Calibration-free Structure-from-Motion with Calibrated Radial Trifocal Tensors

Viktor Larsson[1], Nicolas Zobernig[2], Kasim Taskin[3], and Marc Pollefeys[1,4]

[1] Department of Computer Science, ETH Zürich,
[2] Dept. of Information Technology and Electrical Engineering, ETH Zürich,
[3] KTH Royal Institute of Technology
[4] Microsoft Mixed Reality & AI Zurich Lab

Abstract. In this paper we consider the problem of Structure-from-Motion from images with unknown intrinsic calibration. Instead of estimating the internal camera parameters through some self-calibration procedure, we propose to use a subset of the reprojection constraints that is invariant to radial displacement. This allows us to recover metric 3D reconstructions without explicitly estimating the cameras' focal length or radial distortion parameters. The weaker projection model makes initializing the reconstruction especially difficult. To handle this additional challenge we propose two novel minimal solvers for radial trifocal tensor estimation. We evaluate our approach on real images and show that even for extreme optical systems, such as fisheye or catadioptric, we are able to get accurate reconstructions without performing any calibration.

1 Introduction

In this paper we revisit the classical Structure-from-Motion problem [19], which is to recover the camera poses (the motion) and the 3D scene geometry (the structure) from a set of images. Structure-from-Motion pipelines generally fall into one of two categories; incremental or global. Incremental SfM methods (see e.g. [45,53,42]) work by incrementally growing an initial reconstruction by alternating posing in new images and triangulating new 3D points. Global SfM methods (see e.g. [38,37,9,56]) instead first estimate pairwise epipolar geometries. In a second step the relative poses are then fused into a single reconstruction, typically using some form of rotation averaging [18,7,14]. There are also SfM methods which combine the two approaches (e.g. [8,33]).

In all of the above methods it is necessary to know the cameras' internal parameters (camera intrinsics and lens-distortion) to achieve accurate reconstruction results. These parameters can either be found during an offline calibration procedure (e.g. using some calibration object with known structure such as checkerboards, see [55,44]), or they are estimated during the reconstruction.

The second set of methods can be further divided into two approaches. Methods which first perform a projective reconstruction followed by a self-calibration step (see e.g. [20,21,34]). The self-calibration step entails estimating the Dual Absolute Quadric (see [19,6]) by adding assumptions on the camera intrinsics,

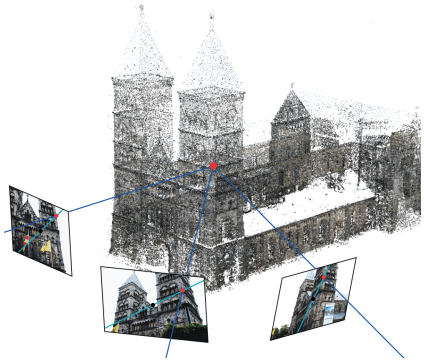

Fig. 1. Structure-from-Motion using radial alignment constraints. Instead of requiring the 3D point to project onto the 2D-point, we only require the projection to lie on the radial lines going through the image point. This makes projection equations invariant to focal length and radial distortion.

such as unit aspect ratio and zero skew. The other approach is to estimate the camera intrinsics during the initial pose estimation process. This can be done either by using solvers which also estimate the internal camera parameters (e.g. [54,28,29,30]) or the camera parameters are initialized with some heuristic guess (e.g. using EXIF tags) followed by bundle adjustment. This approach is used in the open-source framework COLMAP [42,43] that uses focal length sampling [41] and zero-initialized distortion parameters, which are then refined in the bundle-adjustment step. For global SfM with unknown calibration, Sweeney et al. [47] proposed a method which optimizes the consistency of fundamental matrices in order to estimate a consistent focal length for each camera.

While the above approaches can work well in practice, they typically only work reliably for cameras with no or negligible radial distortion. Methods based on finding the calibration during reconstruction are prone to failure for cameras with severe distortion, especially for images where most point correspondences are in regions with high distortion (e.g. close to the image borders).

In this paper we propose a Structure-from-Motion pipeline that does not require knowing or estimating the camera calibration. We only make the assumptions that the camera has square pixels and approximately centered principal point (which is satisfied for essentially all consumer cameras today). The main idea is to use a subset of the geometric constraints which are invariant to any radial change in the projection (such as focal length and most lens-distortion). We show that it is possible to recover high quality reconstructions from this weaker set of constraints even for images from very extreme distortions (e.g. fisheye and catadioptric cameras). In contrast to previous work we do not estimate any distortion model or perform self-calibration.

1.1 Background

The *Radial Alignment Constraint* (RAC) was first introduced by Tsai [51] for camera calibration. The RAC simply states that the projection of a 3D point should lie on the radial line[5] passing through the image point (see Figure 1). This

[5] Radial lines are lines passing through the image center.

constraint has the nice property that it does not depend on the camera's focal length or any purely radial distortion, since these only move the projections along the lines. However, since forward motion also moves the projections radially it is only possible to recover the pose of the camera up to an unknown forward translation using these constraints. This constraint has been used for absolute camera pose estimation with radial distortion, see Kukelova et al. [25] and more recently Camposeco et al. [5] and Larsson et al. [30].

1.2 1D Radial Camera Model

The idea in the RAC later gave rise to the *1D-Radial* camera model which considers the mapping from 3D points to radial lines in the image. Formally, this can be modelled as a projective mapping from $\mathbb{P}^3$ to $\mathbb{P}^1$. Similarly to pinhole cameras, we can describe this mapping with a matrix acting on homogeneous coordinates, i.e. $\mathbf{x} \sim P\mathbf{X}$, where $\mathbf{x} \in \mathbb{P}^1,\ \mathbf{X} \in \mathbb{P}^3,\ P \in \mathbb{R}^{2\times 4}$. Note that in this case the camera matrix P is a 2×4 matrix instead of 3×4. The camera matrix can be thought of as the first two rows of the pinhole camera; giving only the direction of the pinhole projection and not the radial scaling.

As for pinhole cameras we can consider *calibrated* cameras. In the pinhole camera setting we require the first 3×3 block to be a rotation matrix; for radial cameras we require the first 2×3 block to consist of two orthonormal vectors,

$$P = \begin{bmatrix} \mathbf{r}_1^T & t_1 \\ \mathbf{r}_2^T & t_2 \end{bmatrix}, \quad \mathbf{r}_1^T \mathbf{r}_2 = 0, \quad \|\mathbf{r}_1\| = \|\mathbf{r}_2\| = 1. \tag{1}$$

It is important to note that this is not an approximation (like e.g. weak or para-perspective), but instead we essentially consider a subset of the geometric constraints which are independent. This means that for any perspective reconstruction (possibly with non-linear radial distortion), there exists a corresponding 1D radial reconstruction found by just taking the first two rows from each camera. In this paper we show that we can recover this reconstruction without ever estimating the focal length or radial distortion. Note also that the 1D radial camera model is not only valid for central cameras, but any optical system satisfying the RAC, e.g. spherical mirrors (chromeball images) or in general any radially-symmetric mirror, axial cameras, etc. For more details about the 1D-radial camera model see the supplementary material.

1.3 Multiple View Geometry of 1D Radial Cameras

The multiple view geometry of 1D radial cameras was studied by Thirthala and Pollefeys [48] in the framework of multi-focal tensors [49]. Since the radial model only provides a single constraint from each projection, it was shown in [48] that there does not exist any bi- or trifocal tensors for radial cameras in general position, and that it is first in four views that constraints appear. Furthermore, [48] showed that the quadrifocal tensor itself has two internal constraints. Ignoring these constraints the quadrifocal tensor can be linearly estimated from 15

quadruplet point matches. However, as mentioned in [48], this solver is mostly of theoretical interest and not usable for practical purposes due to the high-number of points required. There is currently no known minimal solver for the radial quadrifocal tensor which enforces the internal constraints. In [48] they also consider three special camera configurations where trifocal tensors exist: 1) three principal axes intersect, 2) the scene points are planar and 3) one pinhole camera and two radial cameras. For the tensors they only consider the projective setting, i.e. there are no constraints enforcing that the tensors they estimate can be factorized into calibrated cameras, as in (1).

The tensors above describe projective mappings from $\mathbb{P}^3$ to $\mathbb{P}^1$. There has also been a series of works which consider the multiple view geometry of cameras in lower dimensional spaces, i.e. $\mathbb{P}^2$ to $\mathbb{P}^1$. The trifocal tensor in this setting was first investigated by Quan and Kanade [39]. Faugeras et al. [16] showed that cameras undergoing planar motion can be modeled with 1D cameras by projecting the image measurements onto the ground plane, allowing for estimation with the radial trifocal tensor from [39]. Later, Åström and Oskarsson [4] derived the internal constraint for calibration for this tensor. In this simpler setting the calibration constraint turns out to be linear. These lower dimensional radial trifocal tensors were then used in [11,40,3] for localization of robotic platforms.

Calibrated Multiple View Geometry. In general, enforcing constraints for calibration on multi-focal tensors is very difficult for higher order tensors. For the two-view case (i.e. fundamental vs. essential matrix), these constraints are the well-known trace-constraints[6], $2EE^TE - \mathrm{tr}(EE^T)E = 0$. The constraints for the perspective trifocal tensor have received much attention recently ([32,22,35,13,15]), though currently the minimal solvers are based on homotopy continuation or other iterative methods and have far from practical runtimes, especially compared to their two-view counterparts. In this paper we will show that there exist analogous calibration constraints for the radial trifocal tensor as well as the mixed trifocal tensor. We also show that these constraints can be used to develop fast minimal solvers for calibrated radial reconstruction.

Related work by Kim et al. [23]. Structure-from-Motion with the 1D radial camera model was previously considered by Kim et al. [23]. In [23] the authors presented a method for performing projective reconstruction with 1D radial cameras based on matrix factorization techniques, similar to previous work on projective-factorization for perspective cameras [50,2,10]. In a post-processing step, the method attempts to upgrade the reconstruction to metric by estimating the dual absolute quadric. However, their approach does not handle outlier measurements which limits the applicability on real image sequences. Additionally, due to the matrix factorization based approach, the method does not scale to larger image collections, e.g. the largest reconstruction presented in [23] has 189 3D points from 79 images. For comparison, in Section 4.4 we present 3D reconstructions from over a thousand images and more than 400k 3D points.

[6] also known as the Demazure constraints [12].

2 Calibrated Radial Trifocal Tensors

In this section we will present two new minimal solvers for calibrated radial trifocal tensors. These will be used for initializing our incremental Structure-from-Motion pipeline in Section 3. Next we show that there exists one additional internal constraint for each of the two tensors we consider; the purely radial trifocal tensor with intersecting principal axes, and the mixed trifocal tensor with one central camera and two radial cameras in general position. In the supplementary material we also discuss the third case considered by Thirthala and Pollefeys [48] where the scene is planar.

2.1 Intersecting Principal Axes

First we consider the case of intersecting principal axes. Choosing the world-coordinate frame such that the point of intersection is the origin, then the 2×4 camera matrices will be of the form, $P_k = \begin{bmatrix} A_k & \mathbf{0} \end{bmatrix}$, $k = 1, 2, 3$, $A_k \in \mathbb{R}^{2\times 3}$. The projection equations $\lambda \mathbf{x} = A_1 \mathbf{X}$, $\lambda' \mathbf{x}' = A_2 \mathbf{X}$ and $\lambda'' \mathbf{x}'' = A_3 \mathbf{X}$ can be rewritten

$$\begin{bmatrix} A_1 & \mathbf{x} & 0 & 0 \\ A_2 & 0 & \mathbf{x}' & 0 \\ A_3 & 0 & 0 & \mathbf{x}'' \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ -\lambda \\ -\lambda' \\ -\lambda'' \end{pmatrix} = 0. \tag{2}$$

This 6×6 matrix must thus be rank deficient and its determinant yields an equation which depend on the image points, $\sum_{i,j,k} T_{ijk}\, \mathbf{x}_i \mathbf{x}'_j \mathbf{x}''_k = 0$ where $\mathbf{x}_i$ denotes the ith image coordinate. The coefficients T_{ijk} can be interpreted as the $2 \times 2 \times 2$ multi-focal tensor [49] corresponding to this camera configuration. This is the *radial trifocal tensor* from [48]. In the uncalibrated setting this camera configuration has $3 \cdot (2 \cdot 3 - 1) - (3 \cdot 3 - 1) = 7$ degrees of freedom. Since the corresponding multi-focal tensor is a homogeneous $2 \times 2 \times 2$ tensor (which also has 7 degrees of freedom), the radial trifocal tensor does not have any internal constraint, as was also stated in [48].

Now if we consider the calibrated setting we require each matrix A_k to have orthonormal rows, i.e. $P_k = \begin{bmatrix} R_k & \mathbf{0} \end{bmatrix}$ where $R_k R_k^T = I_2$, $R_k \in \mathbb{R}^{2\times 3}$. In this case each camera only has 3 degrees of freedom and similarly the gauge freedom in the coordinate system is also reduced to 3 (since the projections are scale invariant) resulting in $3 \cdot 3 - 3 = 6$ degrees of freedom. This means that there must exist $7 - 6 = 1$ internal constraint on the corresponding trifocal tensor.

Internal Constraint for Calibration. Using techniques from numerical linear algebra we found that the internal constraint is a homogeneous quartic polynomial in the tensor elements. See the supplementary material for details on the constraint and how we found it. We have verified the validity of the constraint both empirically and symbolically using computer algebra software.

Estimation from Minimal Point Sets. As shown above, each triplet correspondence $(\mathbf{x}, \mathbf{x}', \mathbf{x}'')$ in the images yields one linear constraint on the elements of the radial trifocal tensor (see also [48]). To get a minimal problem we therefore

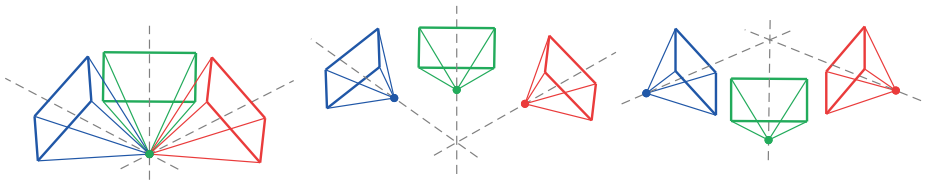

Fig. 2. The radial trifocal tensor describes three views with intersecting principal axes, e.g. from pure rotation (*Left*), panoramic motion (*Middle*) and orbital motion (*Right*).

need we need six triplet correspondences in total. From the six correspondences we can then find the two-dimensional linear subspace of possible $2 \times 2 \times 2$ tensors that satisfy the trifocal constraints, i.e.

$$T = \alpha_1 N_1 + \alpha_2 N_2, \tag{3}$$

where N_1 and N_2 are basis vectors to the nullspace. Since the tensor is homogeneous we can fix the scale by setting $\alpha_2 = 1$. To solve for the remaining unknown we insert (3) into the internal constraint from the previous section, yielding a single univariate quartic polynomial in α_1 that can be solved in closed form. In Section 4.1 we evaluate the proposed minimal solver.

2.2 Mixed Trifocal Tensor

Now we consider heterogeneous camera setups with both radial and pinhole cameras. The different minimal problems for heterogeneous camera setups were listed in Kozuka and Sato [24], though only in the projective setting. For the trifocal case there are two possibilities: 1) one pinhole and two radial cameras, 2) two pinhole and one radial camera. The second case becomes trivial since the minimal problem decouples into independent relative pose estimation between the pinhole cameras followed by pose estimation of the radial camera.

One Pinhole and Two Radial. The minimal solution for this camera case was first presented in [48] in the uncalibrated setting. In this case there are $11 + 7 + 7 - 15 = 10$ degrees of freedom[7]. Since the corresponding tensor is a homogeneous $3 \times 2 \times 2$ tensor with 11 degrees of freedom, there exist a single internal constraint. This constraint was derived in [48] and is a degree 6 polynomial in the tensor.

In the calibrated setting we have 6 d.o.f. in the calibrated pinhole camera, 5 in each of the calibrated 1D radial cameras and the coordinate system has 7 d.o.f., yielding $6 + 5 + 5 - 7 = 9$ degrees of freedom. Thus there must exist one additional internal constraint in the case of calibrated cameras. Similarly to Section 2.1 we used numerical techniques to recover the internal constraint. For this case it was more difficult to recover the constraint, both due to its higher degree, and having to consider the multiples of the original internal constraint from [48]. The internal constraint is a homogeneous degree 8 polynomial in the

[7] The projective coordinate system has 15 d.o.f.

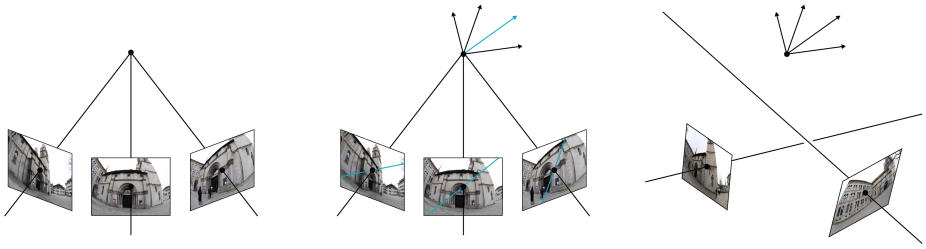

Fig. 3. Initialization for Structure-from-Motion with 1D radial cameras. *Left:* First we estimate a calibrated radial tensor describing the relative motion of three cameras with intersecting principal axes. *Middle:* Intersecting the backprojected feature correspondences of the three cameras we synthesize the image of a central camera with the intersection point as the projection center. *Right:* Finally we estimate a mixed trifocal tensor describing the relative motion of the synthesized central camera and two additional radial cameras.

elements of the tensor. For space reasons we do not print the full polynomial here (it has 3357 monomials). See the supplementary material for more details.

Estimation from Minimal Point Sets Each point correspondence yields a single linear constraint on the 12 elements of the mixed trifocal tensor. From the minimal sample of nine point correspondences we get a three dimensional nullspace where the tensor must lie, $T = \alpha_1 N_1 + \alpha_2 N_2 + \alpha_3 N_3$. Fixing $\alpha_3 = 1$ and inserting into the two internal constraints we get two polynomials in two unknowns of degree 6 and 8. Empirically we found that the coefficients of the two polynomials are completely generic which means that we have 48 solutions in general. Note that in practice many of these solutions end up being complex and only a small subset needs to be verified in the end. Using the generator from Larsson et al. [27] we created a Groebner basis solver for this polynomial system, but other techniques such as resultants could have been used as well.

3 Calibration-free Structure-from-Motion

In this section we present our incremental pipeline for Structure-from-Motion based on the 1D radial camera model (see Section 1.2). We base our method on the incremental SfM pipeline COLMAP [42]. The main steps in our pipeline are: Initialization (Section 3.1), Triangulation (Section 3.2), Camera Resectioning (Section 3.3) and Bundle Adjustment (Section 3.4). The main difference to traditional SfM frameworks is that each point-observation now only gives a single constraint on the reconstruction instead of two. This makes the geometric estimation problems significantly harder, e.g. 3D points require at least three views to triangulate. The benefit of this camera model is that we can perform reconstructions which are invariant to focal length or radial distortion. Note that at no point in our reconstruction pipeline do we estimate these parameters. We

only make the assumption of square pixels and centered principal point. The next sections detail the different parts of our framework.

3.1 Initialization

Initializing Structure-from-Motion is significantly harder for the 1D radial camera model compared to normal pinhole-like camera models. Without additional assumptions on the camera motion, the first constraints on the reconstruction appear in four views, i.e. it is (in general) impossible to estimate the structure and motion from only two or three views. The projective four-view case was investigated by Thirthala and Pollefeys in [48], but due to the high number of points required (15 for the linear solver presented in [48]), it is not useful in practice where we need to deal with outlier-contaminated data.

Now we present our approach for finding the initial reconstruction for the incremental Structure-from-Motion pipeline. It is based on the assumption that we can find three images where the principal axes are (approximately) intersecting (see Figure 2). Note that while a purely rotating camera satisfies this assumption, intersecting principal axes is a weaker constraint since the camera centers are not required to coincide. This also covers the spherical type of motion common in handheld panoramic image capture (see e.g. [52,46]) as well as orbital motion. This camera configuration is also common in photo collections where the cameras are often pointed towards some object of interest. The initialization consists of three stages and is performed using a combination of the minimal solvers described in Section 2.1 and 2.2. See Figure 3 for an overview.

a) Estimate Calibrated Radial Trifocal Tensor. Using the 6 point minimal solver from Section 2.1 in a RANSAC framework [31] we estimate a calibrated radial trifocal tensor for the first three images (which we assume have approximately intersecting principal axes). In Section 3.5 we present a simple heuristic we use for finding such image triplets in an image collection and in Section 4.3 evaluate the quality of the estimated camera poses on real images.

b) Create Synthetic Central Camera. From the three images with intersecting principal axes it is not possible to triangulate any 3D points. Each 2D observation backprojects to a 3D plane which contain the 3D point as well as the principal axis of the camera. If we intersect all three backprojected planes, the intersection will contain both the true 3D point and the intersection point of the three principal axes, and thus also the entire line between them. Thus we can only determine the direction towards the 3D point from the principal axes' intersection point. The idea is now that we can interpret these directions as the viewing rays from a central camera with projection center at the intersection point. Note that this automatically becomes a *calibrated* central image, since the directions were triangulated in the coordinate system defined by the calibrated radial trifocal tensor from the previous step. Note that we only triangulate the directions of the sparse set of correspondences we have and not generate a full synthetic image. The idea of generating synthetic pinhole images

from the radial trifocal tensor was also used in [36] to create undistorted images from three views.

c) Estimate Calibrated Mixed Trifocal Tensor. Finally we use the 9 point solver from Section 2.2 in RANSAC [31] to estimate the calibrated mixed trifocal tensor between the synthetic central camera and two additional views (which are modeled as radial cameras and can be in general position). Once we have a reconstruction with the five radial cameras in the same coordinate system we perform bundle adjustment (Section 3.4). For the refinement we remove the constraint that the first three views have intersecting principal axes.

3.2 Triangulation

Each 2D-3D correspondence yields a single constraint,

$$(-y, x) \begin{bmatrix} \mathbf{r}_1^T & t_1 \\ \mathbf{r}_2^T & t_2 \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} = 0 \tag{4}$$

Geometrically this can be interpreted as restricting the 3D point $\mathbf{X}$ to lie on the plane $\mathbf{n}^T\mathbf{X} + d = 0$, where $\mathbf{n} = x\mathbf{r}_2 - y\mathbf{r}_1$ and $d = xt_2 - yt_1$. Given at least three correspondences (for cameras in general position) we can find the intersection point of the planes by solving the corresponding linear system of equations (possibly in a least squares sense for overconstrained problems). Note that the triangulation problem is minimal with three views which means that the triangulated point will always have zero reprojection error. Therefore it is not possible to determine if the matches used were correct or not. To avoid this problem we only triangulate points seen in at least four views. For pinhole cameras the same number of constraints is achieved from two views.

3.3 Camera Resectioning

Resectioning is the problem of estimating the camera pose given 2D-3D correspondences. For calibrated radial cameras each camera has five degrees of freedom and thus we require at least five correspondences for estimation. The minimal solver for this problem was proposed by Kukelova et al. [25], where it was used in a two-step approach for radial distortion estimation. Note that for the case where the principal point is not known, the 1D radial solver from [29] which also estimates principal point, could in principle be used as a drop-in replacement. However, this solver has significantly larger runtime and requires two additional correspondences. We did not use this solver and found that the method is stable for the principal point offsets observed in practice.

3.4 Bundle Adjustment

We measure the reprojection error as the orthogonal distance from the projected radial line to the 2D point correspondence, i.e. for a camera $[R\ \mathbf{t}] \in \mathbb{R}^{2\times4}$, 3D

point $\mathbf{X} \in \mathbb{R}^3$ and 2D-observation $\mathbf{x} \in \mathbb{R}^2$, we measure

$$\varepsilon = \left\| \left(\frac{\mathbf{n}\mathbf{n}^T}{\mathbf{n}^T\mathbf{n}} - I \right) \mathbf{x} \right\|, \quad \text{where} \quad \mathbf{n} = R\mathbf{X} + \mathbf{t} \tag{5}$$

For the Bundle-Adjustment step in our pipeline we minimize the squared reprojection errors using the Ceres Solver [1]. If we have multiple images from the same camera we also refine the principal point.

3.5 Implementation Details

We have implemented our Structure-from-Motion pipeline by extending the open-source framework COLMAP [42]. The trifocal tensors estimated by the solvers in Section 2.1 and 2.2 can be factorized into the respective camera matrices. To perform this factorization we use the methods from [39,17], see the supplementary material for more detail. The runtimes of the solvers are 3.6μs (radial trifocal) and 0.8 ms (mixed trifocal). In the synthetic experiments the solvers returned 3.09 / 4 and 8.76 / 48 real solutions in average.

Initialization Image Selection. The proposed initialization method (Section 3.1) requires three images with intersecting principal axes. These images can either be manually selected by the user, or we use a simple heuristic for finding suitable image triplets to initialize from. We restrict ourselves to the case where the camera is undergoing purely rotational motion. For normal Structure-from-Motion this is a degenerate case for initialization which is avoided. In [42] this is detected by checking if a homography fits the image pair. We use this to identify potential image triplets for initialization. For a triplet we can then geometrically verify if the image triplet has intersecting principal axes by fitting a radial trifocal tensor. With this simple heuristic we could find good initialization images for all datasets used in the evaluation in Section 4.4.

4 Experimental Evaluation

4.1 Solver Stability, Robustness and Runtime

In this section we evaluate the numerical stability of the two proposed minimal solvers. Figure 4 (Left) shows the $\log_{10}$-residuals for 10,000 synthetically generated instances. For the residuals we compute the ℓ_2-distance to the ground truth tensor after normalizing each tensor to unit length (i.e. $\|\text{vec}(T)\|_2 = 1$). In the experiment 0.03% (radial trifocal) and 4.25% (mixed trifocal) instances had errors larger than 10^{-8}. The mixed trifocal tensor is slightly less numerically stable and had a few failures as can be seen in the figure.

We also performed an experiment where we evaluate how the solutions for the radial trifocal tensor degrade as the assumption of intersecting principal axes is violated. We generate randomized synthetic scenes with three pinhole cameras looking at the origin from unit-distance. We then perturb the cameras by rotating each camera around a random axis with the camera center being fixed. Figure 4 (Right) shows how the rotation estimates from the radial trifocal tensor degrades as the rotation angle increases.

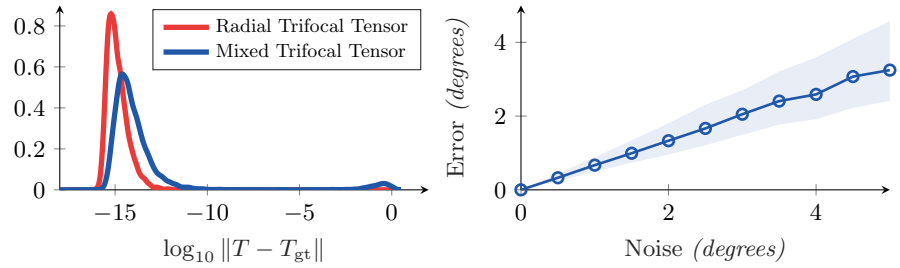

Fig. 4. *Left:* **Numerical stability**. The figure shows the distribution of the errors for 10,000 synthetically generated scenes. *Right:* **Stability to non-intersecting principal axes**. The graph shows the median errors in the relative rotations (in degrees) for the estimated calibrated radial trifocal tensor as the intersection constraint is violated. The shaded regions show the quartiles.

4.2 Comparison with Thirthala & Pollefeys [48]

In [48] the authors propose minimal solvers for estimating the radial trifocal tensor (intersecting principal axes) and the mixed radial trifocal tensor (perspective + two radial cameras) in the projective setting. These solvers do not enforce the additional constraint that ensures the tensors can be factorized into calibrated cameras (see Section 2.1 and Section 2.2). Since they use less constraints on the tensor itself, they also require one additional point correspondence. We generated synthetic scenes with varying levels of noise and compared how close the resulting cameras were to calibrated after factorizing the tensor and attempting metric upgrade (see supplementary material). Figure 5 shows the error in the rotation matrix constraint, $\|R_i R_i^T - I_2\|$, for varying levels of noise added to the image coordinates. Even for low noise levels the solvers from [48] yields solutions which are quite far from calibrated.

4.3 Evaluation of the Initialization on Real Data

The initialization pipeline we propose requires five images where three of them have close to intersecting principal axes. Intersecting principal axes can e.g. come from a purely rotating camera. In Section 3.5 we proposed a simple heuristic for finding this type of motion. To evaluate the initialization method we use the aforementioned method to find potential image triplets to initialize from. We select 1000 triplets from the *Lund Cathedral* dataset from [38]. For each triplet we estimate the trifocal tensor and compute the errors in the relative rotations w.r.t. the reconstruction provided in [38]. This is shown in Figure 6 (Left). Note that some of the selected triplets do not satisfy the assumption of intersecting principal axes, leading to large errors. For the 100 best triplets (highest inlier ratio) we try to further initialize by selecting the two additional images with the most matches. Figure 6 (Right) shows the distribution of the rotation errors for all five cameras used to initialize.

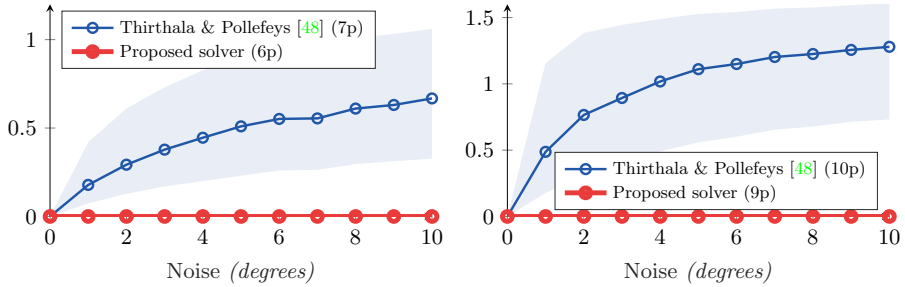

Fig. 5. Comparison with projective solvers from [48]. The graphs show the median error of the rotation matrix constraint (shadowed region shows quartiles) for 10,000 random instances. *Left:* Radial trifocal tensor. (Section 2.1) *Right:* Mixed trifocal tensor. (Section 2.2)

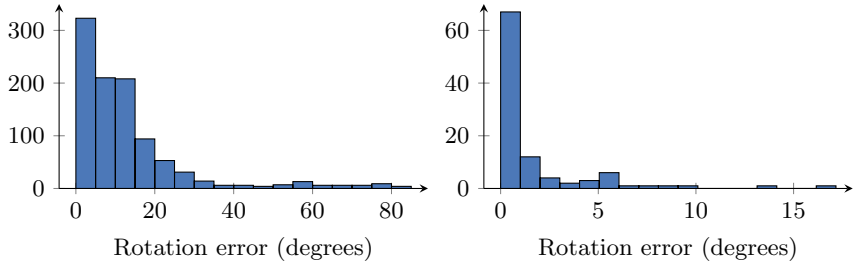

Fig. 6. Rotation errors (in degrees) for the initialization. *Left:* Three-view radial trifocal tensor estimation. *Right:* Full initialization pipeline (five views).

4.4 Structure-from-Motion Evaluation

For the quantitative evaluation of our SfM pipeline we consider five datasets from Olsson et al. [38]. We compare with vanilla COLMAP [42] using the ground truth camera intrinsics. Since we use a subset of the geometric constraints used in COLMAP, this provides an upper bound on the reconstruction quality we can achieve. The reconstructions from [38] are used as a pseudo-ground truth. Table 1 shows the camera pose errors and statistics after robustly aligning the coordinate systems to the ground truth. Since we only recover the camera position up to an unknown forward translation, the position error measures the distance from the ground truth camera center to the principal axis for both our and the baseline method [42]. The scales of the reconstructions from [38] were manually corrected. Image are considered correctly registered if the rotation error is below 5 degrees and it has at least 100 inliers. The table shows that we are able to achieve comparable reconstructions to the state-of-the-art pipeline [42] without knowing the intrinsic calibration. As expected, our reprojection errors are lower since they ignore the radial component of the errors. Figure 9 shows some qualitative results. For the *Spilled Blood* dataset the scene is highly symmetric and some images are being incorrectly registered to the wrong side of the building. Since

Table 1. Quantitative evaluation of the proposed Structure-from-Motion pipeline on the datasets from [38]. The errors are w.r.t. the reconstructions from [38]. Note that we only evaluate on the images that the method from [38] were able to register.

		Reg. Images		3D Points		$\varepsilon_{\text{rotation}}$ (deg)		$\varepsilon_{\text{position}}$ (m)		$\varepsilon_{\text{reproj}}$ (px)	
Dataset	Images	Our	[42]	Our	[42]	Our	[42]	Our	[42]	Our	[42]
Lund Cathedral	1208	99.6%	100%	422k	535k	0.93	0.39	0.929	0.180	0.287	0.578
Orebro Castle	761	100.0%	100%	197k	246k	0.15	0.10	0.387	0.089	0.276	0.532
San Marco	1498	100.0%	100%	293k	325k	0.50	0.29	0.614	0.140	0.443	0.751
Spilled Blood	781	80.3%	100%	285k	328k	0.72	0.26	0.231	0.134	0.409	0.571
Doge Palace	241	100.0%	100%	74k	93k	0.20	0.20	0.154	0.110	0.293	0.605

we use weaker projection constraints it is more difficult to disambiguate these incorrect matches.

Reconstruction with Severe Radial Distortion. Next we present qualitative results of our method applied to highly distorted images and show that we achieve accurate reconstruction without directly modeling the non-linear distortion. Figure 7 shows the reconstruction results from images taken with fisheye camera (from Camposeco et al. [5]) and Figure 8 from a GoPro camera (from Kukelova et al. [26]). In Figure 10 we show a reconstruction from 148 fisheye images. For comparison we also show the result of running COLMAP [42] without providing it intrinsic/distortion parameters, which fails to reconstruct the scene. This experiment shows that COLMAP [42] is not always able to converge to the correct intrinsic/distortion parameters during the bundle adjustment which motivates our method. More results can be found in the supplementary material.

5 Conclusions

We have presented an incremental Structure-from-Motion pipeline using the 1D radial camera model. Since the model is invariant to radial displacements in the image, we can directly perform reconstruction from heavily distorted images without any offline calibration or even explicitly modelling the type of distortion.

In this paper we deliberately focused on the most difficult setup where every camera is modeled as a radial camera, making the initialization more complex. In practice, for heterogeneous image collections it is possible to only model the cameras with high distortion effects as radial cameras and use a pinhole-like model for the others. This would allow for an easier and more general initialization procedure; either from two pinhole cameras, or from one pinhole camera together with two radial cameras (Section 2.2). In principle it is possible to use the reconstructions we recover to calibrate the cameras, e.g. using [25,30] for parametric distortion models or [5] for non-parametric. Even without this post-calibration step we have shown that we can achieve accurate 3D reconstruction.

Acknowledgements: Viktor Larsson was supported by an ETH Zurich Postdoctoral Fellowship.

Fig. 7. *Building dataset* from [5]. 60 images, 8984 3D-points, 0.38 px average reprojection error.

Fig. 8. *Rotunda dataset* from [26]. 62 images, 16292 3D-points, 0.41 px average reprojection error.

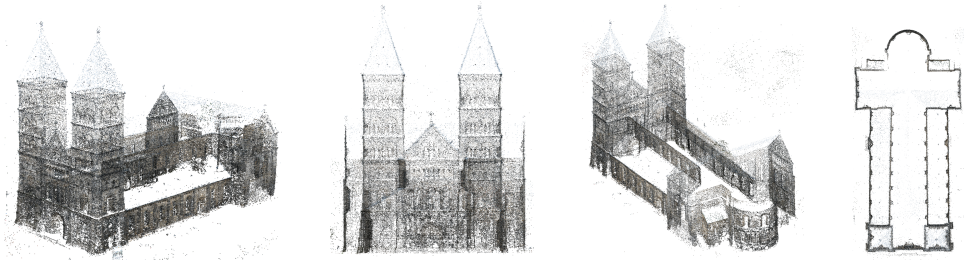

Fig. 9. Qualitative results for *Lund Cathedral* from Section 4.4. 1226 images, 422939 3D-points, 0.29 px average reprojection error.

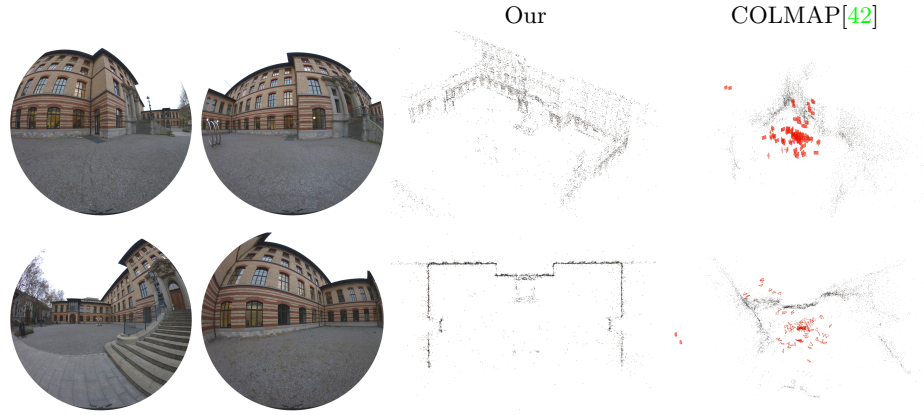

Fig. 10. *Fisheye Dataset*, 148 images, 14893 points, 0.51 px average reprojection error. Without known intrinsic/distortion parameters COLMAP fails to reconstruct the scene, while the proposed method successfully reconstructs it.

References

1. Agarwal, S., Mierle, K., Others: Ceres solver. http://ceres-solver.org 10
2. Angst, R., Zach, C., Pollefeys, M.: The generalized trace-norm and its application to structure-from-motion problems. In: International Conference on Computer Vision (ICCV) (2011) 4
3. Aranda, M., López-Nicolás, G., Sagüés, C.: Omnidirectional visual homing using the 1d trifocal tensor. In: International Conference on Robotics and Automation (ICRA) (2010) 4
4. Åström, K., Oskarsson, M.: Solutions and ambiguities of the structure and motion problem for 1d retinal vision. Journal of Mathematical Imaging and Vision (JMIV) (2000) 4
5. Camposeco, F., Sattler, T., Pollefeys, M.: Non-parametric structure-based calibration of radially symmetric cameras. In: International Conference on Computer Vision (ICCV) (2015) 3, 13, 14
6. Chandraker, M., Agarwal, S., Kahl, F., Nistér, D., Kriegman, D.: Autocalibration via rank-constrained estimation of the absolute quadric. In: Computer Vision and Pattern Recognition (CVPR) (2007) 1
7. Chatterjee, A., Govindu, V.M.: Robust relative rotation averaging. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2017) 1
8. Cui, H., Gao, X., Shen, S., Hu, Z.: Hsfm: Hybrid structure-from-motion. In: Computer Vision and Pattern Recognition (CVPR) (2017) 1
9. Cui, Z., Tan, P.: Global structure-from-motion by similarity averaging. In: International Conference on Computer Vision (ICCV) (2015) 1
10. Dai, Y., Li, H., He, M.: Projective multiview structure and motion from element-wise factorization. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2013) 4
11. Dellaert, F., Stroupe, A.W.: Linear 2d localization and mapping for single and multiple robot scenarios. In: International Conference on Robotics and Automation (ICRA) (2002) 4
12. Demazure, M.: Sur deux problemes de reconstruction. Tech. Rep. RR-0882, INRIA (Jul 1988), https://hal.inria.fr/inria-00075672 4
13. Duff, T., Kohn, K., Leykin, A., Pajdla, T.: Plmp-point-line minimal problems in complete multi-view visibility. In: International Conference on Computer Vision (ICCV) (2019) 4
14. Eriksson, A., Olsson, C., Kahl, F., Chin, T.J.: Rotation averaging and strong duality. In: Computer Vision and Pattern Recognition (CVPR) (2018) 1
15. Fabbri, R., Duff, T., Fan, H., Regan, M., de Pinho, D.d.C., Tsigaridas, E., Wrampler, C., Hauenstein, J., Kimia, B., Leykin, A., et al.: Trifocal relative pose from lines at points and its efficient solution. arXiv preprint arXiv:1903.09755 (2019) 4
16. Faugeras, O., Quan, L., Strum, P.: Self-calibration of a 1d projective camera and its application to the self-calibration of a 2d projective camera. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2000) 4
17. Hartley, R., Schaffalitzky, F.: Reconstruction from projections using grassmann tensors. International Journal of Computer Vision (IJCV) (2009) 10
18. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. International Journal of Computer Vision (IJCV) (2013) 1
19. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003) 1

20. Hong, J.H., Zach, C., Fitzgibbon, A., Cipolla, R.: Projective bundle adjustment from arbitrary initialization using the variable projection method. In: European Conference on Computer Vision (ECCV) (2016) 1
21. Hyeong Hong, J., Zach, C.: pose: Pseudo object space error for initialization-free bundle adjustment. In: Computer Vision and Pattern Recognition (CVPR) (2018) 1
22. Kileel, J.: Minimal problems for the calibrated trifocal variety. SIAM Journal on Applied Algebra and Geometry (2017) 4
23. Kim, J.H., Dai, Y., Li, H., Du, X., Kim, J.: Multi-view 3d reconstruction from uncalibrated radially-symmetric cameras. In: International Conference on Computer Vision (ICCV) (2013) 4
24. Kozuka, K., Sato, J.: Multiple view geometry for mixed dimensional cameras. In: International Conference on Computer Vision Theory and Applications (VISAPP) (2008) 6
25. Kukelova, Z., Bujnak, M., Pajdla, T.: Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In: International Conference on Computer Vision (ICCV) (2013) 3, 9, 13
26. Kukelova, Z., Heller, J., Bujnak, M., Fitzgibbon, A., Pajdla, T.: Efficient solution to the epipolar geometry for radially distorted cameras. In: International Conference on Computer Vision (ICCV) (2015) 13, 14
27. Larsson, V., Astrom, K., Oskarsson, M.: Efficient solvers for minimal problems by syzygy-based reduction. In: Computer Vision and Pattern Recognition (CVPR) (2017) 7
28. Larsson, V., Kukelova, Z., Zheng, Y.: Making minimal solvers for absolute pose estimation compact and robust. In: International Conference on Computer Vision (ICCV) (2017) 2
29. Larsson, V., Kukelova, Z., Zheng, Y.: Camera pose estimation with unknown principal point. In: Computer Vision and Pattern Recognition (CVPR) (2018) 2, 9
30. Larsson, V., Sattler, T., Kukelova, Z., Pollefeys, M.: Revisiting radial distortion absolute pose. In: International Conference on Computer Vision (ICCV) (2019) 2, 3, 13
31. Lebeda, K., Matas, J., Chum, O.: Fixing the Locally Optimized RANSAC. In: British Machine Vision Conference (BMVC) (2012) 8, 9
32. Leonardos, S., Tron, R., Daniilidis, K.: A metric parametrization for trifocal tensors with non-colinear pinholes. In: Computer Vision and Pattern Recognition (CVPR) (2015) 4
33. Locher, A., Havlena, M., Van Gool, L.: Progressive structure from motion. In: European Conference on Computer Vision (ECCV) (2018) 1
34. Magerand, L., Del Bue, A.: Revisiting projective structure for motion: A robust and efficient incremental solution. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2018) 1
35. Martyushev, E.: On some properties of calibrated trifocal tensors. Journal of Mathematical Imaging and Vision (JMIV) (2017) 4
36. Molana, R., Daniilidis, K.: A single-perspective novel panoramic view from radially distorted non-central images. In: British Machine Vision Conference (BMVC) (2007) 9
37. Moulon, P., Monasse, P., Marlet, R.: Global fusion of relative motions for robust, accurate and scalable structure from motion. In: International Conference on Computer Vision (ICCV) (2013) 1

38. Olsson, C., Enqvist, O.: Stable structure from motion for unordered image collections. In: Scandinavian Conference on Image Analysis (SCIA) (2011) 1, 11, 12, 13
39. Quan, L., Kanade, T.: Affine structure from line correspondences with uncalibrated affine cameras. Trans. Pattern Analysis and Machine Intelligence (PAMI) (1997) 4, 10
40. Sagues, C., Murillo, A., Guerrero, J.J., Goedemé, T., Tuytelaars, T., Van Gool, L.: Localization with omnidirectional images using the radial trifocal tensor. In: International Conference on Robotics and Automation (ICRA) (2006) 4
41. Sattler, T., Sweeney, C., Pollefeys, M.: On sampling focal length values to solve the absolute pose problem. In: European Conference on Computer Vision (ECCV) (2014) 2
42. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR) (2016) 1, 2, 7, 10, 12, 13, 14
43. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016) 2
44. Schops, T., Larsson, V., Pollefeys, M., Sattler, T.: Why having 10,000 parameters in your camera model is better than twelve. In: Computer Vision and Pattern Recognition (CVPR) (2020) 1
45. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. International Journal of Computer Vision (IJCV) (2008) 1
46. Sweeney, C., Holynski, A., Curless, B., Seitz, S.M.: Structure from motion for panorama-style videos. arXiv preprint arXiv:1906.03539 (2019) 8
47. Sweeney, C., Sattler, T., Hollerer, T., Turk, M., Pollefeys, M.: Optimizing the viewing graph for structure-from-motion. In: International Conference on Computer Vision (ICCV) (2015) 2
48. Thirthala, S., Pollefeys, M.: Radial multi-focal tensors. International Journal of Computer Vision (IJCV) **96**(2), 195–211 (2012) 3, 4, 5, 6, 8, 11, 12
49. Triggs, B.: Matching constraints and the joint image. In: International Conference on Computer Vision (ICCV) (1995) 3, 5
50. Triggs, B.: Factorization methods for projective structure and motion. In: Computer Vision and Pattern Recognition (CVPR) (1996) 4
51. Tsai, R.: A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. Journal on Robotics and Automation (1987) 2
52. Ventura, J.: Structure from motion on a sphere. In: European Conference on Computer Vision (ECCV) (2016) 8
53. Wu, C.: Towards linear-time incremental structure from motion. In: International Conference on 3D Vision (3DV) (2013) 1
54. Wu, C.: P3.5p: Pose estimation with unknown focal length. In: Computer Vision and Pattern Recognition (CVPR) (2015) 2
55. Zhang, Z., et al.: Flexible camera calibration by viewing a plane from unknown orientations. In: International Conference on Computer Vision (ICCV) (1999) 1
56. Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L.: Very large-scale global sfm by distributed motion averaging. In: Computer Vision and Pattern Recognition (CVPR) (2018) 1