Self-Calibrating Isometric Non-Rigid Structure-from-Motion

Shaifali Parashar, Adrien Bartoli, Daniel Pizarro

Self-Calibrating Isometric Non-Rigid Structure-from-Motion

Shaifali Parashar[1], Adrien Bartoli[1], and Daniel Pizarro[2,1]

[1] Institut Pascal - CNRS/Université Clermont Auvergne, Clermont-Ferrand, France
[2] GEINTRA, Universidad de Alcalá, Alcalá de Henares, Spain
Code available at `http://igt.ip.uca.fr/~ab/`

Abstract. We present self-calibrating isometric non-rigid structure-from-motion (SCIso-NRSfM), the first method to reconstruct a non-rigid object from at least three monocular images with constant but unknown focal length. The majority of NRSfM methods using the perspective camera simply assume that the calibration is known. SCIso-NRSfM leverages the recent powerful differential approaches to NRSfM, based on formulating local polynomial constraints, where local means correspondence-wise. In NRSfM, the local shape may be solved from these constraints. In SCIso-NRSfM, the difficulty is to also solve for the focal length as a global variable. We propose to eliminate the shape using resultants, obtaining univariate polynomials for the focal length only, whose sum of squares can then be globally minimized. SCIso-NRSfM thus solves for the focal length by integrating the constraints for all correspondences and the whole image set. Once this is done, the local shape is easily recovered. Our experiments show that its performance is very close to the state-of-the-art methods that use a calibrated camera.

Keywords: NRSfM, self-calibration, uncalibrated camera, differential geometry, metric tensor, Christoffel symbols, resultants

1 Introduction

Estimating the intrinsic camera parameters from images is known as camera self-calibration. In Structure-from-Motion (SfM), which is a mature technique for the 3D reconstruction of rigid objects from monocular images, the intrinsic parameters are required to achieve Euclidean 3D reconstruction [12]. SfM may use calibrated images directly [23, 29] or uncalibrated images with self-calibration [26, 22]. SfM was extended to handle non-rigid (deformable) objects in the last two decades with Non-Rigid Structure-from-Motion (NRSfM). While most recent SfM methods use the perspective camera, many early NRSfM methods [9, 34, 28, 3, 7, 30] use a metric affine camera, namely the orthographic or weak-perspective camera. They handle uncalibrated images because these metric affine cameras only have a scale factor as intrinsic parameter, which couples with the scale of the 3D structure in the reconstruction equations. However, the use of these metric affine cameras restricts the imaging conditions [12], which

limits practical applicability. Concretely, the affine camera models do not capture the perspective effect and may thus be inaccurate. They may also yield flip ambiguities in the reconstruction. More recent NRSfM methods [17, 6, 5, 18] use the perspective camera. They can thus cope with broader imaging conditions, are generally more accurate and do not suffer from the flip ambiguities. However, they assume that the camera is calibrated, which puts a different limit on their applicability. Two exceptions are [17, 24], which assume that parts of the scene remain rigid. These rigid parts are used to self-calibrate the camera using an SfM method such as [20]. The estimated calibration is then used in a calibrated NRSfM method. Therefore [17, 24] do not solve the problem of self-calibration in NRSfM strictly speaking but use a workaround based on a sensible but very strong assumption on the scene contents. Since there is a positive gain in choosing the perspective camera, self-calibrating NRSfM appears to be a natural and important problem to study.

We study self-calibrating NRSfM for isometrically deforming surfaces, widely used in recent work [18, 5, 24, 34, 28, 4, 33]. Isometry is one of the most intuitive deformation model and approximates the majority of real-life deformations. In order to deal with uncalibrated images, we use the common assumption that the camera has square pixels and a known principal point lying at the image center. Thus, the only intrinsic parameter which needs to be estimated is the focal length. Assuming that the focal length is constant, we propose a solution based on solving a univariate polynomial, modeling the contribution of $N \geq 3$ images in a least-squares fashion. Our method takes inspiration from a recent solution to isometric NRSfM [18]. This solution uses the image warps to constrain the differential 3D structure. The method uses advanced concepts from Riemannian geometry, namely the Metric Tensor (MT) and the Christoffel Symbols (CS). The MT represents the local surface structure and the CS expresses the rate of change of the MT. In addition, the method uses the concept of infinitesimal planarity, which is widely used in differential geometry. According to this assumption, the surface is planar at an infinitesimal level but maintains its curvature at the global level. The method arrives at Partial Differential Equations (PDEs) that can be converted to algebraic equations in two shape variables and solved locally. By locally we mean that the solution is obtained at each point correspondence independently. The two variables, related to the local 3D shape, are computed in [18] by minimizing the sum-of-squares of the algebraic equations using a computationally expensive polynomial optimization engine [13]. This local solution handles both wide and short baseline data and naturally copes with missing data and occlusions.

We introduce the focal length as an additional variable to the Riemannian framework of [18]. This leaves the CS unaltered but changes the MT. The reconstruction equations also change, containing the two local shape variables, similarly to [18], and a global variable representing the focal length. These equations are degree 5 polynomials, which means that the derivative of their sum-of-squares is a degree 9 polynomial in 3 variables, which is by far out of bounds for the existing polynomial optimization engines such as [13]. We propose a solution

by segregating the global focal length from the local shape variables using the resultants. We obtain univariate polynomials in terms of the focal length. In spite of their high degree, they can be easily solved globally by minimizing their sum-of-squares using a standard root finding algorithm. This global formulation accumulates the local constraints for all correspondences and all images, making the focal length well-constrained and the solution stable. We finally use the focal length estimate to solve isometric NRSfM locally. Our solution improves on [18] by dropping the dependency on [13]. Concretely, it minimizes the sum-of-squares of univariate polynomials for each of the two shape variables, obtained using the resultant of the original multivariate reconstruction equations. Our experiments show that the focal length estimated by SCIso-NRSfM is close to the ground truth and the 3D reconstructed shape very close to calibrated NRSfM methods [5, 18]. We also compare with the NRSfM methods [9, 34] that use an orthographic camera and found that these are outperformed by SCIso-NRSfM.

2 State-of-the-Art and Contributions

Self-calibration has been extensively studied for SfM. It follows one of several possible scenarios where the camera intrinsics are partially constrained. The first solution [8] introduced the Kruppa equations, which use the epipolar geometry to draw constraints on the camera intrinsics. However, they suffer from singularities. Later, [21] proposed a stratified approach where a projective reconstruction is upgraded to affine using a modulus constraint, and further upgraded to Euclidean using linear constraints [11]. In contrast, a direct projective to metric upgrade was done by [14]. The most successful approach finds the explicit location of the absolute quadric using its dual [31]. It obtains a global solution to the fixed camera intrinsic scenario by solving algebraic equations. Based on this model, [20] proposed a linear algorithm to estimate a varying focal length.

Self-calibration was scarcely studied for deformable objects, partly because the subject is more recent than SfM and partly because it forms a less constrained problem. A related problem to NRSfM is Shape-from-Template (SfT) which uses a deformable 3D template and a single input image [1, 25]. A solution to isometric SfT with focal length calibration was proposed in [2]. It works by solving for the focal length locally and using the median of these local solutions as final estimate. The local solutions were found to have a large spread across the input image. This is because locally the focal length is weakly constrained.

Self-calibration in NRSfM forms a difficult and open problem. First, the successful algebraic framework of the dual absolute quadric is based on the rigidity constraint and can thus not be borrowed from SfM. Second, the differential method for calibration in SfT showed signs of instabilities, even if SfT is a much more constrained problem than NRSfM. Our solution uses a differential framework in order to deal with deformations but estimates the focal length globally, by combining local constraints from all point correspondences and all images. More precisely, we make the following main contributions. *1)* We show how to form algebraic constraints for each point correspondence and image pair. These

constraints depend on three variables related to the focal length and the local 3D shape and are not directly solvable. We show how to convert these constraints into easily solvable univariate polynomials. *2)* We show how to form a numerically stable global solution to the focal length by integrating the local constraints over all points and images, and minimizing their sum-of-squares optimally. *3)* We show how, given the estimated focal length, the local 3D shape may be recovered by minimizing the sum-of-squares of univariate polynomials. *4)* We give an algorithm based solely on standard numerical tools.

3 Mathematical Background

Notation. Latin letters denote scalars and Greek letters denote functions. Bold Latin letters denote vectors and matrices. There are a few exceptions however, and $\mathbf{\Gamma}$, which denotes the CS matrix, is one of them. We use superscripts to index the $N \geq 3$ images. The reference image has index 1, without loss of generality. The other images have indexes $(j, r) \in \{2, \dots, N\}$. We often drop the reference image index from the equations for the sake of clarity. For instance, the inverse depth function for image 1 will be defined as β^1 but often referred to as β. We use the subscript $i \in \{1, \dots n\}$ to refer to a particular point correspondence, with n the total number of correspondences.

Surface and camera models. We model 3D surfaces as Riemannian manifolds. Fig. 1 shows a surface $\mathcal{M}$ viewed in image $\mathcal{I}$. We use the perspective camera model $\mathit{\Pi}$. It takes as input the 3D point $\mathbf{Q} = \begin{pmatrix} x & y & z \end{pmatrix}^\top$ and outputs its normalized retinal coordinates $\mathbf{r} = \mathit{\Pi}(\mathbf{Q}) = \begin{pmatrix} \frac{x}{z} & \frac{y}{z} \end{pmatrix}^\top$. We translate the image coordinates so that the principal point aligns with the origin. This allows the intrinsic parameter matrix $\mathbf{K}$ to be expressed in terms of the focal length f only, where $f > 0$ is expressed in px, meaning in number of pixels. The pixel coordinates $\mathbf{p} = \begin{pmatrix} u & v \end{pmatrix}^\top$ are then related to the retinal coordinates as $\mathbf{r} = \begin{pmatrix} \frac{u}{f} & \frac{v}{f} \end{pmatrix}^\top = \frac{\mathbf{p}}{f}$.

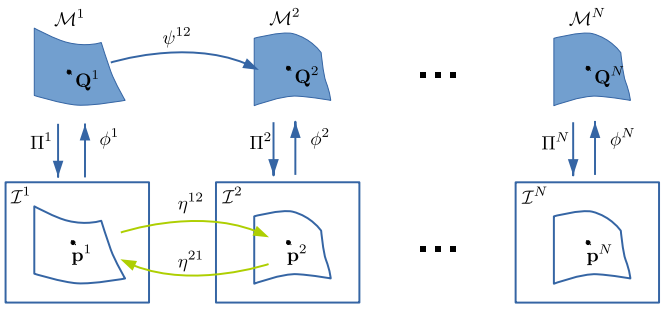


Fig. 1: Principal notations.

The image embedding ϕ is the 'inverse' of the projection Π for the points on $\mathcal{M}$. It maps the retinal coordinates $\mathbf{r}$ to the 3D point $\mathbf{Q}$ as:

$$\phi(\mathbf{r}) = \frac{1}{\beta(\mathbf{r})} \begin{pmatrix} \mathbf{r}^\top & 1 \end{pmatrix}^\top, \tag{1}$$

where $\beta(\mathbf{r})$ is the inverse-depth function. We omit the argument $\mathbf{r}$ in the subsequent use of $\phi(\mathbf{r})$, $\beta(\mathbf{r})$ and most other functions. We use ϕ to express the differential properties of the surface derived from the two concepts of Riemannian geometry, MT and CS, which we describe shortly.

Modeling NRSfM. We use a very similar model to [18], shown in Fig. 1. The goal of [18] was to solve for NRSfM with a calibrated camera but we also solve for the focal length. The model has N isometrically deforming surfaces $\mathcal{M}^1, \ldots, \mathcal{M}^N$ projected in the input images $\mathcal{I}^1, \ldots, \mathcal{I}^N$. The image warp η^{j1} represents the optic flow between $\mathcal{I}^j$ and $\mathcal{I}^1$. We have $\eta^{1j} = \left(\eta^{j1}\right)^{-1}$. We compute the warps from keypoint correspondences using [19]. The surfaces $\mathcal{M}^1$ and $\mathcal{M}^j$ are related by an isometric deformation function ψ^{1j}. Isometricity is the main constraint we use in SCIso-NRSfM.

Metric Tensor. Denoted $\mathbf{g}[\phi]$, the MT is a first-order differential quantity that describes physical surface properties such as lengths, angles and areas [16]. It can be derived from $\mathbf{J}_\phi$, the Jacobian of ϕ. Using ϕ from equation (1), $\mathbf{g}[\phi]$ is shown to be a 2×2 matrix given by:

$$\mathbf{g}[\phi] = \mathbf{J}_\phi^\top \mathbf{J}_\phi \text{ with } \mathbf{J}_\phi = \frac{1}{f\beta} \begin{pmatrix} 1 - u\zeta & -v\zeta & -f\zeta \\ -u\kappa & 1 - v\kappa & -f\kappa \end{pmatrix}^\top, \tag{2}$$

where we define the inverse-depth derivatives as $\beta_u = \frac{\partial \beta}{\partial u}$, $\beta_v = \frac{\partial \beta}{\partial v}$, and their ratio with the inverse-depth as $\zeta = \frac{\beta_u}{\beta}$, $\kappa = \frac{\beta_v}{\beta}$. For isometric surfaces, the MT is transferable across images using the first-order derivatives of the warps [18]:

$$\mathbf{g}[\phi^j] = \mathbf{J}_{\eta^{j1}}^\top \mathbf{g}[\phi^1] \mathbf{J}_{\eta^{j1}}. \tag{3}$$

Christoffel Symbols. Denoted $\mathbf{\Gamma}^u[\phi]$ and $\mathbf{\Gamma}^v[\phi]$, the CS are second-order differential quantities that describe the curvature of a surface [16]. They are defined as the rate of change of the MT. They usually have a very long and complex expression. This is however reduced using the infinitesimally planarity assumption, which allows one to neglect the second-order derivatives of the image embedding. This means that β in equation (1) is infinitesimally linear. Using ϕ from equation (1), $\mathbf{\Gamma}^u[\phi]$ and $\mathbf{\Gamma}^v[\phi]$ are shown to be 2×2 matrices given by:

$$\mathbf{\Gamma}^u[\phi] = -\begin{pmatrix} 2\zeta & \kappa \\ \kappa & 0 \end{pmatrix} \quad \mathbf{\Gamma}^v[\phi] = -\begin{pmatrix} 0 & \zeta \\ \zeta & 2\kappa \end{pmatrix}. \tag{4}$$

For isometric surfaces, the CS are transferable across images using the first- and second-order derivatives of the warps [18]:

$$\zeta^j = \frac{\partial u^1}{\partial u^2}\zeta^1 + \frac{\partial v^1}{\partial u^2}\kappa^1 - \left(\frac{\partial^2 u^1}{\partial u^2 \partial v^2} \frac{\partial v^2}{\partial u^1} + \frac{\partial^2 v^1}{\partial u^2 \partial v^2} \frac{\partial v^2}{\partial v^1} \right)$$

$$\kappa^j = \frac{\partial u^1}{\partial v^2}\zeta^1 + \frac{\partial v^1}{\partial v^2}\kappa^1 - \left(\frac{\partial^2 u^1}{\partial u^2 \partial v^2}\frac{\partial u^2}{\partial u^1} + \frac{\partial^2 v^1}{\partial u^2 \partial v^2}\frac{\partial u^2}{\partial v^1}\right). \tag{5}$$

Resultants. The resultant of two polynomials is a polynomial expression of their coefficients, which is equal to zero if and only if the polynomials have a common root [35]. This allows one to find the common roots of a system of polynomials. Consider as an example two bivariate polynomials $\alpha(t,u)$ and $\gamma(t,u)$ of degree l and m respectively and in variables t,u. Their resultant $\mathrm{Res}_t(\alpha,\gamma)$ with respect to t is a univariate polynomial in u. It is given as the determinant of the so-called Sylvester matrix $\mathbf{S}_t \in \mathbb{R}^{(l+m)\times(l+m)}$ as $\mathrm{Res}_t(\alpha,\gamma) = \det(\mathbf{S}_t)$. The elements of the Sylvester matrix depend on the coefficients of α,γ.

4 Self-calibrating Isometric NRSfM

We first derive the reconstruction equations. These are constraints depending on two local shape variables and the focal length. We then show how these constraints can be optimized globally for just the focal length, and then locally for the shape.

4.1 The Reconstruction Equations

The reconstruction equations are built starting from the MT transfer equation (3). This equation involves the MT $\mathbf{g}[\phi^j]$, which is expressed in terms of the embedding's Jacobian $\mathbf{J}_{\phi^j}$ given by equation (2). The latter involves (ζ^j,κ^j), the ratios of inverse depth derivatives to inverse depth in image j. Because these are elements of the CS, we can express them in terms of the same ratios (ζ,κ) taken in image 1 using the CS transfer equation (5). We thus obtain a new expression of the MT $\mathbf{g}[\phi^j]$ depending on $(f,\beta^1,\beta^j,\zeta,\kappa)$. By substituting this expression in the MT transfer equation (3), we obtain a 2×2 matrix equation. Taking ratios, (β^1,β^j) vanish and we arrive at two independent algebraic PDEs $\mathcal{E}_{1,2}$ in (f,ζ,κ). These PDEs have coefficients (a_t^j, b_t^j) and are given by:

$$\mathcal{E}_1^j(f,\zeta,\kappa) = \sigma_7^j\zeta^3 + \sigma_5^j\zeta^2 + \sigma_3^j\zeta + \sigma_1^j \tag{6}$$

$$\mathcal{E}_2^j(f,\zeta,\kappa) = \sigma_8^j\zeta^3 + \sigma_6^j\zeta^2 + \sigma_4^j\zeta + \sigma_2^j, \tag{7}$$

with:

$$\begin{aligned}
\sigma_1^j &= a_{27}^j + a_{26}^j\kappa + a_{24}^j\kappa^2 + a_{21}^j\kappa^3 + s(a_{11}^j\kappa^3 + a_{14}^j\kappa^2 + a_{16}^j\kappa + a_{17}^j) + s^2(a_4^j\kappa^3 + a_7^j\kappa^2)\\
\sigma_2^j &= b_{27}^j + b_{26}^j\kappa + b_{24}^j\kappa^2 + b_{21}^j\kappa^3 + s(b_{11}^j\kappa^3 + b_{14}^j\kappa^2 + b_{16}^j\kappa + b_{17}^j) + s^2(b_4^j\kappa^3 + b_7^j\kappa^2)\\
\sigma_3^j &= a_{25}^j + a_{23}^j\kappa + a_{20}^j\kappa^2 + s(a_{10}^j\kappa^2 + a_{13}^j\kappa + a_{15}) + s^2(a_6^j\kappa + a_3^j\kappa^2)\\
\sigma_4^j &= b_{25}^j + b_{23}^j\kappa + b_{20}^j\kappa^2 + s(b_{10}^j\kappa^2 + b_{13}^j\kappa + b_{15}) + s^2(b_6^j\kappa + b_3^j\kappa^2)\\
\sigma_5^j &= a_{22}^j + a_{19}^j\kappa + s(a_{12}^j + a_9^j\kappa) + s^2(a_5^j + a_2^j\kappa) \qquad \sigma_7^j = a_{18}^j + sa_8^j + s^2a_1^j\\
\sigma_6^j &= b_{22}^j + b_{19}^j\kappa + s(b_{12}^j + b_9^j\kappa) + s^2(b_5^j + b_2^j\kappa) \qquad \sigma_8^j = b_{18}^j + sb_8^j + s^2b_1^j \qquad s = f^2.
\end{aligned}$$

The coefficients (a_t^j, b_t^j) directly depend on the derivatives of the warp η^{1j}. Choosing an image j, fixing a single point $\mathbf{r}$ and defining $k_1 = \zeta(\mathbf{r})$, $k_2 = \kappa(\mathbf{r})$, we obtain two algebraic equations $\mathcal{E}_{1,2}^j(f, k_1, k_2)$. For N images and a single point, we thus have a set of $2(N-1)$ polynomials $\mathfrak{E}_{12}(f, k_1, k_2) = \{\mathcal{E}_1^j(f, k_1, k_2), \mathcal{E}_2^j(f, k_1, k_2)\}_{j=2}^N$. Similar but simpler equations were obtained in [18] for a known focal length. These were then solved locally by minimizing their sum-of-squares using a computationally expensive polynomial optimization engine. This strategy cannot be used to estimate the focal length however, for two reasons. First, estimating the focal length locally would be extremely unstable. Second, the degree of the equations become prohibitive for the existing optimization engines. We next discuss our approach to obtain a global and tractable solution to f and a local solution to (k_1, k_2).

4.2 Solving for the Focal Length Globally

We show how to use the reconstruction equations $\mathfrak{E}_{12}(f, k_1, k_2)$ to find f globally. We use resultants to eliminate the dependency on (k_1, k_2), starting with k_1.

Eliminating k_1. The resultant of $\mathcal{E}_1^j, \mathcal{E}_2^j$ with respect to k_1 gives a new polynomial $\mathcal{E}_3^j$ depending on (f, k_2). Defining the Sylvester matrix $\mathbf{S}_{k_1} \in \mathbb{R}^{6\times 6}$ as shown in Fig. 2 (left), we have:

$$\begin{aligned}\mathcal{E}_3^j(f, k_2) &= \mathrm{Res}_{k_1}(\mathcal{E}_1(f, k_1, k_2), \mathcal{E}_2(f, k_1, k_2), k_1) = \det(\mathbf{S}_{k_1}) \\ &= c_9^j k_2^9 + c_8^j k_2^8 + c_7^j k_2^7 + c_6^j k_2^6 + c_5^j k_2^5 + c_4^j k_2^4 + c_3^j k_2^3 + c_2^j k_2^2 + c_1^j k_2 + c_0^j, \qquad (8)\end{aligned}$$

where c_t^j are polynomials of degree 12 in $s = f^2$. Numerically, they are often of degree 3 or 4. For N images, we thus obtain $N-1$ polynomial equations $\mathfrak{E}_3(f, k_2) = \{\mathcal{E}_3^j(f, k_2)\}_{j=2}^N$.

Eliminating k_2. We eliminate k_2 by evaluating the resultant of the equation for two image pairs, $(1, j)$ and $(1, r)$, in $\mathfrak{E}_3$. This gives a new polynomial equation $\mathcal{E}_4^{jr}$ depending on f only. Defining the Sylvester matrix $\mathbf{S}_{k_2} \in \mathbb{R}^{18\times 18}$ as shown in Fig. 2 (right), we have:

$$\mathcal{E}_4^{jr}(f) = \mathrm{Res}_{k_2}(\mathcal{E}_3^j, \mathcal{E}_3^r) = \det(\mathbf{S}_{k_2}). \qquad (9)$$

For N images, we obtain $\frac{(N-1)(N-2)}{2}$ univariate polynomial equations $\mathfrak{E}_4(f) = \{\mathcal{E}_4^{jr}(f)\}_{j,r\in[2,N],j\neq r}$ of degree 216. Since c_t^j in equation (8) are numerically of degree 3 or 4, the degree of these polynomials lies between 54-72 instead of 216.

Solving for f. A globally optimal solution can be found by minimizing the sum-of-squares of the equation set $\mathfrak{E}_4$. For n points tracked over N images, we define the sum-of-squares cost as:

$$C(f) = \sum_{i=1}^{n} \sum_{j=2}^{N} \sum_{\substack{r=2 \\ r\neq j}}^{N} \left(\mathcal{E}_4^{jr}(f)\right)^2. \qquad (10)$$

$$\begin{pmatrix} \sigma_7^j & \sigma_5^j & \sigma_3^j & \sigma_1^j & 0 & 0 \\ 0 & \sigma_7^j & \sigma_5^j & \sigma_3^j & \sigma_1^j & 0 \\ 0 & 0 & \sigma_7^j & \sigma_5^j & \sigma_3^j & \sigma_1^j \\ \sigma_8^j & \sigma_6^j & \sigma_4^j & \sigma_2^j & 0 & 0 \\ 0 & \sigma_8^j & \sigma_6^j & \sigma_4^j & \sigma_2^j & 0 \\ 0 & 0 & \sigma_8^j & \sigma_6^j & \sigma_4^j & \sigma_2^j \end{pmatrix}$$

$$\begin{pmatrix} c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_9^j & c_8^j & c_7^j & c_6^j & c_5^j & c_4^j & c_3^j & c_2^j & c_1^j & c_0^j \\ c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_9^r & c_8^r & c_7^r & c_6^r & c_5^r & c_4^r & c_3^r & c_2^r & c_1^r & c_0^r \end{pmatrix}$$

Fig. 2: The Sylvester matrices $\mathbf{S}_{k_1}$ (left) and $\mathbf{S}_{k_2}$ (right).

Using Fermat's interior extremum theorem, a local extrema of C occurs at the critical points, obtained by solving $\frac{\partial C}{\partial f}(f) = 0$. The set of critical points is given by $\mathfrak{F}_c = \{f_c \mid \frac{\partial C}{\partial f}(f_c) = 0\}$. In practice, the cost function C is a univariate polynomial of degree 108-144. We simply find the roots of its derivative polynomial to find $\mathfrak{F}_c$. The local minima are the critical points with a positive value of $\frac{\partial^2 C}{\partial f^2}$. Therefore the set of local minima $\mathfrak{F}_l \subset \mathfrak{F}_c$ is given by $\mathfrak{F}_l = \{f_l \in \mathfrak{F}_c \mid \frac{\partial^2 C}{\partial f^2}(f_l) > 0\}$. Finally, the globally optimal focal length is given by:

$$\hat{f} = \arg\min_{f \in \mathfrak{F}_l} C(f). \tag{11}$$

4.3 Solving for the Local Shape

We show how the local shape, represented by (k_1, k_2), can be solved for given an estimate $\hat{f}$ of the focal length, starting with k_2. Given $\hat{f}$, we have that $\mathfrak{E}_3(\hat{f}, k_2)$ forms a set of univariate polynomials in k_2. We find the optimal solution for k_2 by minimizing the sum-of-squares of these polynomials. For a point tracked over N images, the cost is:

$$C'(k_2) = \sum_{j=2}^{N} \left(\mathcal{E}_3^j(\hat{f}, k_2) \right)^2. \tag{12}$$

Because C' is a univariate polynomial, we find its minimum using the same process as described in the previous section for minimizing C. The optimal solution

$\hat{k}_2$ is thus:

$$\hat{k}_2 = \arg\min_{k_2 \in \mathfrak{K}_2} C'(k_2) \text{ where } \mathfrak{K}_2 = \left\{ k_2 \;\middle|\; \frac{\partial C'}{\partial k_2}(k_2) = 0, \frac{\partial^2 C'}{\partial {k_2}^2}(k_2) > 0 \right\}. \quad (13)$$

Using $(\hat{f}, \hat{k}_2)$, we have that $\mathfrak{E}_{12}(\hat{f}, k_1, \hat{k}_2)$ forms a set of univariate polynomials in k_1. We find the optimal solution for k_1 by minimizing the sum-of-squares of these polynomials. For a point tracked over N images, the cost is:

$$C''(k_1) = \sum_{j=2}^{N} \Big(\mathcal{E}_1(\hat{f}, k_1, \hat{k}_2)\Big)^2 + \Big(\mathcal{E}_2(\hat{f}, k_1, \hat{k}_2)\Big)^2. \quad (14)$$

The optimal $\hat{k}_1$ is then:

$$\hat{k}_1 = \arg\min_{k_1 \in \mathfrak{K}_1} C''(k_1) \text{ where } \mathfrak{K}_1 = \left\{ k_1 \;\middle|\; \frac{\partial C''}{\partial k_1}(k_1) = 0, \frac{\partial^2 C''}{\partial {k_1}^2}(k_1) > 0 \right\}. \quad (15)$$

We arrive at an estimate $\left(\hat{k}_1, \hat{k}_2\right)$ of the local shape for the reference image. By substituting this estimate in equation (5), we obtain an estimate $\left(\hat{k}_1^j, \hat{k}_2^j\right)$ of the local shape for the rest of the images.

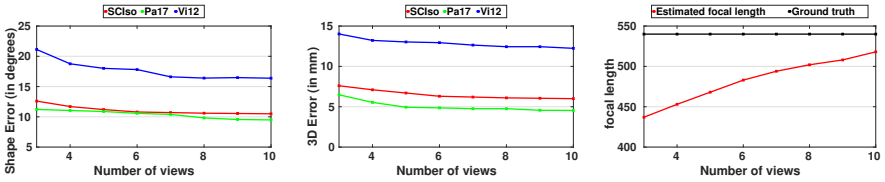
5 Algorithm

We give our algorithm to solve SCIso-NRSfM. For numerical stability, as commonly done in SfM [12], the points' pixel coordinates are standardized using an isotropic scale factor mapping the image boundaries close to $[-1, 1]^2$.

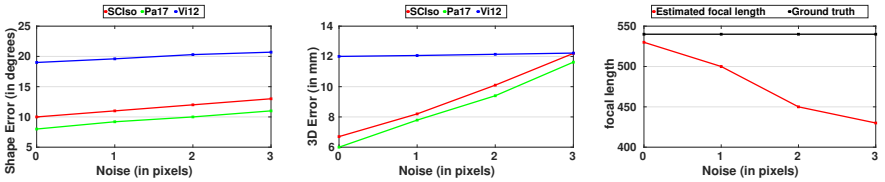
Inputs: Point correspondences $\{\mathbf{p}_i^j\}$ with visibility indicators $\{v_i^j\}$, $i \in [1, n]$, $j \in [1, N]$ ($v_i^j = 1$ means that the ith point is visible in the jth image)

1) *Compute image warps η^{j1}, $j \in [2, N]$.* Use the points visible in the reference and jth images, meaning with indexes $\{i \in [1, n] \mid v_i^1 = v_i^j = 1\}$, to estimate the warp η^{j1} using [19].
2) *Compute the optimal global solution to f.* Find $\hat{f}$ that minimizes C in equation (10).
3) *Compute the optimal local shape (k_1, k_2).* Using the $\hat{f}$ obtained in the previous step, find $\hat{k}_2$ that minimizes C' in equation (12). Then, use $(\hat{f}, \hat{k}_2)$ to find $\hat{k}_1$ that minimizes C'' in equation (14).
4) *Find normals and 3D points.* Find the Jacobian $\mathbf{J}_\phi$ in terms of $(\hat{k}_1, \hat{k}_2)$ using equation (2). Compute the surface normals $\hat{\mathbf{N}}_i^j$ by normalizing the cross-product of the Jacobians columns. Find the inverse depth β^{-1} by integrating the normals using the method in [18]. Apply the embedding ϕ from equation (1) to recover the points $\hat{\mathbf{Q}}_i^j$.

Outputs: Points $\{\hat{\mathbf{Q}}_i^j\}$, normals $\{\hat{\mathbf{N}}_i^j\}$, $i \in [1, n]$, $j \in [1, N]$, focal length $\hat{f}$.



(a) Varying number of images



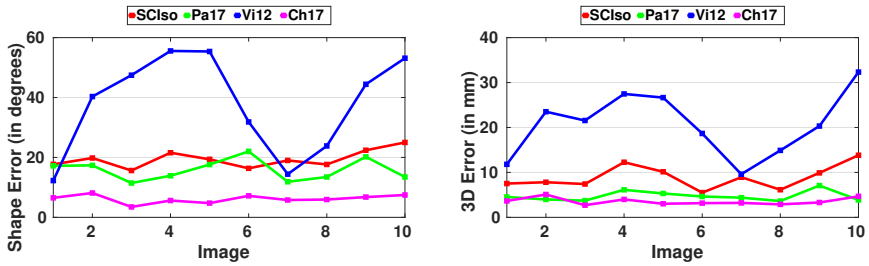
(b) Varying noise

Fig. 3: Results for the *Cylinder* dataset. Mean shape and 3D errors are shown against a varying number of images and noise level. The estimated and true focal lengths are also shown. Best viewed in color.

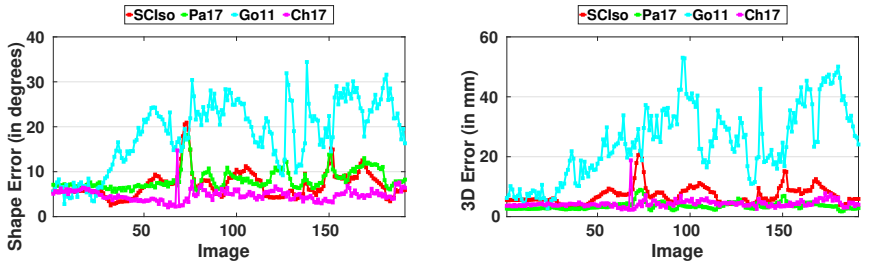
6 Experiments

We tested SCIso-NRSfM (**SCIso**) on a synthetic *Cylinder* dataset [18] and two real datasets, namely *T-shirt* [4] and *Paper* [32] showing objects deforming isometrically. Since self-calibration has not yet been dealt within NRSfM, we compare SCIso-NRSfM with NRSfM methods that assume perspective projection and use calibrated data, **Pa17** [18] and **Ch17** [5]. Also, we compare against methods that assume orthographic projection and avoid the calibration, **Go11** [9] and **Vi12** [34]. For quantitative comparison, we measured the mean shape error (RMSE between computed and ground truth normals in degrees) and the 3D error (RMSE between computed and ground truth 3D points in mm).

Cylinder dataset. This dataset contains randomly generated views of a cylindrical surface deforming isometrically. The cylindrical surface has a radius varying between 2 and 10. The image size is 640×480 px and the camera focal length is 540 px. The number of point correspondences is 400. We vary the number of images and correspondence noise. We compared all methods except **Go11** and **Ch17**. This is because **Go11** uses the low-rank model and requires a large number of images with short baseline and **Ch17** simply failed on this dataset. Figure 3a shows the mean shape and 3D errors for reconstructions performed with 3-10 images. The correspondence noise is chosen to follow a gaussian distribution with standard deviation of 1 px. The performance of **SCIso** is very similar to **Pa17** which solves NRSfM with a calibrated camera, which means using the true focal length, though **Pa17** performs slightly better. On increasing



(a) *T-shirt* image batch



(b) *Paper* video sequence

Fig. 4: Results for the *T-shirt* and *Paper* datasets. Mean shape and 3D errors are shown. Best viewed in color.

the number of images, all methods tend to obtain better results. The performance of **SCIso** and **Pa17** stabilizes with 5 images. However they yield a good reconstruction with 3 images as well. Therefore, they efficiently solve the minimal case. The estimation of focal length in **SCIso** improves with the number of images. For 8-10 images, it is very close to the true focal length. **Vi12** uses an orthographic camera model. Its performance is significantly worse than **SCIso** and **Pa17**.

Figure 3b shows the mean shape and 3D errors for reconstructions performed with 10 images by varying the noise between 0-3 px. The performance of **SCIso** is, again, very similar to **Pa17**, with **Pa17** performing slightly better. On increasing the noise, they both tend to degrade linearly. Interestingly, **Vi12** is barely affected by the noise in the tested range, however, its performance is consistently significantly worse than the other methods. The estimation of focal length in **SCIso** degrades with noise, though remaining reasonable. Interestingly, because the shape and 3D errors of **SCIso** and **Pa17**, which uses the true focal length, are very close, we can conjecture that the focal length estimate cannot be substantially improved without adding extra priors into the problem formulation.

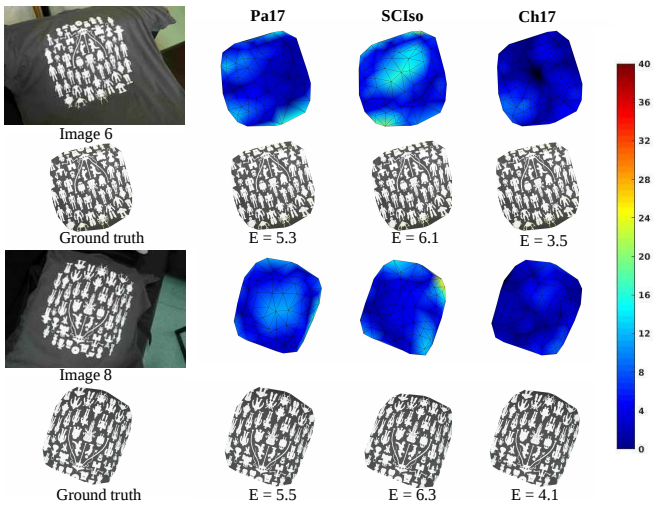


Fig. 5: Error maps and textured rendering of the reconstructed shape for two images of the *T-shirt* dataset. E is mean 3D error (in mm). Best viewed in color.

T-shirt dataset. This dataset was introduced in [4]. It consists of 10 wide-baseline images of an isometrically deforming T-shirt with 85 point correspondences. Camera calibration was obtained carefully using a calibration checkerboard and Matlab's calibration toolbox, yielding a focal length of 3780 px. Figure 4(a) shows the mean shape and 3D errors for all 10 reconstructed surfaces. **Ch17** has the best performance on this dataset, with **Pa17** and **SCIso** being very close. The focal length estimated by **SCIso** is 3954 px which is quite close to the calibrated focal length of 3780 px, with a relative error of 4.6%. **Vi12** does not perform well on this dataset. We did not evaluate **Go11** on this dataset, for the same reason as on the *Cylinder* dataset. Figure 5 shows the renderings of the error maps and textured reconstructed shape for two images.

Paper dataset. This dataset was introduced in [32]. It consists of 191 images from a video sequence with 1500 point correspondences of a paper deforming isometrically. Camera calibration obtained from standard methods is provided, with a focal length of 528 px. Figure 4b shows the mean shape and 3D errors for all the 191 reconstructed surfaces. **Ch17** has the best performance on this dataset, with **Pa17** and **SCIso** being very close. The focal length estimated by **SCIso** using the first 10 images is 498 px, which is close to the actual focal length of 528 px, with a relative error of 5.7%. **Go11** did not perform as well as the other methods. This may be explained by the fact that it uses an orthographic camera or because it is based on the low rank shape model. We could not evaluate **Vi12** on this dataset because of its prohibitive computation time. Figure 6 shows the renderings of the error maps and textured reconstructed shape for three images.

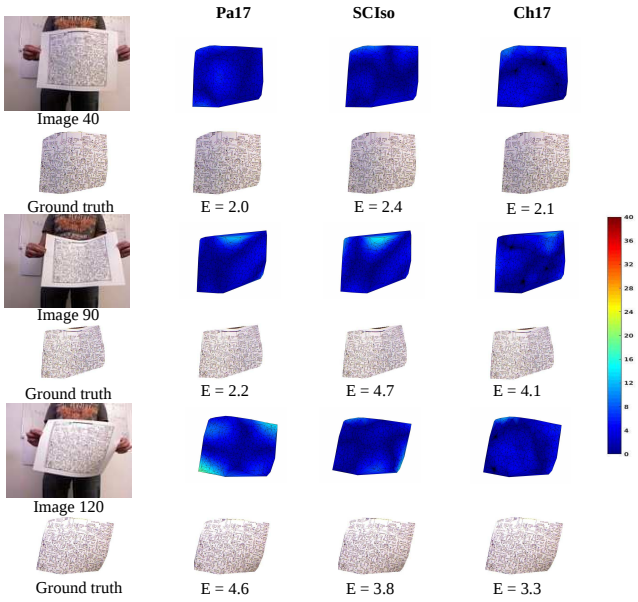


Fig. 6: Error maps and textured rendering of the reconstructed shape for three images of the *Paper* dataset. E is mean 3D error (in mm). Best viewed in color.

	SCIso			**Pa17**		**Ch17**		**Go17**		**Vi12**	
	Es	Ep	%f	Es	Ep	Es	Ep	Es	Ep	Es	Ep
T-shirt	19.1	8.9	4.6	15.8	4.7	6.1	3.5	xx	xx	37.8	20.6
Paper	9.6	7.1	5.7	6.2	3.8	4.8	4.3	18.6	24.7	xx	xx

Table 1: Summary of results on the *T-shirt* and *Paper* datasets. Es and Ep represent mean shape and 3D error respectively. %f denotes the relative error in focal length compared to the calibrated focal length serving as ground truth. xx represents values which could not be computed, as explained in the main text.

Summary of experiments. We compared **SCIso** on a synthetic and two real datasets. We found it to be performing very closely to the current state-of-the-art NRSfM methods, namely **Pa17** and **Ch17**, that use a calibrated perspective camera. **SCIso** estimates the focal length to a good relative accuracy. We also compared with two NRSfM methods that use the orthographic camera, namely **Go11** and **Vi12**. These methods do not perform as well as **Pa17** and **Ch17** on the tested datasets. The performance of all the compared methods on the real datasets is summarized in table 1.

Computation time. We have implemented our method in MATLAB and the code was not optimized. We used a standard PC with i5 CPU and 16 GB RAM. We first solve for f and then for the shape. In order to solve for f, it takes about 76 s to form the polynomial in equation (9) and 90 s to compute its derivatives and

find its roots. Computing the shape takes about 10 ms. It is much faster than Pa17 (1.5 s), which is already significantly faster than the rest of the compared methods, according to [18].

7 Conclusions

We have presented SCIso-NRSfM, a theory and an algorithm to reconstruct an isometrically deforming object from monocular uncalibrated images with constant but unknown focal length. SCIso-NRSfM represents the first step in joining self-calibration and NRSfM. We have used a differential approach and derived a system of polynomial PDEs. Upon eliminating the shape variables from these using resultants, we have then showed that the focal length could be recovered optimally by globally minimizing the constraints arising from all correspondences and all images in a single least-squares cost. Our experimental results have shown that SCIso-NRSfM compares very favorably to existing calibrated NRSfM algorithms against the number of images and correspondence noise, and recovers the focal length with a relative error of a few percents. They have also shown that SCIso-NRSfM works well for the minimal case of three images and improved with the number of images.

We finally give two possible lines of future research. First, in SfM, for rigid objects, there exist Critical Motion Sequences (CMS) [27, 15] in which case self-calibration cannot be resolved. These typically happen when the camera motion is not general enough, for instance when all optical axes intersect. The possible existence of CMS in SCIso-NRSfM is then a very natural question. In the deformable case however, the question must be addressed by considering the pose of the local surface with respect to the camera. In other words, there is not a unique pose for each image, but a continuously varying pose across the surface. This diversity seems to play in favor of dramatically reducing the chance of encountering a degenerate case in SCIso-NRSfM. Nonetheless, this is something we intend to thoroughly study in the near future. The second possible line of future research is to exploit SCIso-NRSfM in plane-based self-calibration. Almost all existing methods, such as [10], take as input a set of homographies relating the input images. It is well known that, when the observed plane is smaller in an image, the computed homography may be unstable. Interestingly, SCIso-NRSfM does not require homographies as inputs but uses the assumption of IP, which suggests that it forms differential constraints for infinitesimal planes. How these constraints relate to the absolute conic formalism now widely accepted in self-calibration and whether these constraints may aid plane-based self-calibration are profound questions which we indeed to study in the near future.

Acknowledgements. This research has received funding from the EU's FP7 through the ERC research grant 307483 FLEXABLE, the Spanish Ministry of Economy, Industry and Competitiveness under project ARTEMISA (TIN2016-80939- R) and the University of Alcalá, Spain under the project SEQUENCE (CCGP2017-EXP/048).

References

1. Bartoli, A., Gérard, Y., Chadebecq, F., Collins, T., Pizarro, D.: Shape-from-template. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(10), 2099–2118 (2015)
2. Bartoli, A., Pizarro, D., Collins, T.: A robust analytical solution to isometric shape-from-template with focal length calibration. In: ICCV (2013)
3. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: CVPR (2000)
4. Chhatkuli, A., Pizarro, D., Bartoli, A.: Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In: BMVC (2014)
5. Chhatkuli, A., Pizarro, D., Collins, T., Bartoli, A.: Inextensible non-rigid structure-from-motion by second-order cone programming. IEEE transactions on pattern analysis and machine intelligence pp. 1–1 (2017)
6. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. International Journal of Computer Vision **107**(2), 101–122 (2014)
7. Del Bue, A., Smeraldi, F., Agapito, L.: Non-rigid structure from motion using non-parametric tracking and non-linear optimization. In: CVPRW (2004)
8. Faugeras, O., Luong, Q.T., Maybank, S.J.: Camera self-calibration: Theory and experiments. In: ECCV (1992)
9. Gotardo, P., Martinez, A.: Kernel non-rigid structure from motion. In: ICCV (2011)
10. Gurdjos, P., Sturm, P.: Methods and geometry for plane-based self-calibration. In: CVPR (2003)
11. Hartley, R.: Self-calibration from multiple views with a rotating camera. In: ECCV (1994)
12. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521623049 (2000)
13. Henrion, D., Lasserre, J.B.: Gloptipoly: Global optimization over polynomials with matlab and sedumi. ACM Transactions on Mathematical Software **29**(2), 165–194 (2003)
14. Heyden, A., Astrom, K.: Euclidean reconstruction from constant intrinsic parameters. In: ICPR (1996)
15. Kahl, F., Triggs, B., Astrom, A.: Critical motions for auto-calibration when some intrinsic parameters can vary. Journal of Mathematical Image and Vision **13**(2), 131146 (2000)
16. Lee, J.: Riemannian manifolds : an introduction to curvature. Springer (1997)
17. Lladó, X., Del Bue, A., Agapito, L.: Non-rigid metric reconstruction from perspective cameras. Image and Vision Computing **28**(9), 1339–1353 (2010)
18. Parashar, S., Pizarro, D., Bartoli, A.: Isometric Non-Rigid Shape-from-Motion with Riemannian Geometry Solved in Linear Time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
19. Pizarro, D., Khan, R., Bartoli, A.: Schwarps: Locally projective image warps based on 2D schwarzian derivatives. International Journal of Computer Vision **119**(2), 93–109 (2016)
20. Pollefeys, M., Koch, R., Van Gool, L.: Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In: ICCV (1998)
21. Pollefeys, M., Van Gool, L.: Stratified self-calibration with the modulus constraint. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(8), 707–724 (1999)

22. Ramachandran, M., Veeraraghavan, A., Chellappa, R.: A fast bilinear structure from motion algorithm using a video sequence and inertial sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1), 186–193 (2011)
23. Ramalingam, S., Lodha, S., Sturm, P.: A generic structure-from-motion framework. Computer Vision and Image Understanding **103**(1), 218–228 (2006)
24. Russell, C., Yu, R., Agapito, L.: Video pop-up: Monocular 3d reconstruction of dynamic scenes. In: ECCV (2014)
25. Salzmann, M., Fua, P.: Linear local models for monocular reconstruction of deformable surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 931–944 (2011)
26. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. International Journal of Computer Vision **80**(1), 189–210 (2008)
27. Sturm, P.: Critical motion sequences for monocular self- calibration and uncalibrated euclidean reconstruction. In: CVPR (1997)
28. Taylor, J., Jepson, A.D., Kutulakos, K.N.: Non-rigid structure from locally-rigid motion. In: CVPR (2010)
29. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision **9**(2), 137–154 (1992)
30. Torresani, L., Hertzmann, A., Bregler, C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30(5)**, 878–892 (2008)
31. Triggs, B.: Autocalibration and the absolute quadric. In: CVPR (1997)
32. Varol, A., Salzmann, M., Fua, P., Urtasun, R.: A constrained latent variable model. In: CVPR (2012)
33. Varol, A., Salzmann, M., Tola, E., Fua, P.: Template-free monocular reconstruction of deformable surfaces. In: ICCV (2009)
34. Vicente, S., Agapito, L.: Soft inextensibility constraints for template-free non-rigid reconstruction. In: ECCV (2012)
35. Van der Waerden, B.L.: Modern Algebra. Springer (2003)