Inextensible Non-Rigid Shape-from-Motion by Second-Order Cone Programming

Ajad Chhatkuli[a], Daniel Pizarro[b,a], Toby Collins[a] and Adrien Bartoli[a]

[a]ISIT - CNRS/Université d'Auvergne, Clermont-Ferrand, France
[b]GEINTRA, Universidad de Alcalá, Alcalá de Henares, Spain

Abstract

We present a global and convex formulation for template-less 3D reconstruction of a deforming object with the perspective camera. We show for the first time how to construct a Second-Order Cone Programming (SOCP) problem for Non-Rigid Shape-from-Motion (NRSfM) using the Maximum-Depth Heuristic (MDH). In this regard, we deviate strongly from the general trend of using affine cameras and factorization-based methods to solve NRSfM. In MDH, the points' depths are maximized so that the distance between neighbouring points in camera space are upper bounded by the geodesic distance. In NRSfM both geodesic and camera space distances are unknown. We show that, nonetheless, given point correspondences and the camera's intrinsics the whole problem is convex and solvable with SOCP. We show with extensive experiments that our method accurately reconstructs quasi-isometric surfaces from partial views under articulated and strong deformations. It naturally handles missing correspondences, non-smooth objects and is very simple to implement compared to previous methods, with only one free parameter (the neighbourhood size).

1. Introduction

Non-Rigid Shape-from-Motion (NRSfM) is the problem of finding the 3D shape of a deforming object given a set of monocular images. This problem is naturally under-constrained because there can be many different deformations that produce the same images. By including deformation constraints one limits the set of solutions. Several methods have been proposed in the last decade to tackle NRSfM with a variety of deformation constraints. There are two main categories of methods based on the deformation constraints: statistics-based [27, 14, 5, 10, 12] and physics-based [26, 30, 7, 29, 2] methods. In the former group one assumes that the space of deformations is low-dimensional. These methods are accurate for deformations such as body gestures, facial expressions and simple smooth deformations. However they tend to perform poorly for objects with high-dimensional deformation spaces or atypical deformations. They can also be difficult to use when there is missing data due to *e.g.* occlusions. In the latter group one finds deformation models based on isometry [7, 26, 30, 29], elasticity [1] or particle-interaction models [2]. The isometric model is especially interesting and is an accurate model for a great variety of real objects. In the related problem of template-based reconstruction (also referred to as Shape-from-Template [4]) it has been proven to make the problem well-posed [23, 18, 4, 8]. However in NRSfM, approaches based on isometry still lack in several aspects. For example solutions tend to be complex and often require very good initialization.

To address the shortcomings of state-of-the-art approaches we propose a method with the following properties: *1)* a perspective camera model is used (unlike in low-rank models and few others), *2)* the isometry constraint is used, *3)* a global solution is guaranteed with a convex problem and no initialization (unlike in the recent methods which use gradient-based energy minimization) *4)* we can handle non-smooth surfaces and do not require temporal continuity *5)* we handle missing correspondences and *6)* the complete set of constraints are tied together in a single problem.

We use the inextensibility constraint for approximating isometry. Inextensibility is a relaxation of isometry where one assumes that the Euclidean distances between points on the surface do not exceed their geodesic distances. Inextensibility alone is insufficient because the reconstruction can arbitrarily shrink to the camera's center. In template-based reconstruction inextensibility has been combined with the so-called Maximum-Depth Heuristic (MDH), where one

Corresponding author email: ajad.chhatkuli@gmail.com

maximizes the average depth of the surface subject to inextensibility constraints. This approach has been successfully applied in [23], providing very accurate results for isometrically deforming objects. The main feature of MDH in template-based scenarios is that it can be efficiently solved with convex optimization. However, in NRSfM, the template is unknown and thus MDH cannot be used out-of-the-box. Our main contribution is to show how to solve NRSfM using MDH for isometric deformations. The problem is solved globally with convex optimization, and handles perspective projection and difficult cases such as non-smooth objects and/or deformations, difficult surface topology and large amounts of missing data (*e.g.* 50% or more due to self-occlusions). Furthermore, our solution is far easier to implement than all state-of-the-art methods and has only one free parameter. It can be implemented in MATLAB using only 25 lines of code[1]. We provide extensive experiments where we show that we outperform existing work by a large margin in most cases.

We discuss the state-of-the-art in section 2, and present our problem modeling in section 3, our MDH-based inextensible NRSfM method in section 4, experimental results in section 5 and finally conclusions in section 6.

2. Previous Work

Among the two broad classes of existing methods, factorization-based approaches using the low-rank deformation model have been the focus of research in NRSfM for a long time. Starting from the work of Bregler et al. [5], many works have been proposed to include priors in resolving the ambiguities of factorization-based NRSfM. Priors are important here even after applying the low-rank constraint because some shape ambiguities remain in affine projections [9, 20]. These include the shape basis priors [11], spatial smoothness prior [27] or spatio-temporal smoothness prior and non-linear modeling [14] to name a few. [10] proposed a method to complete NRSfM factorization with only the low-rank prior by improving on the way low rank is imposed in affine projections. Some works have also been done on shape recovery with factorization and perspective camera [15]. Low-rank based factorization methods are global methods that use all available constraints, *i.e*. the image points are concatenated in a matrix which is decomposed to recover all shapes at once. These methods work well with small linear deformations but require learning [25] or prior knowledge to set the number of shape basis, kernel and its parameters [14]. Some improvements have been made for obtaining the basis size automatically but there is no guarantee that a given collection of shapes can be represented by a low number of shape basis accurately. Additionally, in many cases the affine camera has the problem of local two-fold ambiguity [9].

Physical model-based approaches have been explored in the literature to avoid the difficulties and problems with statistical priors. Primarily, efforts have been made on using isometry or its relaxation to inextensibility to constrain the problem in NRSfM [29, 26, 30, 7], which should allow one to handle larger or more complex deformations. Unlike statistical priors, the isometric prior can be fairly accurate for a large variety of deformations. The isometric prior can be used in NRSfM problem locally (point-wise) or semi-locally (patch-wise) or even globally by considering the whole set of surfaces and image points together. A semi-local method using a perspective camera and homographies is proposed in [29]. It can reconstruct surfaces that are composed of large planar patches where it disambiguates surface normals obtained from homography decomposition using smoothness. [7] is a local method that gives point-wise ambiguous solutions for normals which are disambiguated using other views rather than smoothness. However, it requires a smooth surface and very accurate registration represented by splines for computing second-order derivatives of the registration. [26, 9] solved NRSfM locally using the orthographic camera. [26] did this using sets of three points and four or more images with a convex relaxation. [9] did this without a convex relaxation. It used automatically clustered point sets and solved the general case of three or more images. These methods assume a local rigidity prior, which is similar to an isometric prior. [30] uses an orthographic camera and uses the isometric constraints. The method also provides a way to include the perspective camera. It uses discrete non-convex optimization, however, the solutions are not globally optimal and the optimization requires initialization. Furthermore, it is a complex method to implement and test.

Apart from the low rank statistical prior based methods and the isometric prior based methods, some other methods exist. For example, [2] uses a shape basis as well as an isometry-like prior but the method requires an initialization, obtained from rigid factorization on the first set of frames. In that regard, it could be argued that the core of the method is rather like a template-based approach. [21] proposes an interesting local solution based on local fundamental matrices computed from local point sets. However this is a local method that does not use all available constraints and is very complicated to implement. Compared to existing work, our method is the first to formulate a convex problem by relaxing isometry to inextensibility in NRSfM, from which we obtain a globally optimal solution using SOCP.

Notation. We use small-case Latin or Greek alphabets to denote scalars. Bold and small Latin letters denote 2-D vectors and bold and capital Latin letters denote 3-D vectors. Matrices are denoted by capital Latin letters. We use

[1]Optimized code is available at http://isit.u-clermont1.fr/~ab/Research/

$\|.\|_2$ to denote the L2 norm of a vector and $\|.\|_{\text{fro}}$ to denote the Frobenius norm of a matrix. We index points with $i \in \{1 \dots n\}$ where n is the number of scene points, and we index images with $k \in \{1 \dots m\}$ where m is the number of images. We use a subscript to index the points and a superscript to index the images.

3. Modeling

In figure 1 we illustrate the problem and the associated geometric terms described in this section.

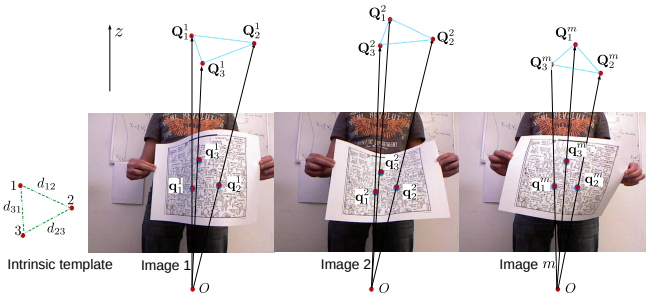

Figure 1: The NRSfM problem and its associated geometric terms. We use O to represent the camera center from which we draw the sight lines. We show only three points for clarity. In practice there can be virtually any number of points and each point can have many neighbours.

3.1. Point-based Reconstruction

We define image measurements as a set of n normalized point correspondences in m images denoted by $\mathcal{C} \triangleq \{\mathbf{q}_i^k\}$. The 2D vector $\mathbf{q}_i^k \triangleq \begin{pmatrix} u_i^k & v_i^k \end{pmatrix}^\top$ denotes the ith point seen in the kth image. We define the unknown set of 3D points by $\mathcal{R} \triangleq \{\mathbf{Q}_i^k\}$, where $\mathbf{Q}_i^k \triangleq \begin{pmatrix} x_i^k & y_i^k & z_i^k \end{pmatrix}^\top$ denotes the unknown 3D position of $\mathbf{q}_i^k$ in camera coordinates. Because we are using the perspective camera, $\mathbf{Q}_i^k$ and $\mathbf{q}_i^k$ are related by

$$\mathbf{Q}_i^k = z_i^k \begin{pmatrix} \mathbf{q}_i^{k\top} & 1 \end{pmatrix}^\top + \epsilon_i^k \tag{1}$$

where ϵ_i^k is measurement noise. The NRSfM problem is solved by determining the unknown set $\mathcal{Z} \triangleq \{z_i^k\}$.

3.2. The Intrinsic Template

We solve $\mathcal{Z}$ using what we call an *intrinsic template*. We use the term intrinsic because it models properties of the surface that are invariant to isometric deformations. The intrinsic template is an undirected graph that links the n scene points through its edges. This is defined by a nearest-neighbourhood graph (NNG) whose edges store the geodesic distances between pairs of points. The NNG is denoted as $\mathcal{N}$ with n points (or *nodes*) and K edges per node. We denote $\mathcal{N}(i)$ as the set of K-neighbours of the ith point. Each edge $e_{ij} \triangleq (i, [\mathcal{N}(i)]_j)$ of the graph has an associated geodesic distance d_{ij}. Because we assume the surface deforms isometrically, we can assume d_{ij} is constant for any deformation. We denote the intrinsic template as the pair $\mathcal{T} \triangleq \{\mathcal{N}, \mathcal{D}\}$, with $\mathcal{D} \triangleq \{d_{ij}\}$.

3.3. Template-Based Reconstruction

In template-based reconstruction (*i.e.* Shape-from-Template), $\mathcal{T}$ is known from the object's reference shape, which is usually built from a geometric mesh. We now describe the MDH for reconstructing an object from a single image. Without loss of generality we assume this is image 1, so the goal is to solve for $\{z_i^1\}$. A solution was first proposed in [19], then solved with convex optimization in [22]. In MDH the deformation model is based on surface inextensibility, which says that the Euclidean distance between any two points $\mathbf{Q}_i^k$ and $\mathbf{Q}_j^k$ is upper bounded by the geodesic distance d_{ij}. For simplicity we neglect the effect of the measurement noise ϵ_i^k as in [22]. The problem formulation is as follows:

$$\begin{aligned} &\underset{\{z_i^1\}}{\text{argmax}} \sum_{i=1}^{n} z_i^1, \\ &\text{s.t.} \quad \forall i \in \{1 \dots n\},\ j \in \mathcal{N}(i) \\ &z_i^1 \geq 0 \\ &\left\| z_i^1 \begin{bmatrix} \mathbf{q}_i^1 \\ 1 \end{bmatrix} - z_j^1 \begin{bmatrix} \mathbf{q}_j^1 \\ 1 \end{bmatrix} \right\|_2 \leq d_{ij}. \end{aligned} \tag{2}$$

The main properties of problem (2) are the following. *1)* It is a Second Order Cone Program (SOCP) that can be solved efficiently and globally with modern optimization tools such as MOSEK and SeDuMi. *2)* The neighbour order K in the intrinsic template can be any. A larger K introduces more cone constraints, however it also significantly increases the computational time. Keeping a lower K is thus important for efficiency purposes.

4. MDH-based NRSfM

4.1. Initial Formulation

The MDH for NRSfM can be expressed as the maximization of the sum of all depths $\{z_i^k\}$ under the inextensibility constraint and the condition that each depth and each distance are positive. Unlike in template-based reconstruction, we require multiple images and in general point correspondences will not be found in all images due to occlusions, missed tracks in optical flow, etc. We therefore introduce the visibility set $\mathcal{V} \triangleq \{v_i^k\}$, where $v_i^k = 1$ if the ith point is visible in the kth image and $v_i^k = 0$ otherwise. We

formulate the problem as follows:

$$\begin{aligned} &\underset{\{z_i^k\},\{d_{ij}\}}{\operatorname{argmax}} \sum_{k=1}^{m}\sum_{i=1}^{n} v_i^k z_i^k, \\ &\text{s.t.} \qquad \forall k \in \{1\dots m\},\ i \in \{1\dots n\},\ j \in \mathcal{N}(i) \\ &z_i^k \geq 0, \quad d_{ij} \geq 0, \\ &v_i^k v_j^k \left\| z_i^k \begin{bmatrix} \mathbf{q}_i^k \\ 1 \end{bmatrix} - z_j^k \begin{bmatrix} \mathbf{q}_j^k \\ 1 \end{bmatrix} \right\|_2 \leq v_i^k v_j^k d_{ij}. \end{aligned} \tag{3}$$

To handle missing correspondences, we fix $z_i^k = 0$ if $v_i^k = 0$ and therefore we do not reconstruct the points that are not visible. The visibility variables are used in problem (2) to disconnect the inextensibility conditions when any of the points involved is not visible. In contrast to the template-based problem (2), in the template-less problem (3) we do not know the intrinsic template $\mathcal{T}$. It is clear that solving problem (3) directly is not possible for two reasons: *1)* the optimization is not well posed because d_{ij} is unbounded (one can keep increasing d_{ij} and the constraints will still be satisfied), *2)* the NNG is an unknown. We now give the solutions to both issues.

4.2. Bounding the Distances

In order to bound the problem, our idea is to fix the scale of the intrinsic template, by fixing the sum of the geodesic distances to a positive scalar (1 in our case). Formally we include in problem (3) the following linear constraint:

$$\sum_{i=1}^{n}\sum_{j\in\mathcal{N}(i)} d_{ij} = 1. \tag{4}$$

By including equation (4), $\{z_i^k\}$ cannot increase indefinitely without violating equation (4), yet the problem is still an SOCP. We illustrate this in figure 2. The effect of equation (4) is to fix the scale of the reconstruction. In NRSfM we are free to fix the scale of the reconstruction arbitrarily, because just like in rigid SfM, it is never recoverable. Having fixed the scale, the reconstructed depths cannot increase arbitrarily, because with a perspective camera as the depths increase so do Euclidean distances between pairs of points. At some point, the Euclidean distances will exceed the geodesic distances and the inextensibility constraints (last line of problem (3)) will be violated.

4.3. The Nearest-Neighbour Graph

The function of the NNG is to constrain the depths between pairs of points on the object's surface (problem (3), last line). These pairs can be any pairs of points, however they give the strongest constraints when the points are close together on the surface. This is because for closer points the inextensibility inequalities become tighter. Of course, we do not know exactly which points are close together *a priori*. A good estimate can be made from the distance of the correspondences in the images, because nearby points on the surface tend to be close in the images. We denote the Euclidean distance between two points $\mathbf{q}_i^k$ and $\mathbf{q}_j^k$ in image k by δ_{ij}^k, and we use these to build the NNG. The specific algorithm we propose is as follows:

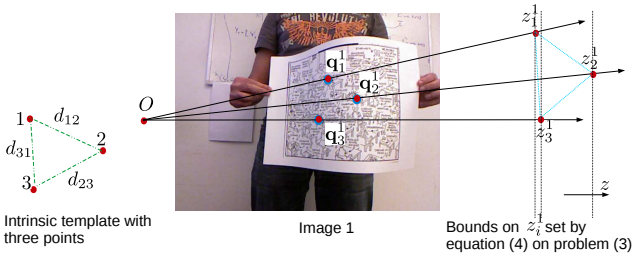

Figure 2: Illustration of the bounds set by equation (4) for NRSfM using three points and one image.

1. Compute distances $\{\delta_{ij}^k\} \quad \forall i \in \{1\dots n\}$, $j \in \{1\dots n\},\ k \in \{1\dots m\}$, and $i \neq j$.
2. If the ith or jth point is not visible in image k, set: $\delta_{ij}^k = -\infty$.
3. Take the maximum distance over the images. $\hat{\delta}_{ij} = \max_k\{\delta_{ij}^k\} \quad \forall i \in \{1\dots n\},\ j \in \{1\dots n\}$.
4. For each point i put into $\mathcal{N}(i)$ the points j with the K smallest values of $\hat{\delta}_{ij}$ $(j \neq i)$.

The only parameter that needs to be selected here is the neighbourhood size K. Our method is not very sensitive to this parameter but a reasonable value (*e.g.*, 20) should be chosen depending on the density of the correspondences and required speed of optimization.

4.4. Implementation Details

We have implemented two versions of our method in MATLAB which uses the MOSEK [3] SOCP solver. MOSEK is faster than many other SOCP solvers, especially for large scale problems. The first version is only 25 lines and uses YALMIP to translate symbolic variables. The second version is longer and does not involve the translation, which can be expensive for large problems. In practice we use the second version because it is significantly faster. For example, we can solve with 50 images, 1000 points and $K = 20$ in about 1 minute in a standard desktop PC. This computation time is the fastest among the compared methods for the number of images and points considered.

5. Experimental Results

5.1. Method Comparison and Error Metrics

We compare our results against five other methods whose source code is provided by the authors. We name our

method as **tlmdh**. We name the non-convex soft inextensibility based method for orthographic camera [30] as **o-sinext** and the local homography method for perspective camera [7] as **p-isolh**. We name the prior free factorization method of [10] as **o-spfac** and the kernel based factorization method [13] as **o-kfac**. We name the locally rigid method based on 3-point SfM [26] as **o-lrigid**. Each method requires one or more parameters to be tuned. We fix these parameters to optimal values for each dataset and keep them constant for all experiments.

We measure a method's accuracy with two metrics: 3D Root Mean Square Error (RMSE) and the normal error. The 3D RMSE is computed from the ground truth 3D point positions. Because NRSfM has a scale ambiguity no method can reconstruct the absolute scale of the object. For methods which use perspective camera (**tlmdh** and **p-isolh**) we scale their reconstructions to best align the with the ground truth. For the methods which use affine cameras (**o-sinext**, **o-lrigid** and **o-spfac**), we transform their reconstructions with a similarity transform to best align them with the ground truth. The normal error is computed by measuring the difference between the ground truth surface normal at each point and the reconstructed normals. We compute the normals by fitting a B-spline and measuring the normals from the B-spline coefficients.

5.2. Developable Surfaces

Most non-rigid reconstruction methods focus on developable surfaces for experiments. A developable surface can be flattened into a planar surface without tearing or stretching, such as a piece of paper. Obtaining continuous tracks of correspondences without partial images is relatively easy for such surfaces. While the surfaces often appear simple, they sometimes have high frequency and non-linear deformations. We experiment with 4 different public datasets representing such surfaces.

The Flag dataset. We use the cloth capture data (mocap) [31] to generate semi-synthetic data. Even though the object is real, the input data for all the methods are generated from a virtual camera with perspective projection. The data shows a flag waving with wind with some changes in the camera viewpoint, making it perhaps the simplest of all datasets. The images are generated with dimensions 640 px $\times$ 480 px using a camera focal length of 640 px. The data has altogether 450 frames. We use this data to test the performance of our method and the competitive methods in several practical scenarios: with changing number of images, changing number of corresponding points and missing correspondences. For changing the number of images, we randomly draw a subset of m images from the 450 images with m varying from 5 to 60. For varying the number of points, we randomly select a subset of n points varying

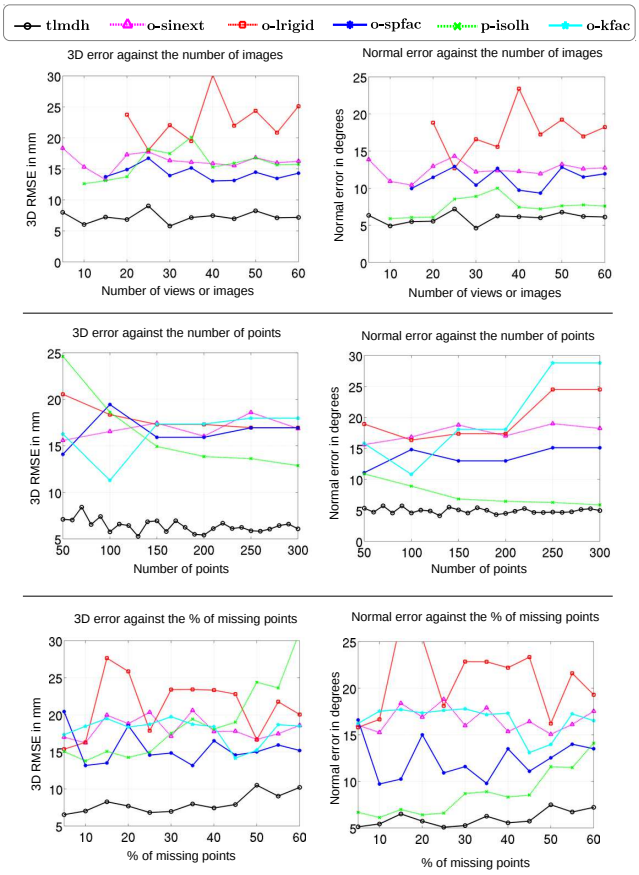

Figure 3: Plots for synthetic Flag dataset. The 3D errors shown in the left column and the normal errors in the right column. Legends are shown on the top.

from 50 to 300. Finally, for varying the amount of missing correspondences for each image we randomly remove a percentage of correspondences ranging from 5 to 60. For the default conditions, we use 40 images, 300 points and no missing data. In order to fill the missing correspondences required by some methods we follow [16] for matrix completion. Note that our method **tlmdh** works with incomplete data and therefore we do not complete missing correspondences for our method. **p-isolh** computes registration functions with B-splines and so we use them to fill in the missing correspondences for that method. Figure 3 shows the plots for the dataset. The results show that our method **tlmdh** performs very well with just 5 images and considerably better than all other methods. The factorization-based method **o-spfac** and the local homography based method **p-isolh** also does better compared to other methods. We obtain an RMSE 3D error of 6.3 mm using 40 images. Similarly, it can be seen that our method is able to reconstruct the surface with as many as 60% random missing data.

The KINECT Paper dataset. We use the KINECT Paper dataset [28] as one of our real datasets for evaluation, originally used for template-based reconstruction [18]. The dataset shows a VGA resolution sequence of a large piece of textured paper undergoing smooth deformations.We generate correspondences by tracking points in the sequence using an optical flow-based method [12] designed for non-rigid surfaces. The tracks are outlier free and semi-dense. Due to the large number of frames we again subsample them for all methods except **o-kfac**, which requires temporal continuity. Figure 4 shows the plots against the number of images for the rest of the methods. We obtain very accurate reconstructions that in fact compares with template-based reconstructions [18, 8].

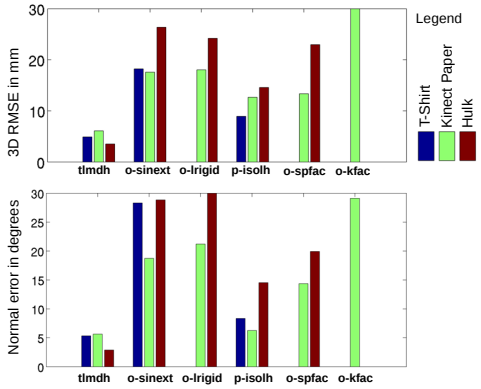

Figure 4: Mean 3D errors for the real developable surfaces.

The Hulk and the T-Shirt dataset. The Hulk dataset [7] consists of a comic cover printed on a piece of paper in 21 different deformations. Similarly, the T-Shirt dataset [7] consists of a textured T-Shirt with 10 different deformations. These datasets provide images with wide-baseline matches. We do not test the factorization-based methods on these datasets as they have very few images and also do not form a temporal sequence. Large number of images $(m > 3/2L)$, where L is the number of shape basis here, are required by **o-spfac** and a continuous video sequence is required by **o-kfac**. We show the results of different methods in the bar plot of figure 4. We obtain a mean depth error of 3.5 mm in the Hulk dataset and 4.9 mm in the T-shirt dataset. The next best performing method is **p-isolh** that gives a mean depth error of 14.53 mm and 8.94 mm for the Hulk and T-shirt datasets respectively. Similarly we obtain a mean depth error of 22.98 mm for **o-spfac** in the Hulk dataset. We do not obtain good results with **o-lrigid** and **o-sinext** in these datasets.

Failure cases. Failure cases occur in NRSfM due to the problem being ill-posed due to lack of motion and deformation. Naturally any method would fail when the problem is ill-posed. However, a method can also fail to give good results with a well-posed problem. We found one such example for our method from [24]. The dataset is a bending piece of paper imaged from a fixed camera viewpoint with a relatively longer focal length, and it contains no ground truth. We use optical flow [6] to obtain correspondences. The qualitative reconstructions for three frames are shown in figure 5. The general shape of the paper looks reasonable but in the first image it is bent when it should be flat and the degree of bending is not properly captured in the second image. We know that better reconstructions are possible on this dataset [30], so the problem is not itself ill-posed. The imperfect reconstruction from our method is probably caused by the lack of change in camera viewpoint.

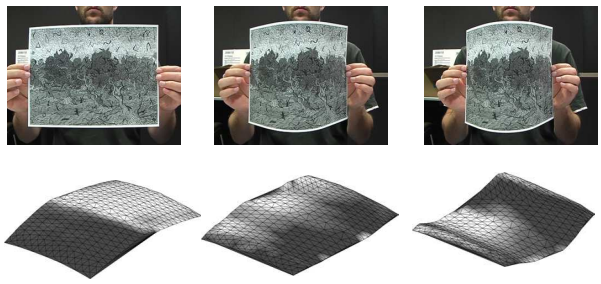

Figure 5: Failure cases: Images (top row) and their respective reconstructions (bottom row). The first two shapes appear largely incorrect.

5.3. Non-Developable Objects

We use two different datasets to perform NRSfM on non-developable surfaces. They are complex objects where some of the compared methods are not even applicable, for example, **p-isolh** requires registration warps, which is non-trivial to implement in volumetric objects. We perform experiments here to show what we can obtain in highly difficult non-rigid reconstruction applications. Below we describe the datasets and the experiments performed.

The Stepping Trousers dataset. The dataset [31] is constructed from motion capture ground truth data with perspective projection. The data shows a pair of trousers stepping around with considerable rapid deformations of the cloth. The images are obtained at a resolution of 640 px $\times$ 480 px with a perspective camera of focal length 320 px. The dataset is semi-synthetic but due to articulations, volume/partial views and rapid nonlinear deformations, it is arguably the most complex data used for NRSfM to date. Unlike the flag dataset, missing correspondences are significant due to self-occlusions. The missing correspondences are handled by filling in the correspondences using [16] for all methods except ours. Figure 6 shows three reconstructed frames. From top to bottom, it shows our best reconstruc-

tion, a reconstruction with medium accuracy and our worst reconstruction. Alongside we show the reconstructions for the compared method **o-spfac**. Note that it is non-trivial to implement the compared methods in the missing data scenario without using a low-rank prior. Thus we only test the best performing low-rank method **o-spfac**. The plots of 3D error for each image for these two methods are shown in figure 7. Because this is a large object, the 3D RMSE error can be large, yet the reconstructions can appear reasonable. We therefore also measure accuracy with a relative 3D reconstruction error, which is defined as follows:

$$\%\ \text{3D error} = \frac{\|\mathbf{P}_{GT} - \mathbf{P}_{REC}\|_{\text{fro}}}{\|\mathbf{P}_{GT}\|_{\text{fro}}} \tag{5}$$

where $\mathbf{P}_{GT}$ represents the ground truth 3D shape ($3 \times n$ matrix) and $\mathbf{P}_{REC}$ represents the reconstructed 3D shape. We obtain a mean 3D error of 22.54 mm and % 3D error of 2.37% for our method while for **o-spfac** those are 51.5 mm and 11.56% respectively. Our results indeed show that large objects with complex deformations in small scale can be reconstructed with our method, although some difficulties can be seen primarily due to high surface curvature. The reconstructions and the plot show that our method can capture a large portion of the deformations correctly even though the parts of the object undergoing deformation are very small in the image, making the projections almost affine. In certain cases, however, it estimates the shapes incorrectly on those parts as shown in the third reconstruction of the sequence in figure 6.

The hand dataset. In tasks such as gesture recognition, several applications require reconstructing a moving hand. When such a task is done, usually a specialized modeling of hand motion and its articulations is used. We show that an accurate reconstruction of a deforming hand can be done solely with the inextensibility prior using our method. We test with two sequences of a deforming hand recorded by an endoscopic camera. The camera images are of dimensions 960×540 px, taken with a focal length of 462 px and capture detailed texture. We obtain ground truth reconstructions of the first and last frame using stereo and post processing. We compute correspondences by densely tracking the hand's texture using [6]. Note that the correspondences are not perfect due to image noise and weak texture. Because most methods cannot handle a huge number of points, we uniformly subsample to 1000 points. Figure 8 shows reconstructions of the hand compared to ground truth for our method, **o-spfac** and **p-isolh** (which were the best performing state-of-the-art methods). The results show that our method can handle complex deformations of a hand. Both the compared methods were unable to capture the second deformation where they gave rather planar or smooth surfaces with 3D error of over 60 mm. On the other hand we

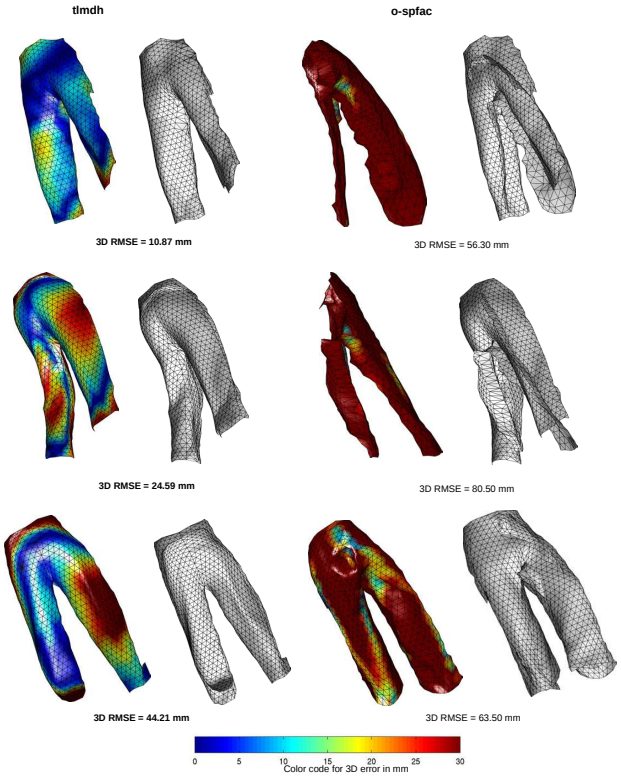

Figure 6: Reconstructions of the stepping trousers dataset for our method and **o-spfac**. Top row shows the reconstructed meshes overlaid on top of the ground truth. Bottom row shows the reconstructed mesh texture mapped with 3D error for each face in the color code shown. Note that we show our best result in the first column and the worst in the last column with a medium accuracy result in the middle.

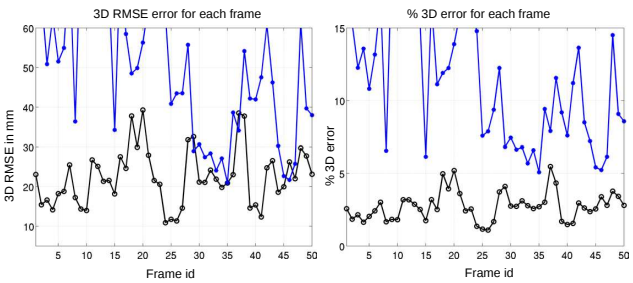

Figure 7: Plot of the depth error in trousers for each sampled image (legend in figure 3).

obtain a slightly higher 3D error of 7.38 mm in the third column.

5.4. NRSfM with Rigid Objects

All rigid objects are isometric, therefore our NRSfM method can be used to reconstruct rigid scenes. However isometry is weaker than rigidity, so it can be expected to perform slightly worse. Nonetheless it is interesting to study such cases for two reasons. First our method gives a convex

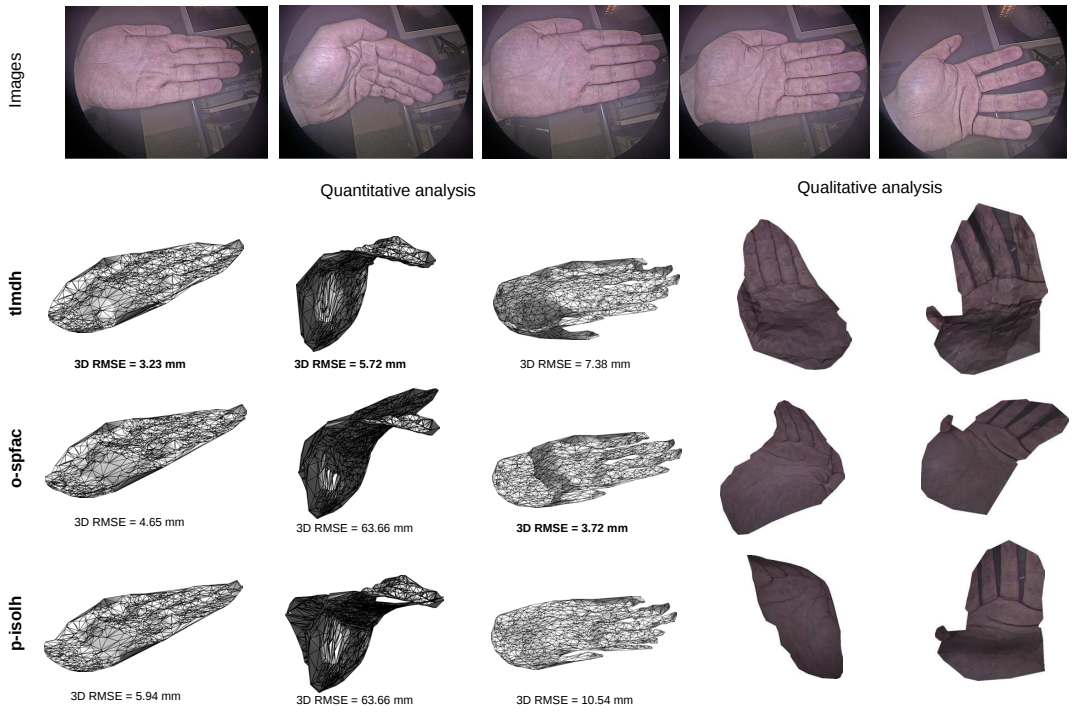

Figure 8: Results on the hand dataset. We use the best performing methods in other datasets for comparison: **o-spfac** and **p-isolh**. Ground truth is shown for three images, overlaid on top of the reconstructions. We texture map the meshes and show qualitative results for the two other images where ground truth 3D is not available.

solution to the problem with a general number of images, which has not been seen before in rigid SfM with perspective cameras. It may therefore find uses for initialising rigid bundle adjustment. The second reason is for a theoretical understanding of our method using rigid scenes, which may be simpler to analyse than for deformable scenes. For example, it may be interesting to study the critical motions associated with the inextensibility relaxation. We show some results from the public dataset [17] on the house sequence using SIFT correspondences. We plot the average % 3D error for each of the 49 images for our method and compare this to a state-of-the-art rigid SfM method (VisualSfM [32]). We see that a reasonable error is obtained for the majority of the images.

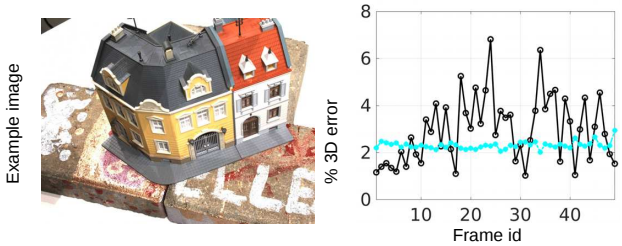

Figure 9: Results on rigid scenes. VisualSfM results are shown in cyan dots.

6. Conclusions

We have brought forward the MDH-based formulation, which has enjoyed great success in inextensible template-based reconstruction, to the more general problem of templateless non-rigid reconstruction known as NRSfM. We have shown that this leads to a convex formulation, which can be solved globally and optimally as an SOCP problem. This forms the first convex, global and optimal NRSfM formulation based on physical constraints. Results on synthetic and real images have shown a great promise and our method outperforms existing ones by a large margin in many cases. In future work, we plan to study the inclusion of outliers in our formulation using slack variables and a theoretical study of the problem's conditioning.

Acknowledgements. This research has received funding from the EUs FP7 through the ERC research grant 307483 FLEXABLE. The work has also been supported by the Spanish Ministry of Economy and Competitiveness under project SPACES-UAH (TIN2013-47630-C2-1-R), and by the University of Alcalá under project ARMIS (CCG2015/EXP-054). The work is also supported by Almerys Corporation.

References

[1] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *CVPR*, 2014.

[2] A. Agudo and F. Moreno-Noguer. Simultaneous pose and non-rigid shape with particle dynamics. In *CVPR*, 2015.

[3] M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).*, 2015.

[4] A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro. Shape-from-template. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(10):2099–2118, 2015.

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

[6] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004.

[7] A. Chhatkuli, D. Pizarro, and A. Bartoli. Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In *BMVC*, 2014.

[8] A. Chhatkuli, D. Pizarro, and A. Bartoli. Stable template-based isometric 3D reconstruction in all imaging conditions by linear least-squares. In *CVPR*, 2014.

[9] T. Collins and A. Bartoli. Locally affine and planar deformable surface reconstruction from video. In *International Workshop on Vision, Modeling and Visualization*, 2010.

[10] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. In *CVPR*, 2012.

[11] A. Del Bue. A factorization approach to structure from motion with shape priors. In *CVPR*, 2008.

[12] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, 2013.

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

[14] P. F. U. Gotardo and A. M. Martínez. Kernel non-rigid structure from motion. In *ICCV*, 2011.

[15] R. Hartley and R. Vidal. Perspective nonrigid shape and motion recovery. In *ECCV*, 2008.

[16] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2117–2130, 2013.

[17] R. R. Jensen, A. L. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014.

[18] J. O. M. Östlund, A. Varol, T. D. Ngo, and P. Fua. Laplacian Meshes for Monocular 3D Shape Recovery. In *ECCV*, 2012.

[19] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. In *BMVC*, 2008.

[20] D. Pizarro, A. Bartoli, and T. Collins. Isowarp and conwarp: Warps that exactly comply with weak-perspective projection of deforming objects. In *BMVC*, 2013.

[21] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *ECCV*, 2014.

[22] M. Salzmann and P. Fua. Reconstructing sharply folding surfaces: A convex formulation. In *CVPR*, 2009.

[23] M. Salzmann and P. Fua. Linear local models for monocular reconstruction of deformable surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):931–944, 2011.

[24] M. Salzmann, R. Hartley, and P. Fua. Convex optimization for deformable surface 3-D tracking. In *ICCV*, 2007.

[25] L. Tao and B. J. Matuszewski. Non-rigid structure from motion with diffusion maps prior. In *CVPR*, 2013.

[26] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *CVPR*, 2010.

[27] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):878–892, 2008.

[28] A. Varol, M. Salzmann, P. Fua, and R. Urtasun. A constrained latent variable model. In *CVPR*, 2012.

[29] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-free monocular reconstruction of deformable surfaces. In *CVPR*, 2009.

[30] S. Vicente and L. Agapito. Soft inextensibility constraints for template-free non-rigid reconstruction. In *ECCV*, 2012.

[31] R. White, K. Crane, and D. Forsyth. Capturing and animating occluded cloth. In *SIGGRAPH*, 2007.

[32] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013.