pOSE: Pseudo Object Space Error for Initialization-Free Bundle Adjustment

Je Hyeong Hong
University of Cambridge
jhh37@cantab.net

Christopher Zach
Toshiba Research Europe
christopher.m.zach@gmail.com

Abstract

Bundle adjustment is a nonlinear refinement method for camera poses and 3D structure requiring sufficiently good initialization. In recent years, it was experimentally observed that useful minima can be reached even from arbitrary initialization for affine bundle adjustment problems (and fixed-rank matrix factorization instances in general). The key success factor lies in the use of the variable projection (VarPro) method, which is known to have a wide basin of convergence for such problems. In this paper, we propose the Pseudo Object Space Error (pOSE), which is an objective with cameras represented as a hybrid between the affine and projective models. This formulation allows us to obtain 3D reconstructions that are close to the true projective reconstructions while retaining a bilinear problem structure suitable for the VarPro method. Experimental results show that using pOSE has a high success rate to yield faithful 3D reconstructions from random initializations, taking one step towards initialization-free structure from motion.

1. Introduction

Structure-from-motion (SfM, visual SLAM or multi-view 3D reconstruction) aims to generate 3D models and camera poses from multiple overlapping images. A complete SfM framework usually consists of several stages, starting from feature extraction and matching, and ranging to a final bundle adjustment step aiming to explain all image observations and correspondences by finding the most probable configuration of 3D structure and camera parameters. The first stages of a typical SfM pipeline (feature extraction, robust matching and relative pose verification) are relatively well understood, and different SfM toolkits and frameworks vary surprisingly little in their respective implementations of these initial steps. Likewise, the final bundle adjustment is generally understood as an instance of a nonlinear least squares problem and implemented accordingly using e.g. the Levenberg-Marquardt (LM) algorithm [26, 31].

It is also well understood that bundle adjustment requires a fairly good initialization for the 3D structure and camera matrices in order to determine a good (local) minimum. The various proposals in the literature for structure-from-motion computation (e.g. [46, 10, 34, 54, 51, 44] among many others) are very diverse in how this initial estimate is obtained. Since no gold-standard method for SfM computation has emerged over several decades of research in this field, we conjecture that a) SfM is a hard problem and b) the sources of its difficulty are not well understood. By looking at a typical bundle adjustment objective

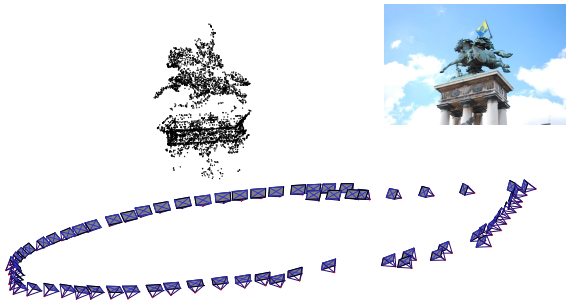


Figure 1. Reconstruction of Vercingetorix [39] from arbitrary initial camera and point parameters.

$$\min_{\substack{\{\mathtt{R}_i\},\{\mathbf{t}_i\},\{\tilde{\mathbf{x}}_j\} \\ \mathtt{R}_i \in SO(3)}} \sum_{(i,j)\in\Omega} \rho\left(\left\| \pi\left(\mathtt{K}_i[\mathtt{R}_i \mid \mathbf{t}_i]\tilde{\mathbf{x}}_j \right) - \mathbf{m}_{i,j} \right\|_2^2 \right) \quad (1)$$

(where ρ is a robust kernel, $\pi([x,y,z]^\top) := [x/z, y/z]^\top$ is the perspective projection, $\{\mathtt{K}_i[\mathtt{R}_i \mid \mathbf{t}_i] = \mathtt{P}_i\}$ are the camera matrices, $\{\mathbf{x}_j\}/\{\tilde{\mathbf{x}}_j\}$ are the inhomogeneous and homogeneous 3D points, $\mathbf{m}_{i,j} \in \mathbb{R}^2$ is the 2D observation of point j in image i respectively and Ω represents the set of visible projections), we conclude that this objective function has several sources of nonlinearities:

1. Each rotation matrix $\mathtt{R}_i$ is constrained to $SO(3)$, which is a non-convex manifold of orthogonal matrices with positive determinants. In this work, we avoid this difficulty by using a stratified approach that first operates in the projective space.
2. The projection function π is non-convex. In general, the (convex) object-space error is a good surrogate for the image-based reprojection error, which we will build on in this work.

3. The robust cost function ρ may (and in almost all cases will) introduce a large number of local minima. If one has to rely strictly on robustified costs (because input data cannot be made sufficiently clean), there is not much one can do from an optimization perspective.
4. There is a bilinear interaction $\mathtt{R}_i\mathbf{x}_j$ (or more generally $\mathtt{P}_i\tilde{\mathbf{x}}_j$) between camera matrices and 3D points. This property is at the core of any image formation model since 3D points in world coordinates have to be projected to the respective camera space. The present work mainly aims to evaluate the difficulty induced by this bilinearity.

Thus, one of the answers we try to provide is the following: how difficult is (non-robust) SfM as an optimization problem? In other words: how wide is the basin of convergence of a modified projective bundle adjustment method started from arbitrary initial values for camera matrices and 3D structure? Since the basin of convergence of any robustified cost function will be inherently narrow, we focus our attention to the non-robust setting (and leave investigations into the robust setting for future work).

It is known that applying bundle adjustment for perspective (pinhole) camera models directly from arbitrary starting points is futile. At the same time it is known that the affine camera model is non-problematic for bundle adjustment (provided the right optimization approach is used), even without a sensible initialization of poses and 3D structure. This observation is leveraged in [23], where it is proposed to solve SfM by a sequence of bundle adjustment tasks with increasing difficulty: the first bundle adjustment round solves SfM for the affine camera model, which is followed by a projective bundle adjustment stage.

In this work, we propose to replace the affine bundle objective by a variant of the object-space error—which we call the "pseudo object-space error" (pOSE, see §3)—better suited for perspective camera models. Empirically, pOSE retains the wide basin of convergence and can therefore be used to initialize our SfM pipeline (see §4) from random initial camera matrices. This also implies that the major reason for SfM being a hard problem is getting the data association step right in order to remove false positive correspondences (further discussed in §5). We upgrade the initial projective reconstruction to the metric frame by employing a simplified self-calibration step (which assumes that at least approximate camera intrinsics are given).

Figure 1 shows an illustrative outcome of our method (i.e. the 3D point cloud and respective camera poses), which was obtained from random initialization. Figure 2 demonstrates that our pseudo object-space error (which is parametrized by a blending weight $\eta \in [0, 1]$) is able to capture the perspective structure of the pinhole camera model ($\eta \approx 0$) much better than the affine model ($\eta \gg 0$).

2. Related work

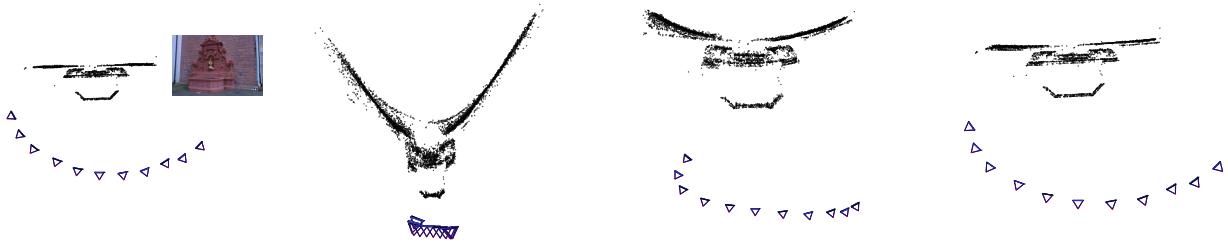
In this section, we first review some work in the literature that forms the basis of this paper.

Factorization approaches in structure-from-motion It was Tomasi and Kanade [52] who first introduced the factorization approach in structure-from-motion. Their very first work showed that, given fully visible and outlier free image observations under the orthographic camera model, it is possible to recover both camera poses and 3D structure using the singular value decomposition (SVD). Their work was later generalized to other affine camera models such as the weak perspective and paraperspective models. Sturm and Triggs [49] proposed projective factorization methods in which projective depths are added as new variables and estimated in an alternating fashion, and [38] presents a variant of the Sturm/Triggs method guaranteed to converge. All these algorithms require every 3D point to be visible in each image, but they can be generalized to problems with missing data by replacing the SVD step in joint estimation of poses and 3D points with iterative matrix factorization with a priori rank [6, 14, 36, 20].

Note that projective factorization by itself an ill-posed problem with degenerate solutions (by collapsing camera matrices or 3D structure to zero) and therefore requires addition of extra constraints (usually on the projective depths) to yield physically meaningful reconstructions. Recently, Nasihatkon et al. [35] reviewed various options how to choose these constraints in order to avoid such degenerate solutions. It is revealed that some conditions proposed in the literature only lead to necessary, but not sufficient conditions for guaranteeing a valid reconstruction, and the authors propose sufficient and necessary conditions on the projective depths to ensure non-degenerate solutions. These conditions are named "generalized projective reconstruction theorem" or GPRT for short. Unfortunately the seemingly strong theoretical insight is limited to fully visible and noise-free (ideal) image observations.

Projective factorization methods are generally global methods working in a non-incremental fashion, simultaneously integrating all image observations. Nevertheless, several incremental methods for projective reconstruction exist, which are designed to handle outliers in the image points as well as missing observations (e.g. [11, 19, 30, 5, 32, 29]). A fundamental shortcoming of incremental, non-global SfM methods is the intrinsic vulnerability to successive accumulation of drift.

Variable projection (VarPro) VarPro dates back to Golub and Pereyra's work in 1973 [13]. In summary, VarPro applies a second order optimizer such as the Gauss-Newton on a reduced objective, after optimally eliminating one set of the unknowns. It is especially applicable to factorization problems, since in these problem instances one of



(a) Ground truth (b) $\eta = 0.5$ (c) $\eta = 0.1$ (d) $\eta = 0.01$

Figure 2. Metric reconstructions of Fountain-P11 obtained by solving pOSE illustrated in §3.

the involved unknown factors can be eliminated in closed form. It has been shown repeatedly [36, 14, 37, 20, 22] that VarPro applied on geometric vision problems has much higher probability of reaching a global optimum (termed *success rate* in this paper) than using a second order method on the full problem (*joint optimization*, i.e. without eliminating one set of unknowns). In [22], it is argued that joint optimization encounters numerical problems in temporarily ill-conditioned solutions and begins to stall. VarPro, which also turns out to be closely connected to joint optimization, has generally no trouble improving the current solution.

To our knowledge, this is the first work aiming for full 3D reconstruction by directly leveraging non-linear refinement without relying on carefully estimated initial structure and motion. Strelow [48] used nonlinear-VarPro, but used noisy ground truth (i.e. not arbitrary) camera matrices and points as initialization. Zheng et al. [58] incorporated metric constraints but their work was limited to weak-perspective camera models and was tested using only the inlier-only dinosaur dataset. The work of Hong et al. [23] is closest to this work: they extended VarPro for affine bundle adjustment to projective bundle adjustment, but did not proceed to the metric reconstruction stage. The utilized image point tracks were guaranteed to be clean and free from outliers. Also, their method is relatively slow as it did not incorporate the recently identified equivalence between non-linear VarPro and the Schur complement [22].

3. Pseudo Object Space Error (pOSE)

We now propose the *pseudo object-space error* (pOSE), which is a surrogate objective for the bundle adjustment cost that keeps the bilinear factorization structure in its residual. In short, pOSE is a convex combination of the object space error (ℓ_{OSE}) and the affine projection error (ℓ_{Affine}), where

$$\ell_{\mathsf{OSE}} := \sum_{(i,j)\in\Omega} \left\| \mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j - (\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j)\mathbf{m}_{i,j} \right\|_2^2 \tag{2}$$

$$\ell_{\mathsf{Affine}} := \sum_{(i,j)\in\Omega} \left\| \mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j - \mathbf{m}_{i,j} \right\|_2^2 \tag{3}$$

$$\ell_{\mathsf{pOSE}} := (1-\eta)\ell_{\mathsf{OSE}} + \eta\ell_{\mathsf{Affine}} \tag{4}$$

with $\eta \in [0,1]$. We use the notation $\mathtt{P}_{i,1:2} \in \mathbb{R}^{2\times 4}$ for the first two rows of the camera matrix $\mathtt{P}_i \in \mathbb{R}^{3\times 4}$, and $\mathbf{p}_{i,3} \in \mathbb{R}^4$ for the last row of $\mathtt{P}_i$. Note that the last element of each homogeneous 3D point $\tilde{\mathbf{x}}_j \in \mathbb{R}^4$ is fixed to 1 when minimizing (4).

The main intuition behind ℓ_{pOSE} is that ℓ_{OSE}, which most closely resembles the bundle adjustment objective (1) and has the desired bilinear problem structure, has an inevitable degenerate global optimum, and therefore the added bilinear ℓ_{Affine} term is a natural choice to prevent this degeneracy.

Calling ℓ_{OSE} an object space error is a slight misnomer, since it penalizes squared point-line distances parallel to the image plane (instead of the shortest point-line distances perpendicular to the line). We nevertheless keep the terminology of object space error based on the fact, that the error is induced by distances in 3D object space. Further, the proper object space error [28] is more suited to model spherical projections rather than projection onto the image plane.

Note that ℓ_{pOSE} can be written as

$$\ell_{\mathsf{pOSE}} = \sum_{(i,j)\in\Omega} \left\| \begin{matrix} \sqrt{1-\eta}\left(\mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j - (\mathbf{p}_{i,3}^\top\tilde{\mathbf{x}}_j)\mathbf{m}_{i,j}\right) \\ \sqrt{\eta}\left(\mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j - \mathbf{m}_{i,j}\right) \end{matrix} \right\|_2^2 \tag{5}$$

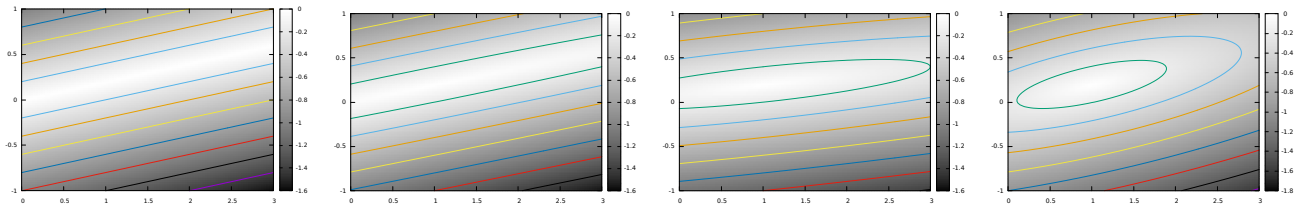
which immediately reveals, that ℓ_{pOSE} is an instance of bilinear factorization problems (see [21] for details). It is also evident that the non-zero pattern of the Hessian (or Gauss-Newton approximated Hessian) is the same as for ℓ_{OSE}.

3.1. Properties of ℓ_{pOSE}

ℓ_{pOSE} (likely) avoids degenerate solutions By expanding and rearranging terms, ℓ_{pOSE} can be written as

$$\begin{aligned} \ell_{\mathsf{pOSE}} = &\sum_{(i,j)\in\Omega} \left\| \mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j - \left((1-\eta)\mathbf{p}_{i,3}^\top\tilde{\mathbf{x}}_j + \eta\right)\mathbf{m}_{i,j} \right\|_2^2 \\ &+ \eta(1-\eta)\sum_{(i,j)\in\Omega} \left\|\mathbf{m}_{i,j}\right\|^2 \left(\mathbf{p}_{i,3}^\top\tilde{\mathbf{x}}_j - 1\right)^2. \end{aligned} \tag{6}$$

The first term is essentially an object space error (measured parallel to the image plane) between the 3D point (transformed into camera space) and a distorted line-of-sight,



(a) OSE, $\eta = 0$ (b) pOSE, $\eta = 0.05$ (c) pOSE, $\eta = 0.5$ (d) rOSE, $\eta = 0.05$

Figure 3. Contour plots of ℓ_{pOSE}, $(1-\eta)(x-uz)^2 + \eta(x-u)^2$ (a-c), and ℓ_{rOSE}, $(1-\eta)(x-uz)^2 + \eta(z-1)^2$ (d), for 1D cameras. u is set to 0.2. For better visibility we actually plot the square root of the function value.

which is the convex combination of pinhole and affine camera rays. The second term favours solutions (as long as $\|\mathbf{m}_{i,j}\|$ is non-zero) that have visible projective depths close to one. Consequently, ℓ_{pOSE} can be understood as regularized and modified object space error.

Since ℓ_{pOSE} intrinsically prefers solutions with positive projective depths (at least for observed image points), it is not expected to produce degenerate solutions for general input data (although it is not guaranteed to avoid degeneracies for all possible inputs, e.g. set $\mathbf{m}_{i,j}$ to 0 for all i and j).

VarPro vs joint optimization In [22] it was argued that for affine factorization-based SfM problems, the VarPro method and joint optimization (i.e. using the LM algorithm jointly w.r.t $\mathtt{P}_i$ and $\tilde{\mathbf{x}}_j$) behave very different: joint optimization suffers from "stalling" behaviour whenever affine camera rays are close to being parallel (since a small update of such camera parameters yield large updates for the 3D points, which is prohibited by Levenberg damping). VarPro avoids this shortcoming by allowing 3D points to freely follow the updates of camera parameters in all cases.

The situation for ℓ_{pOSE} is similar to the affine setting, but one has stronger conditions for camera rays to be parallel: in the affine setting, the parallelity of optical axis is sufficient, whereas in the projective scenario the 3×3 submatrices need to (approximately) satisfy $\mathtt{P}_{i,1:3,1:3} \propto \mathtt{P}_{j,1:3,1:3}$. Hence, one might expect that joint optimization in the projective setting leads to the stalling behaviour less often than in the affine setting, but empirically joint optimization is still clearly inferior to VarPro for our tested choices of η (see Figure 4). Since the success rate of VarPro increases for larger values of η, the selected value for η represents a tradeoff between success rate and the amount of geometric distortion. In view of Figures 2 and 4 we chose $\eta = 0.05$ as the "sweet spot" in our experiments.

Alternative regularizations Degenerate solutions for $\{\mathtt{P}_i\}$ and $\{\tilde{\mathbf{x}}_j\}$ can be avoided by enforcing $\mathtt{P}_i\tilde{\mathbf{x}}_j \geq \delta$ (for all $(i,j) \in \Omega$) for some $\delta > 0$. Adding these constraint to ℓ_{OSE} makes the problem non-smooth and much more difficult to solve. Adding a barrier function such as $-\alpha \sum_{(i,j)\in\Omega} \log(\mathtt{P}_i\tilde{\mathbf{x}}_j - \delta)$ for an $\alpha > 0$ would require either using joint optimization or nonlinear VarPro. As mentioned above, joint optimization frequently leads to stalling behaviour, and nonlinear VarPro has been demonstrated to have a significantly smaller convergence basin [23]. As a result, we rule out adding inequality constraints on the projective depths and respective penalizers or barrier functions.

Instead of defining the target objective as a convex combination of an object-space error and an affine factorization error, one can consider directly a regularized object-space error ℓ_{rOSE} by combining ℓ_{OSE} with a term penalizing visible projective depths,

$$\ell_{\mathsf{rOSE}} = (1-\eta) \sum_{(i,j)\in\Omega} \left\|\mathtt{P}_{i,1:2}^\top \tilde{\mathbf{x}}_j - (\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j)\mathbf{m}_{i,j}\right\|_2^2 + \eta \sum_{(i,j)\in\Omega} (\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j - 1)^2. \tag{7}$$

Note that ℓ_{rOSE} essentially constrains the projective depths of all observed image points, not just for a subset as required by the GPRT (recall §2). If there is a perfect solution with zero object-space error (thus $\mathtt{P}_{i,1:2}^\top \tilde{\mathbf{x}}_j = (\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j)\mathbf{m}_{i,j}$ for all $(i,j) \in \Omega$), then ℓ_{pOSE} and ℓ_{rOSE} reduce to

$$\ell_{\mathsf{pOSE}} = \eta \sum_{(i,j)\in\Omega} \|\mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j - \mathbf{m}_{i,j}\|_2^2 = \eta \sum_{(i,j)\in\Omega} \|\mathbf{m}_{i,j}\|_2^2 \left(\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j - 1\right)^2 \tag{8}$$

$$\ell_{\mathsf{rOSE}} = \eta \sum_{(i,j)\in\Omega} (\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j - 1)^2. \tag{9}$$

Since observations $\mathbf{m}_{i,j}$ are on the image plane (with depth 1), we have for regular, not extremely wide field-of-view cameras $\|\mathbf{m}_{i,j}\| \leq 1$, and ℓ_{pOSE} perturbs the object-space error ℓ_{OSE} less than ℓ_{rOSE} in this case. Assessing the pros and cons of ℓ_{rOSE} over ℓ_{pOSE} is a subject of future work. Inspired by one set of sufficient conditions for projective reconstruction [35], we introduce ℓ_{GPRT} which penalizes projective depths in analogy to ℓ_{rOSE}, but only for indices (i,j) being on a step-like matrix on top of visible elements (see [21]). This choice essentially fixes the projective frame of the reconstruction.

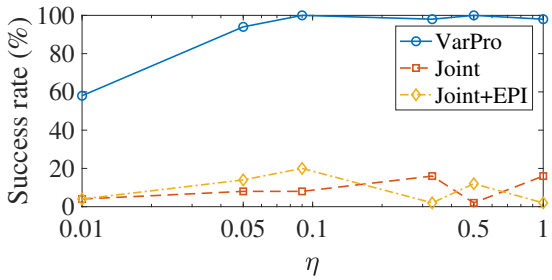


Figure 4. Comparison of VarPro and joint optimization-based algorithms tested on the small Dino sequence in Table 1. For each η, we report the fraction of runs in which each algorithm yields the best seen minimum from random initialization. (See [21] for the results on other datasets.)

Alternative objectives Our choice of using a perturbed object-space error is largely motivated by the bilinear nature of the cost function and the availability of efficient (VarPro) algorithms. As mentioned above, the classical bundle adjustment objective has a very narrow basin of convergence for reaching a good solution [23]. For similar reasons, we ruled out convex objectives such as ℓ_1 and Huber cost functions (e.g. [45, 9, 57]).

4. Stratified bundle adjustment

We now illustrate our multistage pipeline for initialization-free bundle adjustment. The key idea is simple; given a set of point tracks and randomly sampled camera and point parameters, seek for an initial solution (camera poses and 3D structure) by solving the pOSE optimization problem (see §3), refine it in the projective frame, upgrade the solution to metric followed by a final metric refinement step. These steps are also summarized in Algorithm 1.

pOSE optimization As shown in §3, pOSE is designed to closely resemble a projective factorization cost as well as keeping the objective bilinear. This is because bilinear problems have been empirically shown to have wide basin of convergence for the variable projection (VarPro) family of algorithms [8, 36, 14, 20, 22].

We use Ruhe and Wedin algorithm 2 (RW2) [25, 43], which is a variant of the VarPro algorithm that uses an approximated Jacobian, and is therefore easier to implement than Ruhe and Wedin algorithm 1 (RW1) [43], which uses the full Gauss-Newton approximation of the Hessian. This makes RW2 easier to implement than RW1 and also algorithmically very similar (but not equal) to joint optimization with the Schur complement trick [22].

Recall that incorporating robustness at this stage would a) introduce many local minima and b) destroy the bilinear structure convenient for VarPro. We attempt to mitigate this issue by generating mostly-inlier tracks using the approach illustrated in §5.

Algorithm 1 Our initialization-free BA pipeline

Input: a set of geometrically-verified point tracks

1. Solve the L^2-norm pOSE problem (see §3) from arbitrarily sampled cameras and points using VarPro.
2. Refine the solution using a robustified projective BA algorithm incorporating nonlinear VarPro, and discard points with large reprojection errors.
3. Upgrade the above solution to metric and throw away points which do not satisfy cheirality constraints.
4. Refine the solution in the metric space.

Output: metric camera poses and 3D reconstruction

Projective refinement Once cameras and points are obtained by optimizing ℓ_{pOSE}, they are refined by minimizing the gold standard objective [16]

$$\sum_{(i,j)\in\Omega} \rho\left(\left\| \frac{\mathtt{P}_{i,1:2}\tilde{\mathbf{x}}_j}{\mathbf{p}_{i,3}^\top \tilde{\mathbf{x}}_j} - \mathbf{m}_{i,j} \right\|_2^2\right) \tag{10}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is an isotropic robust loss function. This work uses the Cauchy kernel (see [21]). Although (10) can be solved by jointly minimizing over the cameras and points, we implement Strelow's nonlinear VarPro [48], which is an extension of standard VarPro to nonseparable problems (i.e. nonlinear in both sets of variables), and it is demonstrated to have a slightly wider basin of convergence than joint optimization [23]. Furthermore, it has been shown [22] that the iteration complexity of VarPro is approximately equal to that of joint optimization with embedded point iterations [24], which simply amounts to performing additional triangulations after each joint update of cameras and 3D points.

The optimization is carried out in homogeneous coordinates incorporating the Riemannian manifold optimization [1, 23] (algorithmically equivalent to local parameterization in [16, 2]). After the refinement, we discard points having a maximum reprojection error greater than 2 pixels.

Metric upgrade The resulting refined projective camera matrices need to be upgraded to a metric frame to satisfy the $SE(3)$ (group of 3D Euclidean isometries preserving orientation) manifold constraints (after taking out the calibration matrices). This *autocalibration* step is a well-studied topic in 3D computer vision [15, 3, 17, 18, 41, 40, 7, 12]. It is usually carried out by first finding an ambiguity matrix $\mathtt{H} \in \mathbb{R}^{4\times 4}$ which transforms the stack of camera matrices to most closely satisfy the $SE(3)$ constraint followed by the actual manifold projection. Our method is based on [40], with a slight change in the minimized objective to incorporate VarPro once more. The aim is to find $\mathtt{H}$ that satisfies

$$\mathtt{P}_i\mathtt{H} = \mathtt{P}_i \begin{bmatrix} \mathtt{A} & \mathbf{0} \\ \mathbf{c}^\top & 1 \end{bmatrix} \approx \mathtt{K}_i \begin{bmatrix} \mathtt{R}_i & \mathbf{t}_i \end{bmatrix} \quad \forall \quad 1 \le i \le N \tag{11}$$

Algorithm 2 Generating point tracks from images

Input: images and camera intrinsics

1. Obtain pairwise feature matches using SIFT.
2. Verify matches using two view geometric constraints.
3. Verify matches using triplet filtering.
4. Convert pairwise matches to optimal point tracks by using Olsson and Enqvist's algorithm [39].
5. Mask out track segments that are not consistent with the estimated epipolar geometries.

Output: point tracks

where N represents the total number of images, $\mathtt{K}_i \in \mathbb{R}^{3\times3}$ is the upper-triangular calibration matrix of camera i, $[\mathtt{R}_i|\mathbf{t}_i] \in SE(3)$ represents i-th camera's pose and $[\mathbf{c}^\top\ 1]^\top \in \mathbb{R}^4$ represents the plane at infinity. The last column of $\mathtt{H}$ can be set to $[\mathbf{0}^\top 1]^\top$ as it only accounts for global translation and scaling. Since we assume camera intrinsics are known a priori, we utilize the normalized quantity $\tilde{\mathtt{P}}_i := \mathtt{K}_i^{-1}\mathtt{P}_i$. By defining $\tilde{\mathtt{H}} \in \mathbb{R}^{3\times4}$ to comprise the three left-most columns of $\mathtt{H}$, we obtain

$$\tilde{\mathtt{P}}_i\tilde{\mathtt{H}}\tilde{\mathtt{H}}^\top\tilde{\mathtt{P}}_i^\top = \tilde{\mathtt{P}}_i \begin{bmatrix} \mathtt{I} & \mathbf{c} \\ \mathbf{c}^\top & \|\mathbf{c}\|^2 \end{bmatrix} \tilde{\mathtt{P}}_i^\top \simeq \mathtt{R}_i\mathtt{R}_i^\top = \mathtt{I}. \tag{12}$$

Various constraints [41, 40] have been proposed to find $\mathbf{c}$ that most closely satisfies (12). Here, we minimize

$$\min_{\mathbf{c},\{\alpha_i\}} \sum_{i=1}^{F} \left\| \alpha_i\tilde{\mathtt{P}}_i\tilde{\mathtt{H}}(\mathbf{c})\tilde{\mathtt{H}}(\mathbf{c})^\top\tilde{\mathtt{P}}_i - \mathtt{I} \right\|_F^2, \tag{13}$$

where $\{\alpha_i\}$ is the set of individual camera scales and $\|\cdot\|_F$ is the Frobenius norm. Since (13) is linear in $\{\alpha_i\}$, we can use VarPro [13] to solve this efficiently.

Projecting the upgraded solution to the $SE(3)$ manifold is carried out by projecting the rotation part of each camera to $SO(3)$ using Arun et al.'s method [4], followed by flipping all signs of camera translations and 3D points depending on the global cheirality of the reconstructed scene. Any 3D point behind any observing camera is discarded.

Metric refinement The above metric upgrade procedure usually increases the total reprojection error as metric cameras have lower degrees of freedom than projective ones. Hence, an additional step is required to refine camera poses and 3D structure, which is achieved by minimizing the gold standard reprojection error (1). Cameras and points are jointly minimized since there is no visible advantage in employing nonlinear VarPro. Rotations are formulated using the axis-angle representation and Rodrigues' formula [42].

5. Generating point tracks from matches

Our stratified BA pipeline in §4 requires point tracks consisting mostly of inliers. This is due to the limitation that the pOSE objective (4) (or any other reasonable objective) cannot be robustified without sacrificing the convergence basin of VarPro. Hence, the goal of the method described in this section is to generate as clean tracks as possible before feeding them into our BA pipeline. Unfortunately, this is a non-trivial problem on its own for two reasons. First, the estimated epipolar geometries may be grossly incorrect due to perceptual aliasing (e.g. [56]). Second, naively connecting matches across multiple images may form inconsistent tracks in which more than one feature from the same image could participate (c.f. Figure 5). Solving these issues is still an active research problem in computer vision (see e.g. [33, 53] for recent developments). Our approach focuses on scalability and is summarized in Algorithm 2.

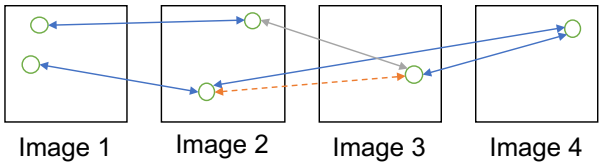


Figure 5. An illustration of a potential issue arising when generating point tracks. The solid blue and grey lines represent 2-view geometrically-verified matches. Naively joining these matches leads to two features participating in image 2. Hence, a match (solid grey) is discarded according to some predefined rule or algorithm. Consequently, this induces a new match between image 2 and 3 (orange). We propose that these implicitly arising matches should be verified to satisfy existing 2-view geometric constraints.

Two view geometric verification The first step is a standard 2-view geometric verification step to remove the "easy" outliers, and we additionally refine the essential matrices using the surviving inlier matches. We set this threshold to be 1.5 px for 2MP images and scale this threshold linearly with the image size.

Triplet filtering Outliers arising from repetitive structures such as windows may remain and may lead to grossly incorrect essential matrices. Many of these false positive relative poses can be detected by checking the self-consistency of relative rotations for image triplets [56, 34, 51]. We require the chained relative rotations to have at most a 5° residual angle. We refrain from including larger loops [56, 10] due to their high computation cost.

Our method iteratively discards the currently most violating image pair (i.e. the one participating in the largest number of inconsistent triplets), until all triplets are consistent. Thus, our triplet filtering method is more aggressive in removing image pairs than [34, 51] (which only require an image pair to participate in at least one consistent triplet).

Point tracking algorithm After matches are geometrically verified using two-view and three-view constraints, we employ Olsson and Enqvist's algorithm [39] to generate algebraically-optimal consistent tracks. In summary, when connecting pairwise matches leads to two or more features

from the same image being joined (e.g. Figure 5), this algorithm selects one based on a track reliability criterion measured by the minimum number of feature matches.

Revisiting the geometric constraints As shown in Figure 5, generated point tracks may induce new matches indirectly, depending on which matches pass the local verification steps. It is possible that the induced feature matches do not satisfy the respective two-view epipolar and cheirality constraints. We propose to revisit each of these newly created matches and verify that they satisfy all desired constraints. If they do not, then we treat these matches as missing data. This is only a partial solution as multiple real point tracks may still be incorrectly merged into a single point track. Ultimately, one should incorporate geometric constraints in measuring the point track reliability.

6. Experimental results

Our experiments have been designed to

1. empirically observe the size of the convergence basin of our pOSE-based stratified BA strategy (proposed in §4) on point tracks with mostly inliers,
2. check whether the GPRT constraints [35] are still sufficient conditions for solving the object space error on sequences with noise and missing data, and
3. whether our pOSE-based BA pipeline equipped with the point track generation algorithm from §5 can accurately solve real SfM problems.

To answer these questions, we carried out two experiments on small to medium-sized real SfM sequences. For convenience, we will use the abbreviation pOSE when referring to our pOSE-based stratified BA pipeline, and GPRT when referring to the same pipeline but with the first stage objective replaced by the object space error (2) with the GPRT constraints (using a step-like matrix, c.f. §3).

Implementation We conducted experiments on a machine with Intel Core i7-7800X CPU (6 cores) and 32GB RAM. For feature detection and matching, we used COLMAP [44] to generate 2-view verified pairwise matches via exhaustive matching. COLMAP was modified to refine and output corresponding essential matrices. All other stages were either implemented in MATLAB or using the Google Ceres Solver [2] library patched to enable the VarPro method [13] according to the guidelines in [22].

Seq.	# img	# pts	Fill (%)	pOSE SRI	pOSE $\bar{t}(s)$	GPRT SRI	GPRT $\bar{t}(s)$
House	10	672	42.4	100	4.2	35	5.8
Corridor	11	737	49.8	100	1.7	15	9.0
Dino (S)	36	319	23.1	96	3.0	0	10.1
Dino (L)	36	4983	9.2	98	6.9	0	32.3
Wilshire	190	411	39.3	100	38.1	0	272.2
Blue bear	196	2480	19.3	80	70.7	0	337.7

Table 1. A list of small classic inlier point tracks used and corresponding results. $\bar{t}$ denotes mean runtime. The pOSE-based BA pipeline has large success rates for these inlier tracks (SRIs) while the GPRT-based pipeline fails on many of these. Fill defines the proportion of visible projections over all possible projections.

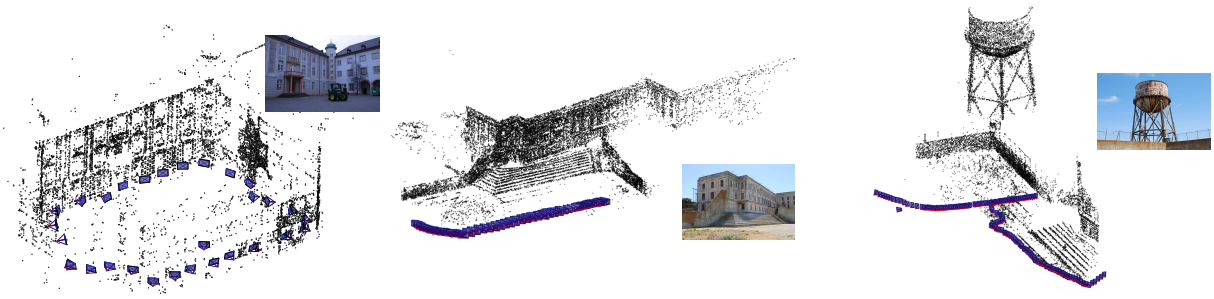
Seq. Pipeline	pOSE	GPRT	Theia	COLMAP
Fountain-P11	2.8	2.8–4.5	**2.4**	2.8
Entry-P10	7.1	6.4–7.0	**6.0**	6.3
Herz-Jesu-P8	3.4–3.9	3.8–4.5	**3.1**	4.1
Herz-Jesu-P25	5.2	5.1	**5.1**	5.2
Castle-P19	**24.5**–41.1	N/A	25.3	24.9
Castle-P30	**21.7**–26.0	N/A	**21.7**	23.2

Table 2. Accuracy comparison of our BA strategy using pOSE against other pipelines on Strecha et al.'s benchmark datasets [47]. The castles are more difficult due to repetitive structures. The reported values are the mean errors in camera positions (in mm). N/A means no successful run. Theia refers to its global SfM pipeline, and its results are referenced from [50].

Initialization Since VarPro optimally eliminates the 3D structure from the pOSE residual (4), we only really need to sample the camera parameters (i.e. $\{\mathtt{P}_i\}$). Other previous work in matrix factorization [6, 8, 36, 20] simply sampled these from an isotropic Gaussian distribution with mean $\mathbf{0}$ and variance $\mathtt{I}$ in the pixel coordinates. We employ a similar sampling method but set the the mean of the sampling distribution to be at the center of the images. We also set the norm of each row of the sampled camera matrix to 1 in order to improve numerical stability (see [21] for details).

Procedures and results In the first experiment, we compare the performance of pOSE and GPRT on point tracks known to be free of outliers. Some of these tracks are publicly available classic datasets (e.g. dinosaur), while others are derived from Olsson's [39] and Strecha et al.'s [47] digital camera images piped through a robust incremental SfM pipeline (COLMAP [44]). We run pOSE and GPRT for a fixed number of runs, each from arbitrarily-sampled cameras and points, and report the fraction of runs (SRI: success rate for inlier tracks) each algorithm reaches the best solution up to predefined tolerance value in terms of camera position errors (see [21] for details). Table 1 and the SRI part of Table 3 shows that pOSE has large basin of convergence across various inlier tracks, and that the GPRT constraints are not sufficient to maintain a large basin of convergence for these datasets with missing entries.

In the second experiment, we build our own tracks using the approach in §5 for each of the image sequences listed in Table 3. We then compare the peformance of pOSE and GRPT on these point tracks. This experiment is carried out to take into consideration that creating consistent inlier tracks is also an essential non-negligible subproblem



(a) Castle-P30 [47] (b) Alcatraz Courtyard [39] (c) Water Tower [39]

Figure 6. Some successful reconstructions obtained from our stratified BA strategy illustrated in §4.

Sequence	# img	# pts	Fill (%)	pOSE SRI (%)	pOSE SRF (%)	pOSE $\bar{t}$ (s)	GPRT SRI (%)	GPRT SRF (%)	GPRT $\bar{t}$ (s)
Fountain-P11	11	9181	50.3	100	100	6.5	15	30	70.3
Entry-P10	10	4270	55.5	100	98	5.2	25	5	27.5
Herz-Jesu-P8	8	3553	60.7	100	100	3.4	30	15	16.1
Herz-Jesu-P25	25	12469	27.3	100	100	12.9	10	5	86.2
Castle-P19	19	5144	27.2	94	88	21.4	0	0	21.3
Castle-P30	30	11531	20.4	94	100	32.1	0	0	73.6
House Martenstorget	12	5934	53.2	100	100	11.3	20	15	47.8
Lund Cathedral (small)	17	9400	30.6	92	96	19.2	10	15	75.6
Gustav II Adolf	57	9562	12.3	100	86	23.7	20	35	82.0
Univ. of West. Ontario	57	6742	14.4	100	98	25.2	5	35	92.6
Vercingetorix	69	5231	10.3	78	92	15.0	10	15	60.0
Lund University Sphinx	70	22770	9.9	96	76	74.7	0	5	150.3
Alcatraz courtyard	133	31558	9.7	94	100	145.5	0	0	1653.8
Water tower	173	25531	8.0	90	76	274.6	40	35	1932.1
Pumpkin	209	25962	4.1	100	100	147.5	0	0	869.7

Table 3. A list of real SfM datasets used and corresponding results. $\bar{t}$ is mean runtime for executing our BA pipelines. Fill defines the proportion of actual visible projections over all possible projections.

in SfM. Similar to the first experiment, we report each algorithm's success rate on each full sequence (SRF) along with average runtime in Table 3. In addition, since Strecha et al.'s datasets (the first six in Table 3) provide ground truth camera poses, we also report the camera position errors of our successful solutions in Table 2 (see [21] for discussions). The benchmark results show that pOSE with custom tracks yields accurate reconstructions and produces state-of-the-art results on the castle datasets, which have repetitive structures. These results, together with the results in Table 3, show that the pOSE-based stratified BA pipeline has large basin of convergence towards accurate reconstructions on small and medium-scale real SfM datasets, if given mostly clean point tracks. Figures 1 and 6 show our successful solutions for some datasets, and others are included in [21].

7. Conclusions

In this paper, we proposed the pseudo object-space error (pOSE) for projective 3D reconstruction, and we have shown that—by using the variable projection (VarPro) method—pOSE has a wide convergence basin and can be efficiently implemented. We also presented a stratified framework, which starts from randomly initialized camera matrices, to obtain ultimately a metric 3D model. We also proposed a combination of algorithms to obtain sufficiently clean correspondences from initial feature matches.

In this work, we have demonstrated competitive results for smaller and medium-scale datasets. It is an open question whether the wide convergence basin of our bundle adjustment formulation is confirmed for large datasets, or if alternative strategies such as applying our framework on medium-sized subsets are necessary.

As clean tracks greatly simplify global SfM pipelines, future work will emphasize on the generation of high quality tracks, for instance by incorporating trifocal tensor relations [27].

Acknowledgement We are grateful for the travel support provided by Roberto Cipolla and Toshiba Research Europe.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. 5

[2] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org, 2014. 5, 7

[3] M. Armstrong, A. Zisserman, and R. Hartley. Self-calibration from image triplets. In *4th European Conference on Computer Vision (ECCV)*, pages 1–16, 1996. 5

[4] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 9(5):698–700, may 1987. 6

[5] S. Avidan and A. Shashua. Threading fundamental matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):73–77, 2001. 2

[6] A. M. Buchanan and A. W. Fitzgibbon. Damped Newton algorithms for matrix factorization with missing data. In *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 316–322, 2005. 2, 7

[7] M. Chandraker, S. Agarwal, D. Kriegman, and S. Belongie. Globally optimal algorithms for stratified autocalibration. *International Journal of Computer Vision (IJCV)*, 90(2):236–254, 2010. 5

[8] P. Chen. Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix. *International Journal of Computer Vision (IJCV)*, 80(1):125–142, 2008. 5, 7

[9] A. Dalalyan and R. Keriven. l_1-penalized robust estimation for a class of inverse problems arising in multiview geometry. In *Advances in Neural Information Processing Systems*, pages 441–449, 2009. 5

[10] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 264–271, 2011. 1, 6

[11] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *European conference on computer vision*, pages 311–326. Springer, 1998. 2

[12] R. Gherardi and A. Fusiello. Practical autocalibration. In *11th European Conference on Computer Vision (ECCV)*, pages 790–801, 2010. 5

[13] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis (SINUM)*, 10(2):413–432, 1973. 2, 6, 7

[14] P. F. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(10):2051–2065, Oct 2011. 2, 3, 5

[15] R. I. Hartley. Euclidean reconstruction from uncalibrated views. In *Joint European-US workshop on applications of invariance in computer vision*, pages 235–256, 1993. 5

[16] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 5

[17] A. Heyden and K. Astrom. Euclidean reconstruction from constant intrinsic parameters. In *13th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 339–343, 1996. 5

[18] A. Heyden and K. Astrom. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 438–443, 1997. 5

[19] A. Heyden, R. Berthilsson, and G. Sparr. An iterative factorization method for projective structure and motion from image sequences. *Image and Vision Computing*, 17(13):981–991, 1999. 2

[20] J. H. Hong and A. W. Fitzgibbon. Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4130–4138, 2015. 2, 3, 5, 7

[21] J. H. Hong and C. Zach. Supplementary document to pOSE: Pseudo Object Space Error for Initialization-Free Bundle Adjustment. https://github.com/jhh37/pose, 2018. 3, 4, 5, 7, 8

[22] J. H. Hong, C. Zach, and A. Fitzgibbon. Revisiting the variable projection method for separable nonlinear least squares problems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3, 4, 5, 7

[23] J. H. Hong, C. Zach, A. W. Fitzgibbon, and R. Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *14th European Conference on Computer Vision (ECCV)*, pages 477–493, 2016. 2, 3, 4, 5

[24] Y. Jeong, D. Nister, D. Steedly, R. Szeliski, and I. S. Kweon. Pushing the envelope of modern methods for bundle adjustment. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1474–1481, 2010. 5

[25] L. Kaufman. A variable projection method for solving separable nonlinear least squares problems. *BIT Numerical Mathematics*, 15(1):49–57, 1975. 5

[26] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathmatics*, 2(2):164–168, 1944. 1

[27] M. I. A. Lourakis and A. A. Argyros. Fast trifocal tensor estimation using virtual parallax. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–93–6, Sept 2005. 8

[28] C. P. Lu, G. D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(6):610–622, Jun 2000. 3

[29] L. Magerand and A. D. Bue. Practical projective structure from motion (p2sfm). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 39–47, Oct 2017. 2

[30] S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 430–437. IEEE, 2000. 2

[31] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 1

[32] D. Martinec and T. Pajdla. Structure from many perspective images with occlusions. *Computer VisionECCV 2002*, pages 542–544, 2002. 2

[33] E. Maset, F. Arrigoni, and A. Fusiello. Practical and efficient multi-view matching. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4578–4586, Oct 2017. 6

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

[35] B. Nasihatkon, R. Hartley, and J. Trumpf. A generalized projective reconstruction theorem and depth constraints for projective factorization. *International Journal of Computer Vision (IJCV)*, 115(2):87–114, 2015. 2, 4, 7

[36] T. Okatani, T. Yoshida, and K. Deguchi. Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 842–849, 2011. 2, 3, 5, 7

[37] D. P. O'Leary and B. W. Rust. Variable projection for nonlinear least squares problems. *Computational Optimization and Applications*, 54(3):579–593, Apr 2013. 3

[38] J. Oliensis and R. Hartley. Iterative extensions of the sturm/triggs algorithm: Convergence and nonconvergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2217–2233, 2007. 2

[39] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *Proceedings of 17th Scandinavian Conference on Image Analysis (SCIA)*, pages 524–535, 2011. 1, 6, 7, 8

[40] M. Pollefeys, R. Koch, and L. V. Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision (IJCV)*, 32(1):7–25, 1999. 5, 6

[41] M. Pollefeys and L. van Gool. Stratified self-calibration with the modulus constraint. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(8):707–724, 1999. 5, 6

[42] O. Rodrigues. De l'attraction des sphéroïdes. Technical report, Apr 1816. 6

[43] A. Ruhe and P. Å. Wedin. Algorithms for separable nonlinear least squares problems. *SIAM Review (SIREV)*, 22(3):318–337, 1980. 5

[44] J. L. Schönberger and J. M. Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 7

[45] Y. Seo, H. Lee, and S. W. Lee. Outlier removal by convex optimization for l-infinity approaches. In *Pacific-Rim Symposium on Image and Video Technology*, pages 203–214. Springer, 2009. 5

[46] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006. 1

[47] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 7, 8

[48] D. Strelow. General and nested Wiberg minimization: L2 and maximum likelihood. In *12th European Conference on Computer Vision (ECCV)*, pages 195–207. 2012. 3, 5

[49] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *4th European Conference on Computer Vision (ECCV)*, pages 709–720. 1996. 2

[50] C. Sweeney. Theia multiview geometry library: Tutorial & reference. http://theia-sfm.org. 7

[51] C. Sweeney, T. Sattler, T. Höllerer, M. Turk, and M. Pollefeys. Optimizing the viewing graph for structure-from-motion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 801–809, 2015. 1, 6

[52] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)*, 9(2):137–154, 1992. 2

[53] R. Tron, X. Zhou, C. Esteves, and K. Daniilidis. Fast multi-image matching via density-based clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4077–4086, Oct 2017. 6

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

[55] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2011.

[56] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1426–1433, 2010. 6

[57] C. Zach and M. Pollefeys. Practical methods for convex multi-view reconstruction. In *European Conference on Computer Vision*, pages 354–367. Springer, 2010. 5

[58] Y. Zheng, S. Sugimoto, S. Yan, and M. Okutomi. Generalizing Wiberg algorithm for rigid and nonrigid factorizations with missing components and metric constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2010–2017, 2012. 3