathcal{V}_j)$ (Section 3.1). The connection implies that relative motion $\{\mathcal{R}_{ji}, \mathbf{tr}_{ji}\}$ is known. We refer to $\mathcal{H}(\mathcal{V}, \mathcal{E})$ as a hypergraph because we have access to a list of connected vertices $\mathcal{V}_k$ for a given $\mathcal{E}_{ij}$. The connected vertices are then explored to construct triplets complying with the respective edges (Section 3.2). We assume that cameras are calibrated, i.e. at least their focal lengths and the principal points are known. The local coordinate frame within an edge $\mathcal{E}_{ij}$ or a triplet $(\mathcal{V}_i,\mathcal{V}_j,\mathcal{V}_k)$ is always associated with the first camera (i.e. $\mathcal{V}_i$). In the text we often refer to the "third" image of a triplet by which we mean the image associated with $\mathcal{V}_k$.

3.1 Pairwise relative motions

Several direct algorithms are tested to compute robust relative motions between pairs of images. We begin with the hypothesis that the corresponding points contain a many outliers, and slowly relax the constraint by allowing more points in the computation. The full set of correspondences entering the method amounts to 500 and was chosen randomly. In total we test 6 calculation variants, see Table 1. The difference between variant T1 and T2 is in the way the points a drawn out the set of 500 points. While T2 is purely random, in T1 we bias the selection by forcing that the points are well distributed across the image plane. This way we avoid sampling points that are clustered in one zone. Variants T1-T4 are solved with 8-point algorithm [Hartley, Zisserman, 2003] throughout several RANSAC samples. After each draw the current solution is refined with a ℓ_2 solver. In T5 we employ the full set of points, and perform ℓ_1 and ℓ_2 estimations. The T6 variant tests the planarity of the scene by computing a homography and decomposing it to a relative motion [Faugeras, Lustman, 1988]. Finally, the variant with the smallest re-projection error is refined in a coplanarity-based bundle adjustment.

Variant	N_{Pts}	N_{RANSAC}	ℓ_1 / ℓ_2	Model
T1	8*	200	ℓ_2	M_{Ess}
T2	8	50	ℓ_2	M_{Ess}
T3	12	100	ℓ_2	M_{Ess}
T4	250	20	ℓ_2	M_{Ess}
T5	500	0	ℓ_1 & ℓ_2	M_{Ess}
T6	150	20	ℓ_2	H

Table 1. Six variants to calculate the pairwise relative motions. M_{Ess} is the 8-point essential matrix algorithm, H is the homography decomposed to a relative motion [Faugeras, Lustman, 1988]. (*) indicates that the points were drawn with a bias forcing a homogenous distribution within the image.

3.2 Triplet motions

The triplet motions are determined ideally between "relevant" triplets of images. In order to select a subset of suitable triplets, we first determine the triplet motions between all possible candidates (see *Quick triplet algorithm*), and use them to define a per-triplet quality index (see *Triplet selection*). The final triplet motions, calculated on a pre-selected set, result from the *Robust triplet algorithm*.

On straight lines intersection For the sake of clarity, we lay out the line intersection algorithm that is the building block of the *Quick triplet algorithm*. Let us define two straight lines as (cf. Figure 1 (a)): $P = P_0 + p(P_1 - P_0)$, $Q = Q_0 + q(Q_1 - Q_0)$. The intersecting 3D point belongs to the line that minimizes the distance between the lines P and Q, hence, it will be found on a line perpendicular to them. We define it as $(P - Q)$, and impose the following: $(P-Q)\cdot(P_1-P_0) = 0$, $(P-Q)\cdot(Q_1-Q_0) = 0$. Using the above, parameters p and q are inferred. The final 3D point is equal to $\frac{1}{2}(P_{Int_p} + P_{Int_q})$, where $P_{Int_p} = P_0 + (P_1 - P_0)\cdot p$, $P_{Int_q} = Q_0 + (Q_1 - Q_0)\cdot q$.

Quick triplet algorithm The rotation of the "third" image in the triplet $(\mathcal{V}_i, \mathcal{V}_j, \mathcal{V}_k)$ is calculated twice: directly as $\mathcal{R}'_3 = \mathcal{R}_{31}$, and indirectly as $\mathcal{R}''_3 = \mathcal{R}_{21}\mathcal{R}_{32}$. The definite rotation being $\mathcal{R}_3 = \frac{1}{2}(\mathcal{R}'_3 + \mathcal{R}''_3)$. The orthogonality of $\mathcal{R}_3$ is enforced by mapping it to its nearest orthogonal rotation with Singular Value Decomposition.
The computation of the perspective center proceeds in two steps as presented in Figure 1. Our objective is to force the perspective center to lie close to both the translation direction $\mathbf{tr}_{31}$ calculated in the preceding step, and the directions associated with the position of 3D points. To achieve this, our first step is to determine the 3D position of two corresponding points using image measurements in $(\mathcal{V}_i, \mathcal{V}_j)$ (cf. Figure 1 (b)). Then, the new perspective center of $\mathcal{V}_k$ results from intersecting the translation direction $\mathbf{tr_{31}}$ and the direction $\mathbf{tr}_{3P_{Int}}$, defined by the image measurement in $\mathcal{V}_k$ and attached at the position of the 3D point. For N corresponding points, one obtains N estimates of the perspective center (or N estimates of p and q, see Figure 1 (a)). The terminal value is then $C_3 = \mathbf{tr}_{31} \cdot \hat{p}$, where $\mathbf{p} = \{p_i\}$, $\hat{p}$ is the median of $\mathbf{p}$ and $i \in [1, N]$. By definition, the proposed algorithm calculates consistent triplets exclusively from image correspondences of manifold 3 (i.e., points visible in 3 images). Statistically speaking, higher manifold points are less susceptible to outliers, therefore they are expected to produce more stable results. It is also required that a minimum of 8 3-manifold points exists. In practice, many more points are available and we decimate them by a factor of 50 when the number exceeds 500. Note that in this formulation, the computation of the perspective center is not degenerate if all three perspective centers are located on a straight line.

Triplet selection Out of a number of initialized triplets, we want to select the K best ones. The notion of a best triplet is translated to a quality index Q and calculated for each triplet in $\mathcal{H}$. The Q privileges wide baselines between images (i.e. base-to-height ratios or bh) as presented here:

$$R = \frac{TQ \cdot bh}{bh + TL}, \tag{1}$$

$$Q = min(R_{V_iV_k}, R_{V_jV_k}), \tag{2}$$

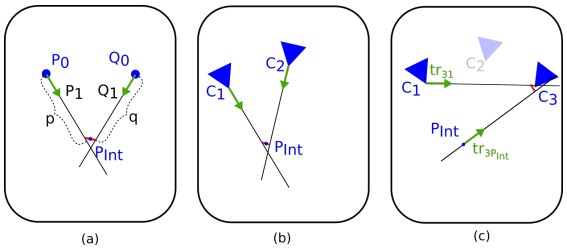

Figure 1. Computation of the perspective center of the "third" image (C_3) in the *Quick triplet algorithm*. C_1 and C_2 are perspective centers of images 1 and 2, respectively. Relative orientations between images 1 and 2 as well as 1 and 3 are known and related by a 7-parameter transformation. (a) A toy example illustrating the intersection of two straight lines defined by vectors $\overrightarrow{P_0P_1}$ and $\overrightarrow{Q_0Q_1}$ (cf. *On straight lines intersection*); (b) intersection of two lines defined by their image observations and respective perspective centers, C_1 and C_2; (c) the sought C_3 lies at the intersection of the vectors $\mathbf{tr}_{31}$ and $\mathbf{tr}_{3P_{Int}}$.

where TQ is the quantification factor and TL is a rough bh limit; $R_{V_iV_k}$ is computed between the first and the second image in a triplet, and $R_{V_jV_k}$ between the second and the third, correspondingly. To rapidly calculate the indices we exploit the triplet motions obtained with *Quick triplet algorithm*. We performed all the experiments with $K = 1$ as with the growing K we didn't get much improvement in accuracy but worsened the running times. The TL and TQ were set to 0.15 and 100.

Robust triplet algorithm At this stage every edge $\mathcal{E}_{ij}$ in $\mathcal{H}$ contains a list of at most K "third" images that were selected in the preceding step. We begin by re-estimating the poses of the "third" images for each $\mathcal{E}_{ij}$ in $\mathcal{H}$. This estimation is embedded in a RANSAC framework, and uses two direct estimation models: the spatial resection algorithm [Grunert, 1841], and the Tomasi *et al.*approach to orientating images with long focal lengths [Tomasi, Kanade, 1992]. We always test both algorithms and choose the better result by comparing their reprojection errors. We set the number of RANSAC draws to 100, and take a random subset of 500 and 30 corresponding points for spatial resection and long focal length algorithms, respectively. Finally, a per-triplet bundle adjustment is run on the randomly reduced points.

4. STRUCTURE APPROXIMATION

Figure 2 illustrates a toy example of the structure approximation algorithm. Given a set of initial 3D points resulting from per-pair or per-triplet intersection (see *On straight lines intersection*), we calculate an ellipsoid inscribed in the points. The ellipsoid is represented by its eigenvectors, eigenvalues as well as its center of gravity. These parameters serve to generate a set of new, fictitious 3D points and their corresponding image measurements. In the same vein, Mayer [Mayer, 2014] suggested a point reduction approach to speed-up the merging process in hierarchical SfM. The ellipsoids are established for both triplets and pairs of images. To begin with, the center of gravity $\mu_{\mathbf{P}}$ and the covariance matrix $K_{\mathbf{P}}$ are determined:

$$\mu_{\mathbf{P}} = \frac{1}{\sum_{i=1}^{N} Pds_{P_i}} \cdot \sum_{i=1}^{N} P_i \cdot Pds_{P_i}, \tag{3}$$

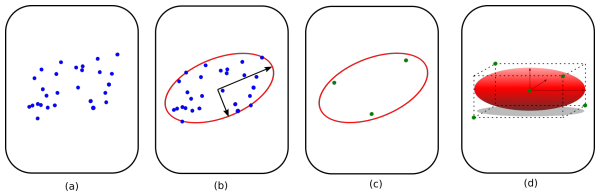

Figure 2. A toy example of structure approximation in 2D and 3D, where (a) shows the initial 2D structure, (b) shows the fitted ellipse, (c) the fictitious 2D structure, and (d) is the equivalent in 3D.

$$K_{\mathbf{P}} = \frac{1}{\sum_{i=1}^{N} Pds_{P_i}} \cdot \sum_{i=1}^{N} Pds_{P_i}^2 \cdot P_i P_i^T - \mu_{\mathbf{P}} \cdot \mu_{\mathbf{P}}^T, \quad (4)$$

where $\mathbf{P} = \{P_i\}$ are the point correspondences, Pds is the weight function that penalizes large reprojection errors (er) and small base-to-height ratios (bh), $Pds = 1/\left[1 + \frac{er}{\alpha \cdot bh}^2\right]$. In our experiments α was fixed to 10. Then, the eigenvalues and eigenvectors of the K matrix are retrieved with the Jacobi method [Press et al., 1992].

We choose to approximate the structure with 5 symmetrically distributed points (see Figure 2 (d)) because it is the minimal number of points sufficient for direct pose estimation algorithms such as the 5-point algorithm (essential matrix), spatial resection as well as for a similarity transform that brings two stereo reconstructions to a common coordinate frame (e.g. in hierarchical SfM).

Given the eigenvalues $\mathbf{e}$ and eigenvectors $\mathbf{E}_i$, we generate a new point H_i from:

$$H_i = \mu + e_1 \mathbf{E_1} \cdot h_{i_1} + e_2 \mathbf{E_2} \cdot h_{i_2} + e_3 \mathbf{E_3} \cdot h_{i_3} \quad (5)$$

where μ is the ellipsoid's center of gravity; $\mathbf{E_1}$, $\mathbf{E_2}$, $\mathbf{E_3}$ are first, second and third unit eigenvectors and e_1, e_2, e_3 are their corresponding eigenvalues; $[h_{i_1}, h_{i_2}, h_{i_3}]$ is the position of a fictitious point in the coordinate frame of the eigenvectors. In our 5-point configuration from Figure 2 (d), the following positions were used: $h_1 = [-1, 1, 1]$, $h_2 = [1, -1, 1]$, $h_3 = [1, 1, -1]$, $h_4 = [0, 0, 0]$, $h_5 = [-1, -1, -1]$. After the first generation of the new structure, we verify that its eigenvectors and eigenvalues were preserved. If they deviate from their initial values, we recalculate the new structure using an appropriate correction factor. Inspired by Mayer [Mayer, 2014], we additionally extended our experiments to generating 5 randomly distributed points. The reported results are always a mean value over 100 or 1000 repeated random calculations. In the bundle adjustment, the reduced points are weighted according to the number of points that contributed to calculating the ellipsoid:

$$Pds = 1 - \frac{Nb_{Max}}{Nb + Nb_{Max}}, \quad (6)$$

where Nb are the initial points, and Nb_{Max} was set to 100.

5. RESULTS

We run experiments on three groups of datasets: the Strecha benchmark [Strecha et al., 2008], a sequential dataset (CAR) acquired by a mobile mapping system, and a drone dataset (UAV). The number of images, number of corresponding points and number of triplet pairs per dataset is shown in Tables 2 and 3. In Tables 4, 6 and Figure 3 the results are evaluated with respect to ground truth data, and with respect to other software solutions. To bring the results to the same coordinate frame, we perform a 7-parameter transformation on the cameras' perspective centers. The *Average positional error* refers to respective perspective centers' differences once the transformation has been applied.

Dataset	N_{images}	N_{tri}
F-P11	11	52
HJ-P8	8	18
HJ-P25	25	228
C-P10	10	44
C-P19	19	68
C-P30	30	214
CAR	647	4167
UAV	73	440

Table 2. Number of images (N_{images}) and triplets (N_{tri}) in respective datasets.

	Image correspondences				Ratio
Dataset	Init.	E_{p+tri}	E_{tri}	E_p	$\frac{Init}{p+tri}$
F-P11	31929	1796	1432	364	18
HJ-P8	17084	736	540	196	23
HJ-P25	96155	8158	6638	1520	12
C-P10	25704	1628	1296	332	16
C-P19	40008	2396	1850	546	16
C-P30	148695	7726	6212	1514	19
CAR	2175980	157878	128688	-	13
UAV	170824	16388	13046	3342	10

Table 3. Number of extracted features in respective datasets. *Init* are the features extracted with SIFT [Lowe, 2004], E_p, E_{p+tri} and E_{tri} are respectively the fictitious image points derived from structure approximated on pairs, pairs and triplets, as well as triplets only.

Strecha images are provided with their ground truth poses. The CAR dataset contains 647 images that amount to an approximately 2km trajectory without loops (see Figure 4). The camera ground truth poses were calculated using highly accurate ground control points measured in the field with classical surveying techniques (point positional accuracy in the range of few mm). To avoid having to measure many ground control points across the entire trajectory length, the vehicle moved in circles around a block of buildings, and during processing the image correspondences were extracted only between immediate cameras. The SfmInit [Wilson, Snavely, 2014] solution did not suceed in orientating the CAR image set, therefore, we needed to resort to an alternative global SfM available in MicMac [Pierrot Deseilligny, Cléry, 2011]. A possible explanation is that the CAR dataset is a linear acquisition, with potentially many cameras located on the same 3D line. Such acquisitions form a degenerate case for algorithms based only on pairwise constraints. For the same reason the BA refinement on structure approximated with pairs of images (E_p) did not converge.

The UAV dataset contains 73 images over an area of 150x160m as illustrated in Figure 5. Analogously to the CAR dataset, the ground truth poses were calculated using ground control points measured in the field. Here as well, SfmInit did not manage to provide an initialisation to the final BA, hence, we resorted to MicMac [Pierrot Deseilligny, Cléry, 2011].

The followed general pipeline includes:

1. Extraction of **correspondences** with SIFT [Lowe, 2004];
2. Building of **relative motions** (Section 3);

3. Approximating the **structure** (Section 4), variants based on:
   (a) pairs, E_p;
   (b) triplets, E_{tri};
   (c) pairs and triplets, E_{p+tri};
   (d) pairs and triplets with random point distribution, E^{rnd}_{p+tri};
4. Computation of **global motions** with SfmInit [Wilson, Snavely, 2014] or MicMac (MM_{Init}) [Pierrot Deseilligny, Cléry, 2011];
5. **Bundle adjustment** on camera poses and image correspondences [Pierrot Deseilligny, Cléry, 2011], with the same number of iterations;
6. Comparison with the **ground truth** data (Table 4-5, Figure 3).

Running times Tables 5, 6 and Figure 3 report on the running times. All our experiments were carried out on a machine with Intel Core i7-8550U, 1.8GHz x 8 processor. The *Ours* approaches concern time spent on the final BA. To deduce the integral processing time, one should add it to the time spent on calculating the initial solution. For a just evaluation, in Table 5, we look at the ratio $\frac{SfmIBA}{Ours}$, which compares the BA processing times with SIFT to our approximated structure approach. We are faster in all instances by a few factors.

Structure approximation variants Approximating the structure with triplets only (E_{tri}), or triplets and pairs (E_{p+tri}) of images turns out to be the best solution throughout all processed acquisitions. We can assume that for well connected images, which is the case for the three acquisitions tested, using triplets only is accurate enough and fast. For sparser connections, adding pairs may be indispensable. We have also observed a mediocre performance when the structure was approximated only with the pairs (E_p). In general, the refinement step succeeds in improving accuracy with respect to initial poses, and the variants with structure approximated by triplets of images perform at least as good as the sequential SfMs solution.

Deterministic or random fictitious points The results of the three experiments using random points do not stand out from the 5-point configuration. On the Strecha benchmark as well as the UAV dataset, the random distribution is only slightly worse than the best solutions. On the other hand, in the CAR dataset, the random selection sometimes outperforms the other three variants. We perceive that what discriminates the acquisitions is the scene geometry. In the CAR dataset, the scene is characterised by very large depth variations, which allows it to fully model the local structure even of randomly sampled.

Limitations The structure approximation approach is not meant for image sets with unstructured connectivity graphs (e.g. Internet photo collections). The overwhelming redundancy of images does not reduce the information content across the scene, hence, does not accelerate the processing.

6. CONCLUSION

There are two leading contributions in this publication. First, we introduce a combinatorial approach to estimating relative motions which enhances the robustness to identified degenerate cases. Second, for a given relative motion, we propose a way of abstracting its 3D structure, such that the motion's estimation properties are preserved. The abstracted structure injected into a bundle adjustment effectively refines the initial solution at very low computational cost. The concept is validated with respect to ground truth data as well as other software solutions. Future work will concentrate on the purely structureless bundle adjustment, with exclusively view-dependent constraints and a thorough propagation of the minimal structure via relative and absolute motion covariance matrices.

REFERENCES

Agarwal, S., Snavely, N., Seitz, S. M., Szeliski, R., 2010. Bundle adjustment in the large. *European conference on computer vision*, Springer, 29–42.

Arie-Nachimson, M., Kovalsky, S. Z., Kemelmacher-Shlizerman, I., Singer, A., Basri, R., 2012. Global motion estimation from point matches. *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, IEEE, 81–88.

Cefalu, A., Haala, N., Fritsch, D., 2016. STRUCTURELESS BUNDLE ADJUSTMENT WITH SELF-CALIBRATION USING ACCUMULATED CONSTRAINTS. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 3(3).

Chatterjee, A., Govindu, V. M., 2013. Efficient and robust large-scale rotation averaging. *Proceedings of the IEEE International Conference on Computer Vision*, 521–528.

Courchay, J., Dalalyan, A., Keriven, R., Sturm, P., 2010. Exploiting loops in the graph of trifocal tensors for calibrating a network of cameras. *European Conference on Computer Vision*, Springer, 85–99.

Dalalyan, A., Keriven, R., 2009. l_1-penalized robust estimation for a class of inverse problems arising in multiview geometry. *Advances in Neural Information Processing Systems*, 441–449.

Enqvist, O., Kahl, F., Olsson, C., 2011. Non-sequential structure from motion. *2011 IEEE International Conference on Computer Vision Workshops*, IEEE, 264–271.

Faugeras, O. D., Lustman, F., 1988. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(03), 485–508.

Govindu, V. M., 2001. Combining two-view constraints for motion estimation. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2, IEEE, II–II.

Govindu, V. M., 2006. Robustness in motion averaging. *Asian Conference on Computer Vision*, Springer, 457–466.

Grunert, J., 1841. Das pothenotsche problem, in erweiterter gestalt, nebst bemerkungen über seine anwendung in der Geodäsie. *Archiv der Mathematik und Physik*, 238–248.

Hartley, R., Aftab, K., Trumpf, J., 2011. L1 rotation averaging using the weiszfeld algorithm. *CVPR 2011*, IEEE, 3041–3048.

Dataset	Accuracy err [mm] SfmI	Colmap	MM	SfmI After BA	Ours E^{rnd}_{p+tri}	Ours E_p	Ours E_{tri}	Ours E_{p+tri}
F-P11	163.8	1.4	1.8	1.3	1.3	33.6	1.0	1.0
HJ-P8	70.5	2.8	5.2	2.7	2.8	2.0	2.6	2.5
HJ-P25	95.3	3.9	5.8	4.3	3.8	4.8	3.5	3.4
C-P10	105.5	5.2	3.3	5.5	5.7	44.4	5.9	5.5
C-P19	1630.8	31.3	16.3	20.0	55.0	350.1	21.3	56.2
C-P30	1346.3	14.7	11.2	13.3	16.7	263.5	12.7	12.6

Table 4. Strecha benchmark. Average position error (err) w.r.t ground truth for different sequential [Pierrot Deseilligny, Cléry, 2011], [Schonberger, Frahm, 2016b] SfMs, as well as our approach (Ours). SfmI is the initial solution found with [Wilson, Snavely, 2014] that we relied upon in our BA. The E^{rnd}_{p+tri} corresponds the solution with structure approximated by pairs an triplets with randomly distributed points; a mean value over 1000 repeated calculations is reported; E_p, E_{tri}, E_{p+tri} are the solutions where the structure is approximated with 5 points (see Figure 2(d)), and structure is approximated with pairs, triplets and the fusion of both, respectively.

Dataset	Running time τ [s] SfmI	Colmap	MM	SfmI BA	Ours E^{rnd}_{p+tri}	Ours E_p	Ours E_{tri}	Ours E_{p+tri}
F-P11	3.0	4.3*	11.2*	1.9	0.7	0.5	0.6	0.7
HJ-P8	2.9	2.5*	6.8*	1	0.4	0.3	0.4	0.4
HJ-P25	3.1	18.2*	40.2*	8.9	2.5	1.5	1.6	2.1
C-P10	3.0	4.0*	9.4*	1.7	0.9	0.5	0.6	0.8
C-P19	3.2	8.2*	14.5*	8.3	1.7	0.7	0.8	1.5
C-P30	3.3	24.3*	195.0*	9.7	4.5	1.4	2.4	3.8

Table 5. Strecha benchmark. Running times τ in conjuction with the results presented in Table 4. Note that the times for Ours concern only the BA step. Times markes with (*) correspond to the total incremental SfM.

		Accuracy err [mm]	Running time τ [s]
Colmap		17.5	25.0
MM		27.2	39.1
MM_{Init}		41.0	50.0
MM_{Init} BA		30.0	61.0
Ours	E^{rnd}_{p+tri}	21.0	9.8
Ours	E_p	27.4	3.8
Ours	E_{tri}	18.2	5.3
Ours	E_{p+tri}	17.0	7.2

Table 6. UAV dataset. Average position error (err) and running times (τ). MM_{Init} and MM are the global and sequential SfMs in MicMac. MM_{Init} BA corresponds to the BA refinement on the initial solution. Note that the times for Ours concern only the BA step. In E^{rnd}_{p+tri} a mean value over 1000 repeated calculations is reported.

Hartley, R., Trumpf, J., Dai, Y., Li, H., 2013. Rotation averaging. *International Journal of Computer Vision*, 103(3), 267–305.

Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Indelman, V., Roberts, R., Beall, C., Dellaert, F., 2012. Incremental light bundle adjustment. Georgia Institute of Technology.

Jeong, Y., Nister, D., Steedly, D., Szeliski, R., Kweon, I. S., 2011. Pushing the envelope of modern methods for bundle adjustment. *IEEE transactions on pattern analysis and machine intelligence*, 34(8), 1605–1617.

Jiang, N., Cui, Z., Tan, P., 2013. A global linear method for camera pose registration. *Proceedings of the IEEE International Conference on Computer Vision*, 481–488.

Julià, L. F., Monasse, P., 2017. A critical review of the trifocal tensor estimation. *Pacific-Rim Symposium on Image and Video Technology*, Springer, 337–349.

Kahl, F., Hartley, R., 2008. Multiple-View Geometry Under the L_∞ -Norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9), 1603–1617.

Klopschitz, M., Irschara, A., Reitmayr, G., Schmalstieg, D., 2010. Robust incremental structure from motion. *Proc. 3DPVT*, 2, 1–8.

Kushal, A., Agarwal, S., 2012. Visibility based preconditioning for bundle adjustment. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1442–1449.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Martinec, D., Pajdla, T., 2007. Robust rotation and translation estimation in multiview reconstruction. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1–8.

Mayer, H., 2014. Efficient hierarchical triplet merging for camera pose estimation. *German Conference on Pattern Recognition*, Springer, 399–409.

Moulon, P., Monasse, P., Marlet, R., 2013. Global fusion of relative motions for robust, accurate and scalable structure from motion. *Proceedings of the IEEE International Conference on Computer Vision*, 3248–3255.

Moulon, P., Monasse, P., Perrot, R., Marlet, P., 2016. OpenMVG: Open multiple view geometry. *International Workshop on Reproducible Research in Pattern Recognition*, Springer, 60–74.

Olsson, C., Eriksson, A., Hartley, R., 2010. Outlier removal using duality. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 1450–1457.

Pierrot Deseilligny, M., Cléry, I., 2011. Apero, an open source bundle adjusment software for automatic calibration and orientation of set of images. *Proceedings of the ISPRS Symposium, 3DARCH11*, 269277.

Pierrot-Deseilligny, M., Rupnik, E., Girod, L., Belvaux, J., Maillet, G., Deveau, M., Choqueux, G., 2016. Micmac, apero, pastis and other beverages in a nutshell.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1992. Numerical recipes in C++. *The art of scientific computing*, 2, 1002.

Rodriguez, A. L., López-de Teruel, P. E., Ruiz, A., 2011. Gea optimization for live structureless motion estimation. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE, 715–718.

Rupnik, E., Daakir, M., Pierrot Deseilligny, M., 2017. MicMac–a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards*, 2(1), 14.

Schonberger, J. L., Frahm, J., 2016a. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Schonberger, J. L., Frahm, J.-M., 2016b. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Sim, K., Hartley, R., 2006. Removing outliers using the l\infty norm. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, IEEE, 485–494.

Snavely, N., Seitz, S. M., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3D. *ACM transactions on graphics (TOG)*, 25number 3, ACM, 835–846.

Steffen, R., Frahm, J.-M., Förstner, W., 2010. Relative bundle adjustment based on trifocal constraints. *European Conference on Computer Vision*, Springer, 282–295.

Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Ieee, 1–8.

Sweeney, C., Sattler, T., Hollerer, T., Turk, M., Pollefeys, M., 2015. Optimizing the viewing graph for structure-from-motion. *Proceedings of the IEEE International Conference on Computer Vision*, 801–809.

Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A., 2015. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140, 127–143.

Tomasi, C., Kanade, T., 1992. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2), 137–154.

Triggs, B., McLauchlan, P. F., Hartley, R. I., Fitzgibbon, A. W., 1999. Bundle adjustmenta modern synthesis. *International workshop on vision algorithms*, Springer, 298–372.

Wilson, K., Snavely, N., 2014. Robust global translations with 1dsfm. *European Conference on Computer Vision*, Springer, 61–75.

Zach, C., Klopschitz, M., Pollefeys, M., 2010. Disambiguating visual relations using loop constraints. *Computer Vision and Pattern Recognition (CVPR)*, IEEE, 1426–1433.

Zhu, S., Shen, T., Zhou, L., Zhang, R., Wang, J., Fang, T., Quan, L., 2017. Parallel structure from motion from local increment to global averaging. *arXiv preprint arXiv:1702.08601*.

Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L., 2018. Very large-scale global sfm by distributed motion averaging. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4568–4577.

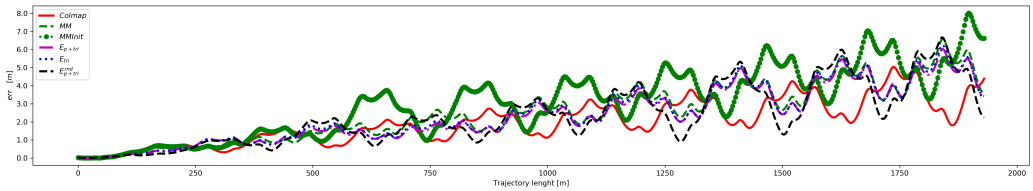

Figure 3. CAR dataset. Average position error (err) along the trajectory w.r.t. ground truth for several structure approximation variants: E^{rnd}_{p+tri} in black dashed line, E_{p+tri} in continuous magenta line, E_{tri} in dotted blue line. MM (green dashed line) and MM_{Init} (green dash-dotted line) are the sequential and global SfMs in MicMac [Pierrot Deseilligny, Cléry, 2011], and in red continuous like it is Colmap's SfM [Schonberger, Frahm, 2016a]. MicMac and Colmap results were computed with with SIFT [Lowe, 2004]. In variant E^{rnd}_{p+tri} a mean value over 100 repeated calculations is reported. Running times for E_{p+tri}, E_{tri}, E^{rnd}_{p+tri}, MM, MM_{Init} and Colmap are 172s, 130s, 170s, 2153s, 715s and 640s respectively. The high frequency drift is due to the cyclic nature of the acquistion geometry.

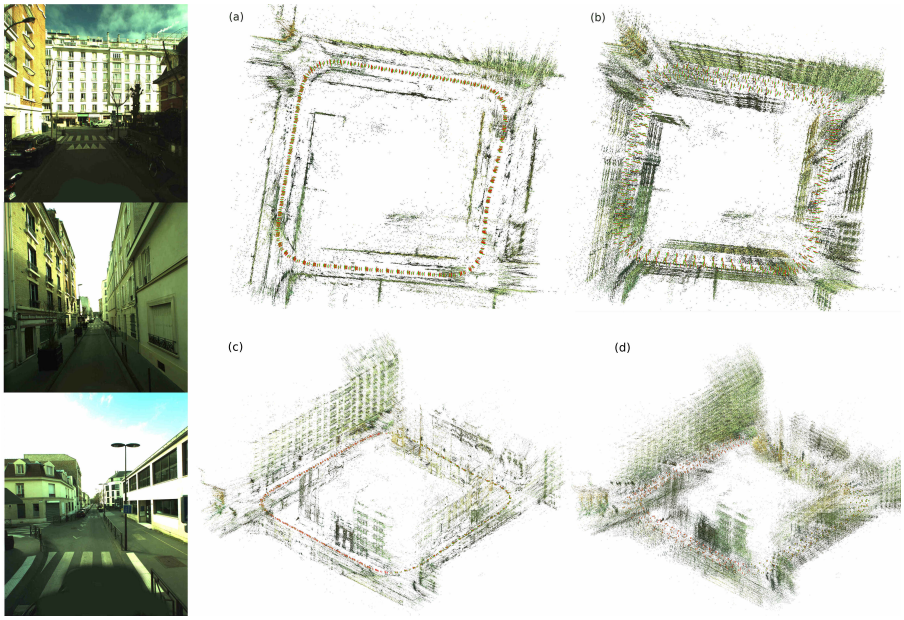

Figure 4. Left: excerpt from the CAR dataset. Right: (a) and (c) is the ground truth in top and side view; (b) and (d) is the equivalent E_{p+tri} result.

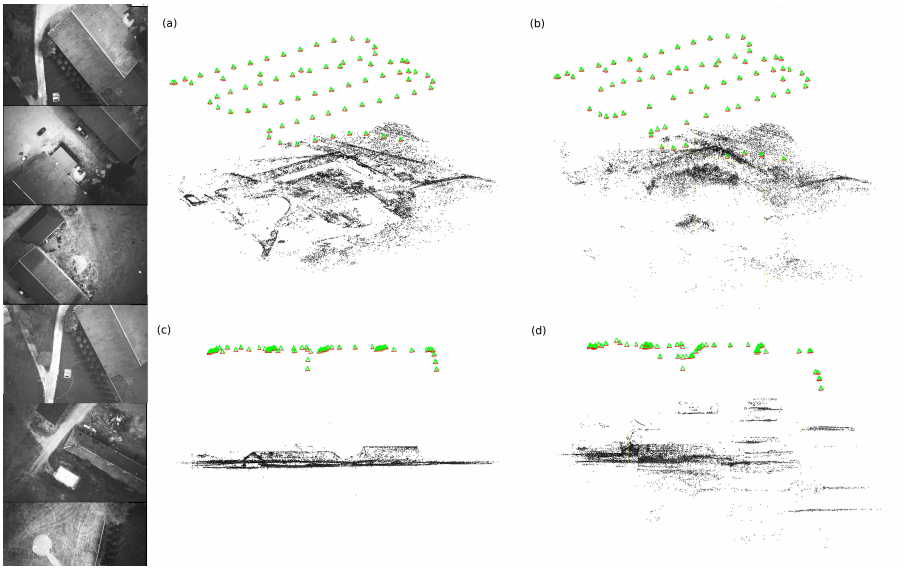

Figure 5. Left: excerpt from the UAV dataset. Right: (a) and (c) is the ground truth in oblique and side view; (b) and (d) is the equivalent SfmInit result.