Radial Distortion Invariant Factorization for Structure from Motion

José Pedro Iglesias[1]
[1]Department of Electrical Engineering
Chalmers University of Technology

Carl Olsson[1,2]
[2]Centre for Mathematical Sciences
Lund University

Abstract

Factorization methods are frequently used for structure from motion problems (SfM). In the presence of noise they are able to jointly estimate camera matrices and scene points in overdetermined settings, without the need for accurate initial solutions. While the early formulations were restricted to affine models, recent approaches have been show to work with pinhole cameras by minimizing object space errors.

In this paper we propose a factorization approach using the so called radial camera, which is invariant to radial distortion and changes in focal length. Assuming a known principal point our approach can reconstruct the 3D scene in settings with unknown and varying radial distortion and focal length. We show on both real and synthetic data that our approach outperforms state-of-the-art factorization methods under these conditions. [1]

1. Introduction

Factorization methods have been employed for SfM problems since the seminal work [30]. Suppose that M is a measurement matrix where each pair of rows correspond to the x- and y- coordinates of the observed 2D points in one image. The approach relies on the fact that the observed projections should form a rank 3 matrix. To remove noise they therefore solve

$$\min_{P,X} \|M - PX\|_F^2, \tag{1}$$

with matrices P having 3 columns and X having 3 rows, using the singular value decomposition (SVD). The interpretation of the result is that a pair of rows P_i in P correspond to an (affine) camera matrix projecting the 3D points in X into image i. The solution is ambiguous since $PX = PHH^{-1}X$ for any invertible H. In [30] a metric upgrade is found by selecting H so that PH consists of pairwise orthonormal rows giving an orthographic projection model. To allow scale changes [25] generalizes the approach to weak perspective cameras.

Since then a number of variations of the SfM problem has been tackled with factorization approaches. Strum and Triggs [28] used a measurement matrix consisting of homogeneous coordinates, scaled with (known) point depths, to extract 3×4 cameras matrices. Iterative approaches alternating estimation of the depths Λ with factorization was proposed in [10, 31]. To avoid solutions where many of the elements of Λ are zero, [22] derives linear constraints that permits a projectively correct reconstruction. Margerand *et al*. [21] eliminate the projective depths and show that the linear constraints can be transferred to the remaining parameters. A very recent method [16], factorizes a block matrix consisting of estimated pairwise fundamental matrices to solve SfM.

One reason for the popularity of the original (affine) factorization approaches was the availability of a closed form solution using the SVD. This is however only possible if all scene point are visible in all images. If this is not the case we are limited to iterative methods. For this purpose splitting methods have become popular because of their simplicity [3, 5, 18]. These allow regularization of the singular values when a proximal operator can be computed. On the other hand recent results [15, 23] have shown that they can give relatively inaccurate results due to slow convergence near the optimum. Convex formulations with the nuclear norm have also been considered [5, 4] but are in general too weak for SfM problems with noise [23, 15]. We note that a number of papers that give conditions which ensure that direct bilinear optimization has no local minimum (other than the global one) [1, 24, 8, 7]. The assumptions these papers make are however too restrictive for the SfM problem, where local minima are known to exist e.g. [2].

A recent series of papers by Hong *et al*. [11, 14, 13, 12] show that when the rank is known direct bilinear estimation of P and X can be made remarkably robust to local minima, when using the VarPro method. In [12] the pOSE formulation is introduced. This method uses a pinhole model and optimizes a trade-off between an object space error (OSE) and an affine term (which fixes the scale). It is shown to converge to the global minimum from random starting solution in the vast majority of cases.

[1]This work has been supported by the Swedish Research Council (grant 2018-05375), the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots), the Wallenberg AI, Autonomous Systems and Software Program (WASP)

One shortcoming of factorization methods is that they cannot handle radial distortion since this introduces a non-linear transformation of the measurements. In this paper we address this by optimizing over 1D radial camera projections [32] which are invariant to radial distortion and changes in focal length. We propose RpOSE, which optimizes a trade-off between a radial OSE and an affine term. The result is a formulation that enables simultaneous estimation of cameras and structure in over-determined settings with unknown and varying focal length and radial distortion. We further show that the pOSE and RpOSE models can be seen as local approximations of reprojection error, which opens up the possibility of iterative refinement.

Radial cameras have previously mostly been studied for minimal solvers. The radial constraint was first proposed in [32]. The multiple view geometry for these cameras was studied in [29]. Solvers for the absolute pose problem were presented in [19]. In a recent work [20] Larsson *et al*. present a sequential reconstruction pipeline using radial cameras. They show that accurate metric solutions can be constructed from unordered image collections with varying and unknown distortion parameters and focal lengths. The most closely related work to ours is [17] where a factorization approach was proposed. It is based on a convex relaxation using the nuclear norm and as such it is sensitive to noise. In addition it uses a splitting scheme which can converge slowly [15, 23]. Furthermore the formulation does not take into account that noisy point projections that end up near the principal point result in directional vectors that are very uncertain. Hence without a proper weighting these points can degrade the quality of the reconstruction.

In summary the main contributions of this paper are:

- A new pOSE formulation, named RpOSE, that can handle radially distorted images.
- We show that the new formulation can be robustly optimized using VarPro converging to the globally optimal solution in the vast majority of cases from random starting solutions.
- We show that the pOSE formulations can be seen as local approximations of reprojection error opening up the possibility of robustly solving the maximum likelihood formulation.

2. The 1D-radial Camera Model

2.1. Distortion Models

A common way of modeling radial distortion is through the so called division model [6]. If we assume that the coordinate system in the image has its origin in the distortion center we can write the projection model as

$$\lambda_{ij} \begin{bmatrix} m_{ij} \\ 1 + \kappa(m_{ij}) \end{bmatrix} = K_i [R_i \quad t_i] \begin{bmatrix} X_j \\ 1 \end{bmatrix}. \tag{2}$$

Here m_{ij} represents the observed image point (in regular 2D Cartesian coordinates) X_j is the 3D point and $K_i \begin{bmatrix} R_i & t_i \end{bmatrix}$ is the camera matrix. Throughout this paper we will assume that the distortion center is the principal point, that the skew is 0 and that the aspect ratio is 1. Therefore K is of the form $\mathrm{diag}(f, f, 1)$. The main difference to a pure pinhole camera is the polynomial $\kappa(m_{ij})$ that distorts the point locations by rescaling the coordinates (through division) based on the distance to a distortion center $(0, 0)$. The polynomial is usually of the form $\kappa(m_{ij}) = \sum_l k_l r^{2l}$, where $r^2 = \|m_{ij}\|^2$.

Under no radial distortion, we have that $\kappa(m_{ij}) \equiv 0$ and (2) simplifies to the perspective projection of a 3D point into an image. Note that in an uncalibrated SfM problem, only the image measurements m_{ij} in (2) are known, and all other variables/parameters are unknown.

2.2. The 1D Radial Camera Model

The main benefit of the division model presented in the previous section is that it makes it possible to estimate radial distortion using linear methods for a class of minimal absolute [19] and relative pose problems [6]. On the other hand, it introduces additional non-linearities to the projection that are difficult to handle in over-determined settings. In this paper we circumvent these issues by using the 1D radial camera model [29]. This model uses a weaker projection where each 3D point gives an image line going through the projection center. The interpretation is that this line contains the true point projection. Under the model described in the previous section it is easy to see that radial distortion moves points along such lines thereby making the radial camera invariant. While the projection becomes weaker, basically dropping location information in one direction, the non-linear effects of the distortion also disappear.

When the distortion center and principal point are both in the origin, the 1D-radial projection is given by the two first coordinates of (2) and therefore becomes

$$\lambda_{ij} m_{ij} = P_i \begin{bmatrix} X_j \\ 1 \end{bmatrix}, \tag{3}$$

where $P_i = \mathrm{diag}(f_i, f_i) \begin{bmatrix} r_i^1 & t_i^1 \\ r_i^2 & t_i^2 \end{bmatrix}$ only depends on the first two rows of $[R \quad t]$. Note that, under these assumptions, the explicit dependence of the distortion parameters disappear, and radial distortion is now implicitly modeled by λ_{ij}. Furthermore, the camera matrix P_i is of size 2×4, giving a significant reduction of the number of parameters to be estimated.

We will write $z_{ij} := P_i \begin{bmatrix} X_j \\ 1 \end{bmatrix}$ and interpret this as a directional vector of the reprojected line that contains the measured point m_{ij}. By Z we mean the block matrix formed by putting the vector z_{ij} in block (i, j). This matrix

can be factorized to estimate $P \in \mathbb{R}^{2F\times 4}$ and $X \in \mathbb{R}^{3\times N}$, where

$$P = [P_1^T \quad \dots \quad P_F^T]^T \quad \text{and} \quad X = [X_1 \quad \dots \quad X_N]. \tag{4}$$

2.3. The ML-estimate

We now suppose that our measurement m_{ij} is a noisy observation of a point from the line with directional vector z_{ij}. Under the Gaussian noise assumption the maximum likelihood estimation is to minimize the distance between the line and the measured point leading to the problem

$$\begin{aligned} \min \sum_{ij} \left\| \left(I - \frac{z_{ij} z_{ij}^T}{\|z_{ij}\|^2} \right) m_{ij} \right\|^2 . \\ \text{s.t.} \quad Z = PX. \end{aligned} \tag{5}$$

The objective function was observed to work well for large scale reconstruction in [20]. This work did however use minimal solvers for sequentially building a good starting solution to be refined with bundle adjustment. Here we are aiming to directly optimize in an over-determined setting from random starting solutions.

The matrix $\mathcal{P} := \left(I - \frac{z_{ij} z_{ij}^T}{\|z_{ij}\|^2} \right)$ represents projection onto the plane orthogonal to the line with directional vector z_{ij}. This matrix is a symmetric projection which makes it possible to simplify the terms of the objective function in (5) to

$$\|\mathcal{P} m_{ij}\|^2 = m_{ij}^T \mathcal{P} m_{ij} = m_{ij}^T \left(\frac{\|z_{ij}\|^2 I - z_{ij} z_{ij}^T}{\|z_{ij}\|^2} \right) m_{ij}. \tag{6}$$

Because of symmetry it is not difficult to see that we can swap places between z_{ij} and m_{ij} in the expression $m_{ij}^T \left(\|z_{ij}\|^2 I - z_{ij} z_{ij}^T \right) m_{ij}$. Furthermore, the matrix $\|m_{ij}\|^2 I - m_{ij} m_{ij}^T$ is of rank 1 and can be written as an outer product $\bar{m}_{ij} \bar{m}_{ij}^T$, where $\bar{m}_{ij}$ is a vector that is perpendicular to m_{ij} and has $\|\bar{m}_{ij}\| = \|m_{ij}\|$. Therefore we can write the terms of the ML estimate as

$$\ell_{\text{ML}} = \left\| \bar{m}_{ij}^T \frac{z_{ij}}{\|z_{ij}\|} \right\|^2 . \tag{7}$$

Since this projection is independent of which directional vector is used it is clear that the objective is scale invariant with respect to Z. Indeed the inner residual $\bar{m}_{ij}^T \frac{z_{ij}}{\|z_{ij}\|}$ is non-linear in the unknown Z. While non-linear factorization methods based on the VarPro algorithm have been proposed [27, 13] it has been observed in the context of SfM that linear residuals give formulations that are far more resilient to local minima [12]. In the following section we present such a formulation.

2.4. A pOSE model

In [12] the non-linear residuals where avoided by switching from an image error to an object space error (OSE) by effectively multiplying the residual with its denominator. The same applies in our setting. Removing the dependence on z_{ij} from the denominator of $\bar{m}_{ij}^T \frac{z_{ij}}{\|z_{ij}\|}$ clearly makes the inner residual linear and turns (7) into a linear least squares problem (albeit with a trivial solution).

One way to achieve linearity is to simply replace the term $\|z_{ij}\|$ in the denominator by 1. However since the length of z_{ij} and m_{ij} have some correlation for typical scenes we instead use $\|m_{ij}\|$, which gives the radial object space error (ROSE)

$$\ell_{\text{ROSE}} = \left\| \frac{\bar{m}_{ij}^T}{\|m_{ij}\|} z_{ij} \right\|^2 . \tag{8}$$

Figure 1 shows level-sets of the ℓ_{ROSE} (green dashed lines) and the ℓ_{ML} (blue dashed lines). For the ML estimate the coordinates of directional vector z_{ij} should be in the cone (spanned by the blue dashed lines) to achieve an error less than ϵ. Approximating with a constant denominator removes the dependence on the distance to the principal point (middle of the image). The ℓ_{ROSE} term therefore measures the perpendicular distance between z_{ij} and the line containing m_{ij}.

Note, that with this the approximation it appears as if z_{ij} is a (theoretical) point projection that should be optimized to be close to a measured line with directional vector m_{ij}. This is however not the case since z_{ij} is not a proper pinhole projection, but only the result of multiplying a 3D point with a 2×4 matrix. It therefore still represents the directional vector of a line known to contain the true projection.

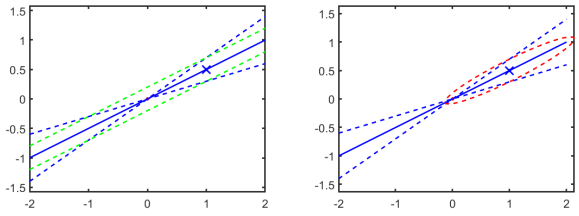

Figure 1. Level-sets of the ℓ_{ROSE} (green dashed lines) and the ℓ_{ML} (blue dashed lines) and the ℓ_{RpOSE}, with the point m_{ij} marked with a cross.

It is clear that resulting term ℓ_{ROSE} is not scale invariant. To prevent the solution from collapsing to zero we therefore add the affine term

$$\ell_{\text{Affine}} = \|m_{ij} - z_{ij}\|^2 \tag{9}$$

and use a convex combination of the two terms giving the objective function $\ell_{RpOSE} = (1-\eta)\ell_{\text{ROSE}} + \eta \ell_{\text{Affine}}$. The affine term penalizes deviations from m_{ij} in all directions.

It is not difficult to see that it can be divided into a term that measures deviations from the line containing m_{ij} and a perpendicular term that measures deviations from m_{ij} along the same line. The exact expression becomes

$$\ell_{\text{RpOSE}} = \ell_{\text{ROSE}} + \eta \left\| m_{ij} - \frac{m_{ij}^T z_{ij}}{\|m_{ij}\|^2} m_{ij} \right\|^2 . \qquad (10)$$

The level-sets of this function are ellipses with half axes that are parallel and perpendicular to to the line containing m_{ij}. Figure 1 shows one such ellipse (red dashed curve). When η is reduced this ellipse is extended in the direction of the line containing m_{ij}.

The resulting objective function is a least squares problem in the unknowns z_{ij}. We can therefore write it in matrix form as $\|\mathcal{A}Z - b\|^2$, where $\mathcal{A}$ is some linear operator on Z.

Since our goal is a factorization of Z into cameras P and 3D points X we solve the problem

$$\underset{P,X}{\text{minimize}} \quad \left\| \mathcal{A}\left(P \begin{bmatrix} X \\ 1 \end{bmatrix} \right) - b \right\|^2 \qquad (11)$$

For this purpose we employ the VarPro formulation [11]. This method uses the fact that linearity of $\mathcal{A}$ makes it possible to marginalize over one of the matrices X and P and express its optimum as a function of the other. In contrast to Levenberg-Marquart this enables us to avoid the use of a dampening term for the eliminated variable which empirically has been shown to greatly reduce its sensitivity to local minima. As shown by Hong *et al.* [12], the use of VarPro combined with the affine term in the loss results in method with a very wide convergence basin. Algorithm 1 outlines the method that we use for solving (11), see [11] for more details.

The affine term restricts the z_{ij} to a neighborhood around the point m_{ij}. To limit this influence and allow solutions where z_{ij} and m_{ij} are roughly parallel but have different scales the η should typically be selected low. Figure 2 shows the shape of the level curves for different η as well as the resulting reconstructions obtained from factorization with this model. (Note that the obtained reconstruction is uncalibrated. Therefore we have registered the resulting point cloud to a ground truth point cloud to resolve the inherent projective ambiguity).

2.5. Evaluation

Next we empirically evaluate our RpOSE model. We first compare to a standard uncalibrated 1D-radial bundle adjustment implementation for the purpose of illustrating the increased convergence basin of our method. We then compare to the state-of-the-art factorization method in [17]. This method is based on a convex relaxation of the problem and therefore independent of initialization. While our

Algorithm 1: VarPro for solving (11)

Select the inputs η, and randomly initialize $P^{(0)}$;
Set up $\mathcal{A}$ and b;
Compute $X^{(0)}$ by minimizing (11) with $P = P^{(0)}$ fixed;
while *true* **do**
  Compute the Jacobians
  $J_P = A\left[X^T \otimes \mathcal{I}, \mathbb{1}^T \otimes \mathcal{I}\right]$; $J_X = A(\mathcal{I} \otimes P)$;
  and the residuals $r = A\text{vec}\left(P \begin{bmatrix} X \\ 1 \end{bmatrix}\right) - b$;
  Compute P_{new} and X_{new} from J_P, J_X, and r as
  $P_{\text{new}} = P + \Delta P$ and $X_{\text{new}} = P + \Delta X$, with
  $\Delta P = (J_P^T(\mathcal{I} - J_X J_X^\dagger)J_P + \lambda\mathcal{I})^{-1} J_P^T r$, and
  $\Delta X = -J_X^\dagger(r + J_P \Delta P)$;
  Evaluate the loss $\ell_{\text{new}} = \left\| \mathcal{A}\left(P_{\text{new}} \begin{bmatrix} X_{\text{new}} \\ 1 \end{bmatrix}\right) - b \right\|^2$;
  if $\ell_{new} < \ell_{best}$ **then**
    $\ell_{\text{best}} = \ell_{\text{new}}$;
    $P \leftarrow P_{\text{new}}$; and $X \leftarrow X_{\text{new}}$;
  end
  if *stopping criterion* **then**
    break;
  end
end

VarPro formulation will be part of a larger stratified reconstruction pipeline here we only consider properties of the RpOSE model. Note that this is an uncalibrated formulation. For qualitative visual evaluation of the results we therefore register the obtained reconstructions to a ground truth point cloud to remove the unknown projective ambiguity. In this way we can fairly evaluate the uncalibrated models without the results being affected by subsequent upgrade steps that are often sub-optimal. For evaluation of the whole pipeline we refer the reader to Section 4.

Basin of convergence In [12] it is shown that pOSE has a wide basin of convergence when solved with VarPro, and can be initialized with random camera matrices. Here we show that this property is retained in the proposed bilinear factorization for 1D radial cameras. For this purpose, we use 3 datasets [20] that aim to represent different camera models, motions, scenes, and percentages of available measurements. Each dataset is run several times for different values of η and the average 3D reconstruction error is used as evaluation metric, after projective registration to the 3D ground-truth. In each instance, the starting solution is constructed by sampling the elements of the camera matrices from a Gaussian distribution with mean 0 and variance ϵ. An example of the obtained reconstruction for the Door dataset is in Figure 2. The trade-off between the quality of the reconstruction and the sensitivity to local minima introduced by the choice of η is shown in Figure 3. In Section 3 we address a strategy to reduce the sensitivity of the method to the initial choice of η.

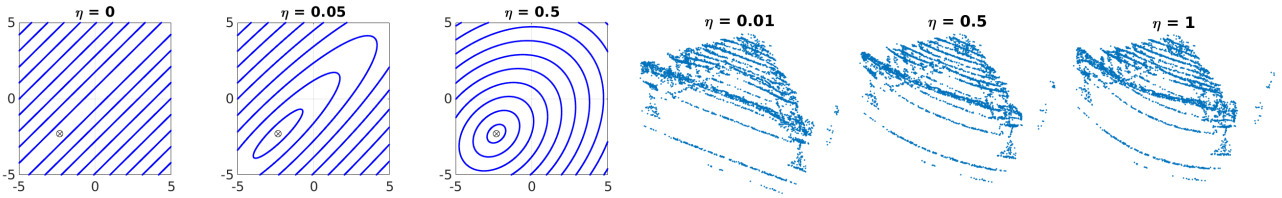

Figure 2. (Left) Level curves of the pOSE for 1-D radial cameras. (Right) Top view of Door datset.

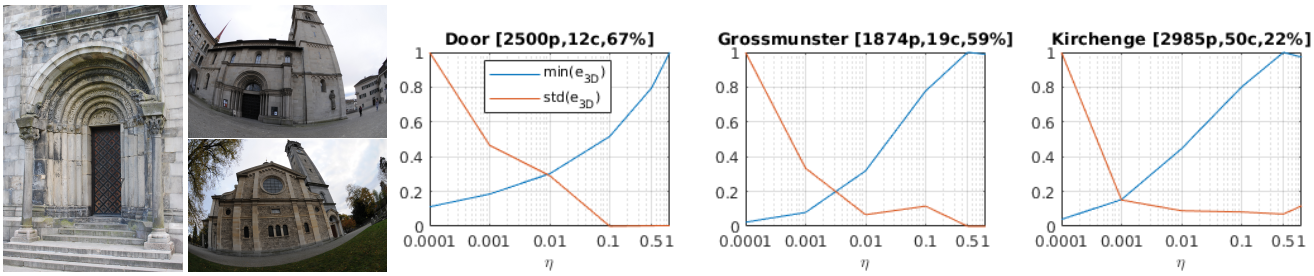

Figure 3. (Left) Examples of the 3 datasets (Door: 2500points, 12 viewpoints, 66% of available data; Grossmunster: 1874points, 19 viewpoints, 59% available data; and Kirchenge: 2985 points, 50 viewpoints, 22% available data) used for evaluating the basin of convergence of RpOSE solved with VarPro. (Right) Plots of the best errors (blue) and standard deviation of the error of the solutions (red) obtained in all instances for each value of η when initializing RpOSE with random camera matrices. The values are normalized by their maximum for better visualization. There is a clear trade-off between stability and accuracy of the algorithm, defined by the parameter η. A strategy to reduce the sensibility of the algorithm to the choice of η is proposed in Section 3.

Comparison with state-of-the-art factorization Factorization with a 1D-radial camera model has previously only been addressed by Kim *et al.* [17]. In this section we compare our proposed method to their approach. The inputs to the two methods are the same - image measurements and assumed principal point. For this experiment we generate a synthetic dataset using the 1000 ground-truth 3D points and 12 camera matrices from the Door dataset and around 66% of the measurements are available. The 3D points are projected to the normalized image plane, from which 3 datasets with different levels of radial distortion are generated. The three levels correspond three different distortion models. 1) Regular pinhole camera with no distortion using the intrinsic camera parameters from the original door dataset. 2) Polynomial model where a point x in the normalized image plane are distorted by multiplication with $1 + \sum_i k_i \|x\|^{2i}$, using the parameters k_i described in [20]. 3) Fisheye model, using the parameters from the cameras of the low-resolution datasets in [26]. Noise sampled from a Gaussian distribution with mean 0 and standard deviation ω is added to the image measurements over several problem instances, and we evaluate the performance of the methods by computing the normalized average 3D error $\frac{\|X - X_{\text{gt}}\|_F}{\|X_{\text{gt}}\|}$, the average 2D angle error $\sum_{ij} \text{acos}\left(\frac{m_{ij}^T z_{ij}}{\|m_{ij}\|\|z_{ij}\|}\right)$ and runtime. Figure 4 shows the results. To illustrate the effect of the radial distortion when no measures are taken to compensate for it we also show the results obtained when using the pOSE method for regular pinhole cameras, with $\eta = 0.05$. Both RpOSE and pOSE are initialized with random camera matrices with zeros translations, while ALM is initialized as proposed by the authors. A stopping criteria of 5×10^{-5} is defined for ALM to avoid high runtimes.

3. Connection between pOSE and ML

In this section we observe that the OSE errors (both ours and that of [12]) are closely connected to the ML in the sense that the OSE residuals can be seen as a linearization of the non-linear reprojection errors at a certain point z_{ij}. This offers some explanation as to why pOSE error works so well as a starting solution for bundle adjustment in [12].

Here we restrict ourselves to linearization of the 1D-radial projection. In the supplementary material we show that the same holds for regular pinhole model of [12]. Taking the first order Taylor expansion of the inner term at an arbitrary point v we see that (7) can be approximated by

$$\ell_{\text{ROSE}} := \left\| \left(\bar{m}_{ij}^T \frac{v}{\|v\|} \right) + J(v)^T (z_{ij} - v) \right\|^2, \qquad (12)$$

where the Jacobian is given by $J(v) = \frac{\bar{m}_{ij}}{\|v\|} - \frac{\bar{m}_{ij}^T v}{\|v\|^3} v$. The connection to the OSE becomes clear when noting that linearization around $v = m_{ij}$ results in $\ell_{\text{ROSE}} = \left\| \frac{\bar{m}_{ij}^T}{\|m_{ij}\|} z_{ij} \right\|^2$. The affine term can be seen as a dampening term, restricting the unknown z_{ij} to a neighborhood around v. When adding it we get a more general approximation

$$\ell_{RpOSE} = (1 - \eta)\ell_{\text{ROSE}} + \eta \ell_{\text{affine}}, \qquad (13)$$

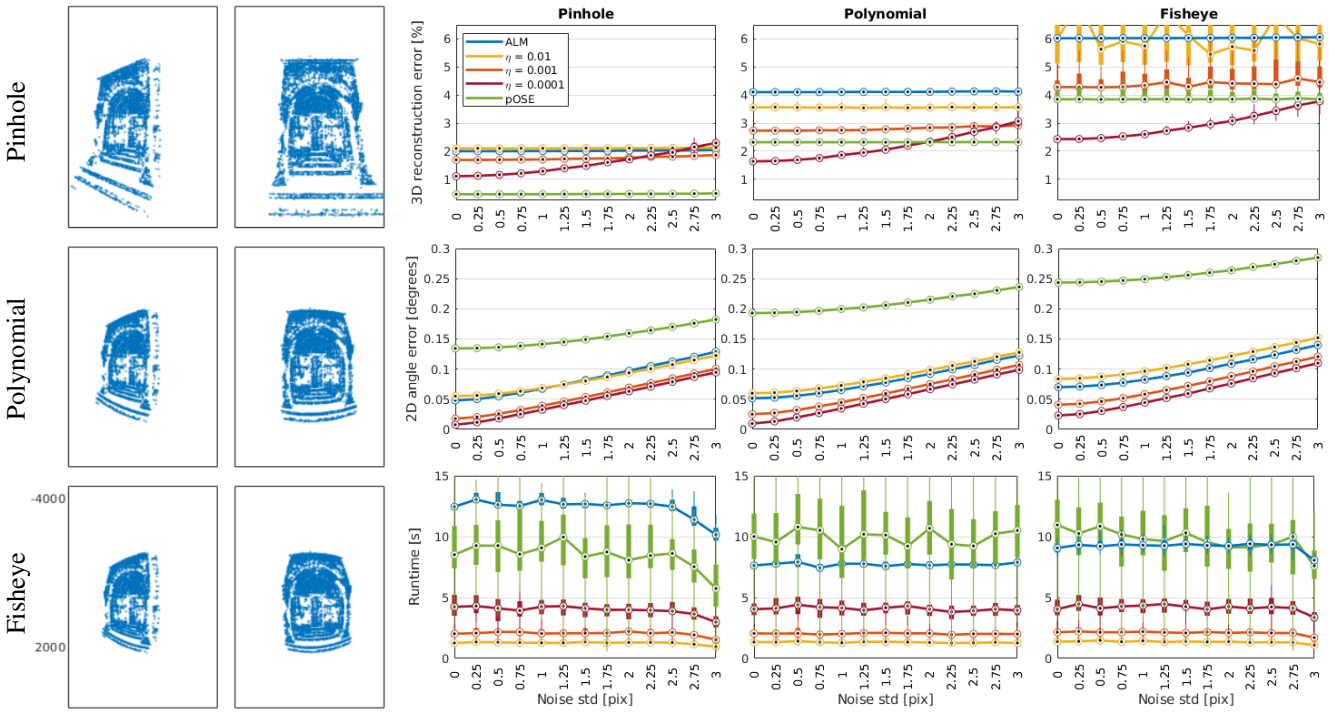

Figure 4. (Left) Examples of the synthetic datasets used in the experiments. (Right) Plots of 3D reconstruction error, 2D angle error and runtime of RpOSE vs ALM [17] and pOSE [12], as a function of the noise level added to the image points.

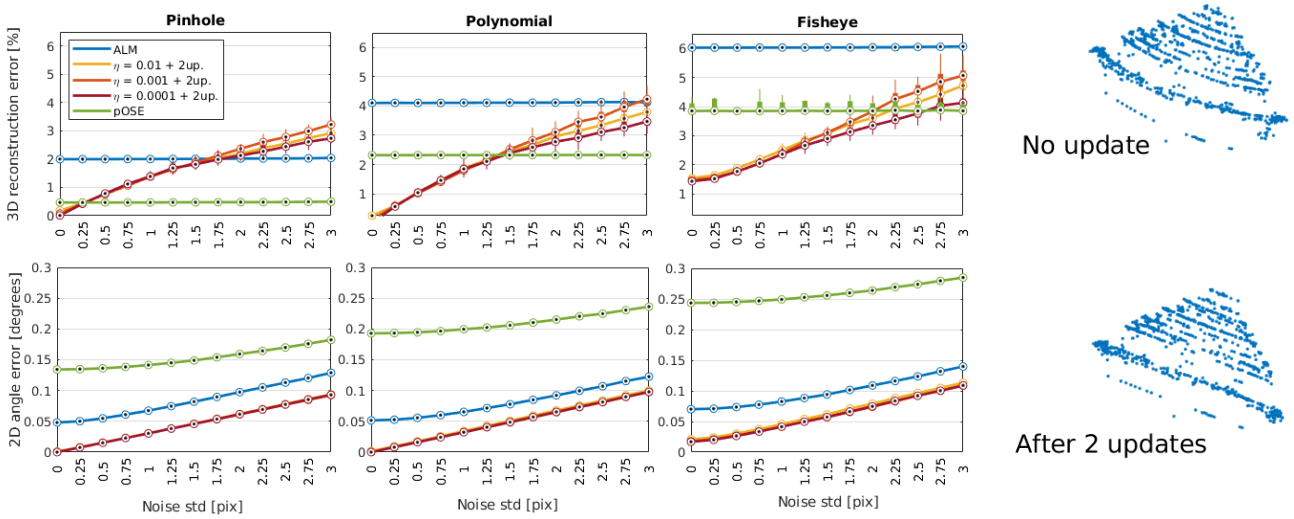

Figure 5. (Left) Reconstruction and angle errors obtained with RpOSE after 2 refinement steps, compared with the results from ALM and pOSE. The relevance of the initial choice of η is significantly reduced by refining the solution from the previous linearization, given that the methods initialized with different values of η achieved similar performance. (Right) Example of the effect of the refinement on the reconstruction itself.

where $\ell_{\text{affine}} = \|z_{ij} - v\|^2$, that could be refined to better approximate the ML estimate.

Note that when v is not parallel to m_{ij} the constant term of (12) will not vanish. Therefore the weight of the affine term can be reduced when better estimates of v become available. Besides preventing the solution from collapsing to zero, the affine term also compensates for the measurement normalization in ℓ_{ROSE} when linearized around $v = m_{ij}$. This avoids a degradation of the reconstruction caused by noisy measurements closer to the center of distor-

tion. Linearizing (7) around points $v \neq m_{ij}$ has a similar effect, and either of them can be used to achieve more accurate reconstructions.

3.1. Algorithm with Refinement

Algorithm 2: Iterated VarPro for solving (5)

```
Select the inputs η;
Linearize around v = m_ij and set up A and b;
Find X and P using VarPro (Alg. 1);
while true do
    Linearize around v = z_ij and set up A and b;
    Find X and P using VarPro (Alg. 1);
    if stopping criterion then
        break;
    end
end
```

Using the linearization above we can extend Algorithm 1 to a method for optimizing the ML estimate by adding an outer loop iteratively refining the pOSE estimate. In each iteration we solve a VarPro formulation minimizing the linearized objective. Note that we do not update the linearization until it has been full optimized. This way we are able to benefit from VarPro's large basin of convergence. In practice we observe that in many cases the solution estimated is very close to the ML estimate already after the first couple of iterations. Algorithm 2 outlines the approach with refinement. One of the benefits of refinement is that the resulting framework becomes less dependent on η. As illustrated in [12] this parameter should be small, decreasing the influence of the affine term, in order to give visually appealing reconstructions. On the other hand problems with a larger η are more well conditioned. The interpretation of a dampening instead of a model parameter enables us to some extent to compensate for an imperfectly selected η though updated linearizations. Empirical results show that reducing the value of η in each iteration of the outer loop of Algorithm 2 leads to better reconstructions.

3.2. Evaluation

Increased accuracy through refinement In order to show that the update of the affine term in (13) can be used to increase the accuracy of the reconstructions, we repeat the experiments with synthetic data in Section 2.5 with 2 updates of the affine term a described in Algorithm 2. In each update, the weight η is decreased by a factor of 10. The results, plotted in Figure 5, show that performing consecutive linearizations allows us to keep a wide basin of convergence without compromising the accuracy of the algorithm.

Refinement versus Local Optimization Our RpOSE formulation can be combined with local optimization of (7). To show the advantages of doing so, we compared the performance of the proposed method, as described in Algorithm 2, with local optimization initialized with RpOSE solution. Additionally, we also measure the accuracy when local optimization is initialized with a refined solution, ALM or pOSE. In these experiments we select $\eta = 0.05$, and in each linearization its value its decreased by a factor of 10. For pOSE, $\eta = 0.05$ is used, as suggested by the authors, and both RpOSE and pOSE are initialized with random camera matrices and zero translations. Table 1 shows the normalized 3D reconstruction error $\frac{\|X - X_{GT}\|_F}{\|X_{GT}\|_F}$ after projective registration to ground-truth 3D points from multiple datasets. The results show that there is a clear benefit of initializing local optimization with RpOSE when compared to state-of-the-art methods. Furthermore, performing 2 linearization steps followed by local optimization achieved the best overall performance, and therefore this is the combination chosen for future experiments. It is also important to note that initializing local optimization directly with random camera matrices and 3D points yield no reasonable results, and that the proposed refinement process often provides faster convergence than local optimization.

4. Full System Outline

In order to achieve a complete reconstruction, the proposed method is incorporated into a SfM pipeline that allows the recovery of the full model as described in (2). The SfM pipeline used has the following structure:

1. RpOSE factorization with 2 refinement steps and local optimization. Given a set of image points tracked along several images, we use Algorithm 2 with two refinement steps followed by local optimization of (7) to obtain estimations of the first two rows of the uncalibrated camera matrix, and the 3D points, up to projective ambiguity.

2. Camera matrix completion and radial distortion estimation. From the solution of the local optimization of (7), the distortion parameters and third row of the uncalibrated camera matrix are estimated from the equations in (2). Note that by assuming a distortion model with $\kappa(m) = \sum_j k_j \|m\|^{2j}$, for each camera a system of equations of the form

$$M_i \begin{bmatrix} p_i^{(3)} \\ \mathbf{k} \end{bmatrix} = b_i \tag{14}$$

can be obtained, where $p_i^{(3)}$ is the third row of the ith camera matrix, and $\mathbf{k}$ is a vector of the distortion parameters. Here we use a distortion model with three parameters, $k_j, j = 1, ..., 3$. Assuming that the distortion model is constant along all views, the overall system of equations can be

Table 1. Average 3D reconstruction errors obtained with the different combination of methods over 5 datasets. Door(S) represents a smaller version of Door with 0% of missing data. The details of Door, Grossmunster and Kirchenge are shown in Figure 3, while Munsterhof contains 2821 points, 50 viewpoints and 24% of available data.

	Door(S)	Door	Grossmunster	Kirchenge	Munsterhof
RpOSE	0.72%	0.63%	4.00%	21.77%	29.01%
pOSE	**0.09%**	**0.05%**	> 100%	> 100%	> 100%
ALM	0.22%	> 100%	> 100%	79.21%	47.74%
RpOSE+LO	0.41%	0.41%	2.55%	10.12%	17.24%
ALM+LO	0.12%	33.94%	> 100%	80.98%	40.08%
RpOSE+1up	0.12%	0.39%	2.41%	6.94%	10.69%
RpOSE+1up+LO	0.13%	0.39%	**1.59%**	2.02%	7.74%
RpOSE+2up	0.12%	0.38%	1.80%	2.47%	6.67%
RpOSE+2up+LO	0.13%	0.39%	1.62%	**0.80%**	**4.89%**

written as

$$M \begin{bmatrix} p^{(3)} \\ \mathbf{k} \end{bmatrix} = b \tag{15}$$

with p being a $4\times$ #views vector with all third rows of the camera matrices.

3. Bundle adjustment. We perform local optimization of

$$\sum_{ij} w_{ij} \left\| m_{ij} - (1 + \kappa(m_{ij}))\, \Pi \left(P_i \begin{bmatrix} X_j \\ 1 \end{bmatrix} \right) \right\|^2 \tag{16}$$

where $\Pi(z) = [\frac{z_1}{z_3}, \frac{z_2}{z_3}]^T$, starting from the estimations of P, X, and $\mathbf{k}$ found with the previous steps.

4. Euclidean update. The bundle adjustment allow us to obtain a refined estimation of the uncalibrated case, so in order to achieve an Euclidean reconstruction, an additional update step needs to be done. The update consists of estimating the projective transformation $H \in \mathbb{R}^{4\times4}$ such that the factorization $\{PH, H^{-1}X\}$ is a Euclidean reconstruction. Given a valid H, we have that the first three columns of $P_iH_{1:3} = K_iR_i$, and consequently $P_iH_{1:3}H_{1:3}^TP_i^T = K_iK_i^T = \omega_i^*$, which corresponds to the dual absolute conic (DAC) [9]. Here we assume that $K_i = \text{diag}(f_i, f_i, 1)$, and by defining the symmetric matrix $Q = H_{1:3}H_{1:3}^T$ we get that $\left(P_iQP_i^T\right)_{1,1} - \left(P_iQP_i^T\right)_{2,2} = 0$ and all off-diagonal elements of $P_iQP_i^T$ are zero. We refer the reader to [9] for further details.

4.1. Experiments

The whole pipeline is applied to several datasets in order to obtain a complete reconstruction. We compare the accuracy of the pipeline when initialized with RpOSE versus ALM and pOSE. As evaluation metrics, we use camera rotation errors, camera positions (normalized by path length), normalized 3D reconstruction error, reprojection errors (which also depends on the distortion parameters), and focal length estimation error. Since the output of the pipeline is an Euclidean reconstruction, we compared to ground-truth through a similarity transformation. A summary of the results in presented in Table 2. We refer the reader to the supplementary material for visualizations of the reconstructions.

Table 2. Evaluation metrics for the performance of the proposed pipeline for SfM when initilized from RpOSE and ALM, for the Door (D), Fountain-p11 (F), Kirchenge (K), and Grossmunster (G) datasets. Note that the high level of missing data and scene complexity makes other factorization methods basically unusable, while RpOSE allow high-accuracy reconstructions to be achieved. With 'N/A' we mark the metrics which were not provided by the corresponding dataset.

	Method	D	F	K	G
Rot. [deg]	RpOSE	**2.74**	**2.060**	**0.589**	**0.113**
	ALM	90.171	103.7	122.4	2.097
Pos. [%]	RpOSE	**119**	**0.970**	**1.538**	N/A
	ALM	795.1	291.3	241.7	N/A
3D [%]	RpOSE	**2.905**	N/A	**0.558**	**1.586**
	ALM	11.275	N/A	1034.9	13.742
2D [pix]	RpOSE	**0.101**	**0.264**	**0.582**	**0.742**
	ALM	7.708	11.50	3.386	0.754
Focal [%]	RpOSE	**23.6**	**4.928**	N/A	N/A
	ALM	90.677	96.7	N/A	N/A

5. Conclusions

This paper presents a factorization methods invariant to radial distortion and changes in focal length. The factorization is based on 1-D radial cameras, and combined with VarPro provides a stable method that can be initialized with random camera matrices, making it it a reliable starting solution for more complete SfM pipelines. The results show that the proposed solution outperforms state-of-the-art factorization methods. The proposed approach can potentially be extended to Non-rigid SfM problems, for which we refer the readers to the supplementary material for some preliminary results.

References

[1] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3873–3881. Curran Associates, Inc., 2016. 1

[2] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 1

[3] Alessio Del Bue, João M. F. Xavier, Lourdes Agapito, and Marco Paladini. Bilinear modeling via augmented lagrange multipliers (BALM). *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(8):1496–1508, 2012. 1

[4] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *International Conference on Computer Vision (ICCV)*, 2013. 1

[5] Y. Dai, H. Li, and M. He. Projective multiview structure and motion from element-wise factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2238–2251, 2013. 1

[6] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. 2

[7] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint*, arxiv:1704.00708, 2017. 1

[8] Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2016. 1

[9] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 8

[10] Anders Heyden. Projective structure and motion from image sequences using subspace methods. In Michael Frydrych, Jussi Parkkinen, and Ari Visa, editors, *Proceedings of the 10th Scandinavian Conference on Image Analysis*, pages 963–968, 1997. 1

[11] Je Hyeong Hong and Andrew Fitzgibbon. Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts. In *Int. Conf. on Computer Vision*, 2015. 1, 4

[12] Je Hyeong Hong and Christopher Zach. pose: Pseudo object space error for initialization-free bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3, 4, 5, 6, 7

[13] 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)*, pages 5939–5947, 2017. 1, 3

[14] Je Hyeong Hong, Christopher Zach, Andrew W. Fitzgibbon, and Roberto Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *European Conf. on Computer Vision*, 2016. 1

[15] José Pedro Iglesias, Carl Olsson, and Marcus Valtonen Örnhag. Accurate optimization of weighted nuclear norm for non-rigid structure from motion. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[16] Yoni Kasten, Amnon Geifman, Meirav Galun, and Ronen Basri. Gpsfm: Global projective sfm using algebraic constraints on multi-view fundamental matrices. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3259–3267, 2019. 1

[17] Jae-Hak Kim, Yuchao Dai, Hongdong li, Xin Du, and Jonghyuk Kim. Multi-view 3d reconstruction from uncalibrated radially-symmetric cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1896–1903, 12 2013. 2, 4, 5, 6

[18] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 51–60. IEEE, 2020. 1

[19] Viktor Larsson, Torsten Sattler, Zuzana Kukelova, and Marc Pollefeys. Revisiting radial distortion absolute pose. In *International Conference on Computer Vision (ICCV)*. IEEE, September 2019. 2

[20] Viktor Larsson, Nicolcas Zobernig, Kasim Taskin, and Marc Pellefeys. Calibration-free structure-from-motion with calibrated radial trifocal tensors. In *European Conference of Computer Vision*, 2020. 2, 3, 4, 5

[21] Ludovic Magerand and Alessio Del Bue. Practical projective structure from motion (p2sfm). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 39–47, 2017. 1

[22] Behrooz Nasihatkon, Richard I. Hartley, and Jochen Trumpf. A generalized projective reconstruction theorem and depth constraints for projective factorization. *Int. J. Comput. Vis.*, 115(2):87–114, 2015. 1

[23] Marcus Valtonen Ornhag, Carl Olsson, and Anders Heyden. Bilinear parameterization for differentiable rank-regularization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun 2020. 1, 2

[24] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 65–74, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. 1

[25] Conrad J. Poelman and Takeo Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(3):206–218, 1997. 1

[26] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference*

on Computer Vision and Pattern Recognition (CVPR), 2017. 5

[27] D. Strelow, Q. Wang, L. Si, and A. Eriksson. General, nested, and constrained wiberg minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1803–1815, 2016. 3

[28] Peter F. Sturm and Bill Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proceedings of the 4th European Conference on Computer Vision-Volume II - Volume II*, ECCV '96, page 709–720, Berlin, Heidelberg, 1996. Springer-Verlag. 1

[29] Sriram Thirthala and Marc Pollefeys. Radial multi-focal tensors. *Int. J. Comput. Vision*, 96(2):195–211, Jan. 2012. 2

[30] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 1

[31] B. Triggs. Factorization methods for projective structure and motion. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 845–851, 1996. 1

[32] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, August 1987. 2