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

Ajad Chhatkuli[1], Daniel Pizarro[2,1], Toby Collins[1] and Adrien Bartoli[1]

[1]Institut Pascal - CNRS/Université Clermont Auvergne, Clermont-Ferrand, France
[2]GEINTRA, Universidad de Alcalá, Alcalá de Henares, Spain

Abstract—We present a global and convex formulation for the templateless 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 Structure-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, which do not perform well with complex nonlinear deformations. 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 can be solved with SOCP. This is the first convex formulation for NRSfM with physical constraints. We further present how robustness and temporal continuity can be included in the formulation to handle outliers and decrease the problem size, respectively. We show with extensive experiments that our methods accurately reconstruct quasi-isometric objects from partial views under articulated and strong deformations. Compared to the previous methods, our approach gives better or similar accuracy. 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).

Code release. We have made our MATLAB implementation available at http://igt.ip.uca.fr/~ab/.

1 Introduction

Non-Rigid Structure-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 underconstrained 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 to tackle NRSfM with a variety of deformation constraints. There are two main categories of methods based on the deformation constraints: statistics-based [Bregler et al., 2000; Dai et al., 2012; Garg et al., 2013; Gotardo and Martínez, 2011; Torresani et al., 2008] and physical model-based [Agudo and Moreno-Noguer, 2015; Chhatkuli et al., 2014b; Taylor et al., 2010; Varol et al., 2009; Vicente and Agapito, 2012] 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 *e.g.*, due to occlusions. In the latter group one finds deformation models based on isometry [Chhatkuli et al., 2014b; Taylor et al., 2010; Varol et al., 2009; Vicente and Agapito, 2012], elasticity [Agudo et al., 2014] or particle-interaction models [Agudo and Moreno-Noguer, 2015]. The isometric model is especially interesting and is an accurate model for a great variety of real object deformations. In the related problem of template-based reconstruction (also referred to as Shape-from-Template [Bartoli et al., 2015]) it has been proven to make the problem well-posed [Bartoli et al., 2015; Chhatkuli et al., 2014a; Ngo et al., 2016; Salzmann and Fua, 2011]. However in NRSfM, approaches based on isometry still lack in several aspects. In particular, the existing solution methods tend to be complex in their design 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)* the perspective camera model is used (unlike in most low-rank model methods 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 energy minimization) *4)* it handles non-smooth objects and does not require temporal continuity *5)* it handles 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) [Perriollat et al., 2011; Salzmann and Fua, 2011], where one maximizes the average depth of the surface subject to inextensibility constraints. This approach has been successfully applied

. Corresponding author email: ajad.chhatkuli@vision.ee.ethz.ch in [Salzmann and Fua, 2011], 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 that we show how to solve NRSfM using MDH for isometric deformations. The problem is solved globally with convex optimization (SOCP), 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). Figure 1 shows the reconstructions obtained from our method for a deforming piece of paper. Our solution is far easier to implement than all state-of-the-art methods and has only one free parameter. The parameter value is not critical and a higher value only translates to a larger problem size but no reduction in solution accuracy. The proposed method can be implemented in MATLAB using only 25 lines of code. We also provide a robust formulation of our method that can handle noisy and erroneous image correspondences. To encode temporal smoothness we represent the depth function as a one-dimensional spline. We design all proposed methods to be SOCP problems so that they can be solved very efficiently and optimally by off-the-shelf solvers. We provide extensive experiments where we show that we outperform existing work by a large margin in most cases. Additionally, inextensibility is also a convex relaxation of rigidity. With this, we can express a rigid SfM problem as a single SOCP. Although for obvious reasons, it cannot solve rigid SfM with the same accuracy as conventional approaches, we show an experiment which proves that our method also generalizes to rigid scenes. A related approach [Li, 2010] uses preservation of Euclidean distance in rigid objects to formulate a Semi-Definite Program (SDP) and solves for a single rigid object without explicitly modeling motion. We differ from this approach by considering the fact that Euclidean distances between 3D points in non-rigid objects are not preserved with deformations but are upper-bounded by the geodesic distances.

This paper represents an extension of our previous work [Chhatkuli et al., 2016] where we presented the global convex formulation using MDH. We here extend the formulation in two ways: one having robustness embedded into the formulation and the other by adding the temporal smoothness prior based on splines. We also present new experiments on additional objects. We organize the paper as follows. 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 and experimental results in section 6. We discuss on the practical aspects of the proposed methods in section 7 and finally conclude in section 8.

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., 2000], many works have been proposed to include priors in resolving the ambiguities of factorization-based NRSfM. Priors are important even after applying the low-rank constraint because some shape ambiguities remain in affine projections [Collins and Bartoli, 2010; Pizarro et al., 2013]. These include the shape basis priors [Del Bue, 2008], spatial smoothness prior [Torresani et al., 2008] or spatio-temporal smoothness prior and non-linear modeling [Gotardo and Martínez, 2011] to name a few. [Dai et al., 2012] 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 the perspective camera [Hartley and Vidal, 2008]. Low-rank based factorization methods are global methods that use all the 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 [Tao and Matuszewski, 2013] or prior knowledge to set the number of shape bases, kernel and its parameters [Gotardo and Martínez, 2011]. Some improvements have been made for obtaining the basis size automatically [Garg et al., 2013] but there is no guarantee that a given collection of shapes can be represented by a low number of shape bases accurately. Additionally, in many cases the affine camera has the problem of local two-fold ambiguity [Collins and Bartoli, 2010].

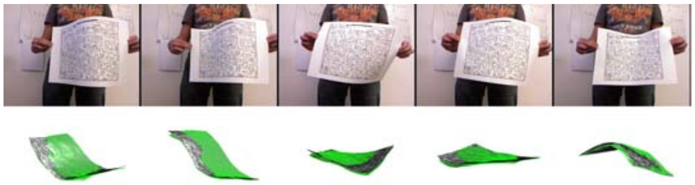

Figure 1: Example reconstructions with our method on the KINECT Paper [Varol et al., 2012a] images. The top row shows the input images and the bottom row shows the groundtruth in green overlaid on top of the reconstruction in white. Our best method gives a 3D error of **4.62** mm while the best compared method [Parashar et al., 2016] has an error of 7.63 mm. This is remarkable if we note that even the best performing SfT method in [Chhatkuli et al., 2017] produces an error of **3.82** mm on the dataset.

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 [Chhatkuli et al., 2014b; Taylor et al., 2010; Varol et al., 2009; Vicente and Agapito, 2012], 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 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 [Varol et al., 2009]. It can reconstruct surfaces that are composed of large planar patches where it disambiguates surface normals obtained from homography decomposition using smoothness. [Chhatkuli et al., 2014b] is a local method that assumes surfaces to be only locally planar at each point. It gives point-wise ambiguous so-

TABLE 1: NRSfM methods and their characteristics.

Methods	Surface Representation	Surface Prior	Camera Model	Constraint type	Primary computation
[Gotardo and Martínez, 2011]	Point sets	Low-rank and temporal smoothness	Orthographic	Global	Non convex
[Dai et al., 2012]	Point sets	Low-rank	Orthographic	Global	Convex with non-convex refinement
[Taylor et al., 2010]	Mesh	Isometry	Orthographic	Local	Small systems
[Vicente and Agapito, 2012]	Point sets with neighborhood	Isometry	Orthographic and perspective	Global	Non-convex
[Parashar et al., 2016]	2D Riemannian Manifold	Isometry	Perspective	Local	Small quartic systems
[Chhatkuli et al., 2014b]	2D Riemannian Manifold (implicit)	Isometry	Perspective	Local	Small systems
Proposed method	Point sets with neighborhood	Inextensibility	Perspective	Global	Convex

lutions for normals which are disambiguated using other views rather than smoothness. The 3D shape is then obtained by surface integration of the normals. However, it only works for smooth surfaces and requires very accurate registration represented by splines for computing second-order derivatives of the registration. A recent local solution for NRSfM [Parashar et al., 2016] gives a much better way to obtain surface normals using local planarity at each point. One remarkable feature of the method is the fact that the computational complexity, which comes from solving a local quartic system, is largely independent of the number of images. [Collins and Bartoli, 2010; Taylor et al., 2010] solved NRSfM locally using the orthographic camera. [Taylor et al., 2010] did this using sets of three points and four or more images with a convex relaxation. [Collins and Bartoli, 2010] 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. [Vicente and Agapito, 2012] uses the isometric constraints under the assumption of an orthographic camera. The method also provides a way to include the perspective camera. However, the solutions are obtained with discrete non-convex optimization on an initial solution and are not globally optimal. Furthermore, it is a complex method to implement and test. Table 1 lists some important methods and their characteristics in comparison to the proposed methods.

Apart from the low rank statistical prior based methods and the isometric prior based methods, some other methods exist. For example, [Agudo and Moreno-Noguer, 2015] uses a shape basis as well as an isometry-like prior but the method requires an initial 3D shape, 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. [Russell et al., 2014] 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. Our method is fast, accurate, simple to understand and uses the perspective camera model.

3 Modeling

In figure 2, we illustrate the problem and the associated geometric terms described in this section. We use Latin and Greek letters in italics to denote scalars. Bold and lower case Latin and Greek letters denote vectors. Matrices are denoted by bold upper case Latin letters. We use a Greek letter to emphasize that a given quantity is a function. We use $\|.\|_2$ to denote the L2 norm of a vector and $\|.\|_{\text{fro}}$ to denote the Frobenius norm of a matrix. We index points with $i \in \{1 \ldots n\}$ where n is the number of scene points, and we index images with $k \in \{1 \ldots m\}$ where m is the number of images. We use a subscript to index the points and a superscript to index the images.

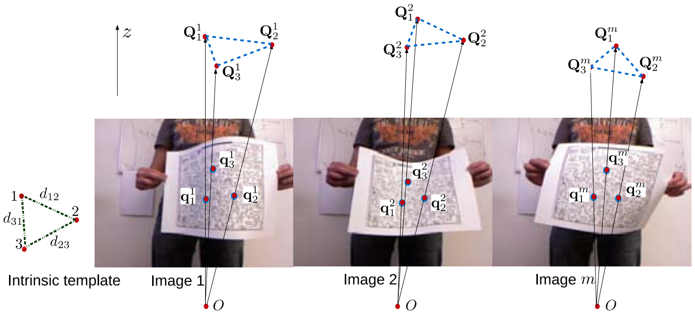

Figure 2: The NRSfM problem and its associated geometric terms. We use **O** to represent the camera center from which we draw the sight lines. We show only three points for clarity. In practice there can be virtually any number of points and each point can have many neighbours.

3.1 Point-based reconstruction

We define image measurements as a set of n point correspondences expressed in the camera frame 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 $\mathcal{R} \triangleq \{\mathbf{p}_i^k\}$, where $\mathbf{p}_i^k \triangleq \begin{pmatrix} x_i^k & y_i^k & z_i^k \end{pmatrix}^\top$ denotes the unknown 3D position of

$\mathbf{q}_i^k$ in camera coordinates. Because we use the perspective camera, $\mathbf{p}_i^k$ and $\mathbf{q}_i^k$ are related by

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

where $\mathbf{e}_i^k$ is measurement noise. We do not explicitly parametrize the camera motion in our model. This frees the method from dealing with the ambiguities between the camera motion and the object deformation. The NRSfM problem is solved by determining the unknown set $\mathcal{Z} \triangleq \{z_i^k\}$.

3.2 The intrinsic template

We start with the MDH-based SfT problem and then migrate to NRSfM. We formalize the 3D template with what we call the *intrinsic template*. This is used to solve the set of point depths $\mathcal{Z}$. We use the term intrinsic because it models properties of the object 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 object 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

MDH for reconstructing a deformable surface was first proposed in the template-based scenario. We therefore first describe template-based reconstruction with MDH and then move to the generic NRSfM problem. 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 [Perriollat et al., 2008], then solved with convex optimization in [Salzmann and Fua, 2009]. In MDH the deformation model is based on surface inextensibility, which says that the Euclidean distance between any two points $\mathbf{p}_i^k$ and $\mathbf{p}_j^k$ is upper bounded by the geodesic distance d_{ij}. The geodesic distance d_{ij} and the NNG $\mathcal{N}$ can be computed easily as the template shape is given. For simplicity we neglect the effect of the measurement noise $\mathbf{e}_i^k$ as in [Salzmann and Fua, 2011]. The problem formulation is as follows:

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

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 is non-critical and can be a number greater than or equal to 2, $K \geq 2$, since each edge provides one inequality. Having $K = 2$ translates to slightly more constraints than variables. In practice, it is better to keep $K > 2$ for each point because we have inequalities rather than equalities. A very large value of K, however implies that inextensibility constraints between distant points will be included in problem (2). Such constraints between distant points do not strongly constrain the problem and including them only amounts to an increase in the computation 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, it uses 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 assume the visibility set to be known, meaning that we know which points are missing in each image. We formulate the problem as follows:

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

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 known visibility set is 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 a solution 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 an arbitrary 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 3. 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 (final constraint of problem (3)) will be violated.

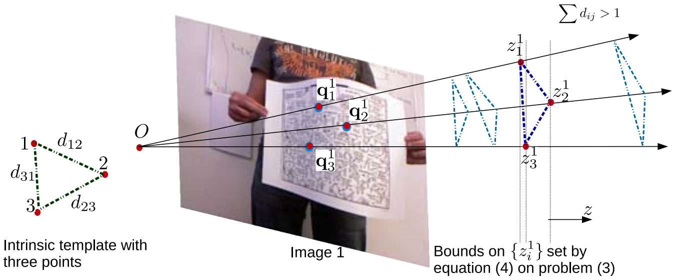

Figure 3: Illustration of the bounds set by equation (4) for NRSfM using three points and one image. The depth values cannot increase to the shaded region on the right because this would violate equation (4).

4.3 The nearest-neighbour graph

The function of the NNG is to select pairs of points on the object's surface which give strong inextensibility constraints. 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 object's 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:

1) Compute distances $\{\delta_{ij}^k\} \quad \forall i \in \{1 \dots n\}, \quad j \in \{1 \dots n\}, \ k \in \{1 \dots m\}$, 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 augment $\mathcal{N}(i)$ with the points j with the K smallest values of $\hat{\delta}_{ij}$ $(j \neq i)$.
5) Find the connected components using each point index i and its neighborhood $\mathcal{N}(i)$ and reconstruct each component separately.

The above algorithm keeps only those points in a neighborhood that are close to each other in all the images. This implies that if a material is torn apart or an object splits, we treat the parts as separate objects. In that case, they could be reconstructed separately and the scale could be fixed after the reconstruction to merge them in images where they form a single object. The only parameter that needs to be selected here is the neighbourhood size K. As explained in the end of section 3.3, 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 NRSfM with temporal smoothness

One potential application of NRSfM is to reconstruct a deforming object from its video. In such a setup, the object points can be assumed to move smoothly over time. This can be expressed by replacing the maximization term in problem (3) with the following:

$$\underset{\{z_i^k\},\{d_{ij}\}}{\text{maximize}} \sum_{k=1}^{m}\sum_{i=1}^{n} v_i^k z_i^k - \lambda_t \sum_{k=1}^{m-1}\sum_{i=1}^{n} \| v_i^{k+1} v_i^k (z_i^{k+1} - z_i^k) \|_1 \tag{5}$$

subject to the same constraints as in problem (3). We use the hyperparameter $\lambda_t \in \mathbb{R}$ to balance the two costs of problem (5). The added term in problem (5) causes the depth values to change slowly between consecutive views, albeit with an added computational complexity. The added complexity comes from the use of slack variables required for implementing the L1 cost. Many methods including [Vicente and Agapito, 2012] use such first-order approach to impose temporal smoothness. However, using a large number of images (say, greater than 100) can increase the size of problem (3) making it very time consuming to solve. Using the formulation of problem (5) can make it possibly intractable in such situation. We introduce a different approach to impose temporal smoothness that attempts on reduction of the size of problem (3). We define temporal smoothness as the smooth evolution of depth over time and use uniform cubic B-splines to represent depth as a function of time. Thus for each 3D point over the time sequence, the unknown variables are the set of control points representing the evolution of depth in the sequence.

B-splines can be used to parametrize an N-D function using weighting parameters known as the control points. We use a 1-D spline to parametrize the depth function $z_i(k) \in \mathbb{R}^+$. Note that it is a function of a single variable, *i.e.*, the image index k. The spline is evaluated as a linear function of its control points at each image, given by:

$$z_i^k = z_i(k) = \boldsymbol{\eta}_k^\top \mathbf{w}_i, \quad i = 1 \dots n, \ k = 1 \dots m \tag{6}$$

where $\boldsymbol{\eta}_k : k \to \mathbb{R}^{m_c}$ is a function of time (image index) k and $\mathbf{w}_i$ is the vector of control points for the point i. Given that we use $m_c < m$ control points to represent each point depth on the object's surface, the set of control points is $\mathbf{w}_i = [w_1 \ w_2 \dots w_{m_c}]^\top \in \mathbb{R}^{m_c}$. The lifting function $\boldsymbol{\eta}_k$ can be precomputed. A good description of the lifting function and its computation can be found in [Brunet, 2010]. For our purpose, it produces a sparse vector with at most 4 non-zero values and of the same size as the vector of control points. Using equation (6), we can rewrite the NRSfM problem in terms of the new unknowns as below:

$$\begin{aligned}
&\underset{\{\mathbf{w}_i\},\{d_{ij}\}}{\text{maximize}} \sum_{k=1}^{m}\sum_{i=1}^{n} v_i^k \boldsymbol{\eta}_k^\top \mathbf{w}_i \\
&\text{subject to,} \\
&\boldsymbol{\eta}_k^\top \mathbf{w}_i \geq 0 \\
&d_{ij} \geq 0 \\
&\sum_{i=1}^{n}\sum_{j\in\mathcal{N}(i)} d_{ij} = 1 \\
&v_i^k v_j^k \left\| \boldsymbol{\eta}_k^\top \mathbf{w}_i \begin{bmatrix}\mathbf{q}_i^k \\ 1\end{bmatrix} - \boldsymbol{\eta}_k^\top \mathbf{w}_j \begin{bmatrix}\mathbf{q}_j^k \\ 1\end{bmatrix} \right\|_2 \leq v_i^k v_j^k d_{ij} \\
&\forall k \in \{1\ldots m\},\ i\in\{1\ldots n\},\ j\in\mathcal{N}(i).
\end{aligned} \tag{7}$$

We solve for the set of unknown control points $\{\mathbf{w}_i\}$ and the set of geodesic distances $\{d_{ij}\}$. The final depth values are obtained from equation (6) after the control points are obtained by solving problem (7). The total number of unknowns in problem (7) is thus $Kn + nm_c$ instead of $Kn + nm$. Usually we set $m_c < 0.3m$ and thus for a large problem this can result in a significant reduction of computation time as well as memory usage with a negligible drop in accuracy.

5 MDH-based Robust NRSfM

The basic problem formulation presented in section 4 gives very good reconstructions when the input correspondences have no outliers. However in the presence of a few outlier correspondences, they break down easily. This is because the method does not model noise or errors in the point correspondences. Thus the constraints at an outlier point can affect the solution of all other points. This is in contrast to local methods [Chhatkuli et al., 2014b] that solve the NRSfM problem one point at a time independently. Several strategies exist on dealing with outlier correspondences. Recovering inlier correspondences is most efficient with a dedicated outlier removal method such as [Pilet et al., 2008; Pizarro and Bartoli, 2012]. However these methods often miss a few outlier points. Consequently, an outlier rejection strategy is necessary but not sufficient for the MDH-based NRSfM, as even very few missed outliers can result in an incorrect solution. We thus require a method that gives good reconstructions even in the presence of a small percentage of outlier image correspondences or small amount of noise in the correspondences. In the SfT method [Ngo et al., 2016], the authors use an outlier removal strategy based on the mesh Laplacian; they then solve the final step of reconstruction using an iterative non linear refinement with slack variables to handle outliers. We here show that robustness with slack variables can be added into problem (3) without losing its convexity so that a global solution is still obtained. We achieve robustness by introducing slack variables in the inextensibility constraint that can 'capture' outliers.

We introduce sets of scalar variables $\{a_i^k\}$ and $\{b_i^k\}$ for each point in each view so that the back projection is:

$$\mathbf{p}_i^k = \begin{bmatrix} a_i^k \\ b_i^k \\ 0 \end{bmatrix} + z_i^k \begin{bmatrix}\mathbf{q}_i^k \\ 1\end{bmatrix}. \tag{8}$$

Equation (8) allows the sighlines from the corresponding point on image $\mathbf{q}_i^k$ to move in order to 'correct' for the outlier correspondences. The angle a given sightline moves with the above correction can be measured using the following cross-product vector:

$$\mathbf{c}_i^k = \begin{bmatrix} a_i^k \\ b_i^k \\ 0 \end{bmatrix} \times \begin{bmatrix} x_i^k \\ y_i^k \\ 1 \end{bmatrix} = \begin{bmatrix} b_i^k \\ a_i^k \\ x_i^k b_i^k - y_i^k a_i^k \end{bmatrix}. \tag{9}$$

Given that only few of the points are actually outliers requiring small corrections, a correct NRSfM solution should result in sparse sets of $\mathbf{c}_i^k$ and therefore we require minimizing the L1-norm of $\mathbf{c}_i^k$: $|a_i^k|+|b_i^k|+|x_i^k b_i^k - y_i^k a_i^k|$. We modify problem (3) to include equation (8) and add the above L1-cost as:

$$\begin{aligned}
&\underset{\{z_i^k\},\{d_{ij}\},\{a_i^k\},\{b_i^k\}}{\text{maximize}} \sum_{k=1}^{m}\sum_{i=1}^{n} v_i^k z_i^k \\
&- \lambda_r \sum_{k=1}^{m}\sum_{i=1}^{n} v_i^k \left(\left|a_i^k\right| + \left|b_i^k\right| + \left|x_i^k b_i^k - y_i^k a_i^k\right| \right) \\
&\text{subject to,} \\
&z_i^k \geq 0,\ d_{ij} \geq 0 \\
&a_i^1 = 0,\ b_i^1 = 0 \\
&\sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} d_{ij} = 1 \\
&v_i^k v_j^k \left\| z_i^k \begin{bmatrix}\mathbf{q}_i^k \\ 1\end{bmatrix} + \begin{bmatrix} a_i^k \\ b_i^k \\ 0 \end{bmatrix} - z_j^k \begin{bmatrix}\mathbf{q}_j^k \\ 1\end{bmatrix} - \begin{bmatrix} a_j^k \\ b_j^k \\ 0 \end{bmatrix} \right\|_2 \leq d_{ij} \\
&\forall k \in \{1\ldots m\},\ i\in\{1\ldots n\},\ j\in\mathcal{N}(i).
\end{aligned} \tag{10}$$

When point correspondences are obtained by tracking or wide-baseline matching with a single image (say, the first image), a further constraint can be added that no outliers exist in the first image. Thus, we additionally set $a_i^1 = 0$ and $b_i^1 = 0$. The first image point correspondences act as the reference on the basis of which the reconstructed points as well as the correspondences in other images can move to correct for outlier mismatches. We additionally require a single hyperparameter λ_r to balance the depth maximization with respect to the correction for outliers. Problem (10) is much better constrained than problem (3) when the image point correspondences have noise or outliers.

6 Experimental Results

6.1 Implementation details

We have implemented all of our methods in MATLAB using the MOSEK SOCP solver [ApS, 2015]. MOSEK is faster than many other SOCP solvers, especially for large scale problems. All of the methods can be implemented in very few lines of code (25 to 35) with the YALMIP interface [Löfberg, 2004] for MATLAB. However we use our optimized interface to call the MOSEK solver for the proposed methods in favor of speed. We can solve an NRSfM problem with 60 images, 300 points and $K = 20$ in about 4 minutes in a 2012 desktop PC. This computation time is among the fastest of the NRSfM methods for the number of images and points considered. The robust version of the method takes about 13 minutes for the same problem. On the other hand, the method imposing temporal smoothness based on splines as in problem (7) takes only 130 seconds for the same task.

6.2 Method comparison and error metrics

We compare our results against five other methods whose source code is provided by the authors. We name our first NRSfM formulation that implements problem (3) and equation (4) as **tlmdh** and its robust version of problem (10) as **r-tlmdh**. We name the implementation of our NRSfM with temporal smoothness described by equation (5) as **t-tlmdh** and our NRSfM with temporal smoothness based on 1D splines as **s-tlmdh**. We name the non-convex soft inextensibility based method for orthographic camera [Vicente and Agapito, 2012] as **o-sinext** and the local homography method for perspective camera [Chhatkuli et al., 2014b] as **p-isolh**. We write the local method of [Parashar et al., 2016] based on the metric tensor as **p-isomet**. We name the prior free factorization method of [Dai et al., 2012] as **o-spfac** and the kernel based factorization method [Gotardo and Martinez, 2011] as **o-kfac**. We name the locally rigid method based on 3-point SfM [Taylor et al., 2010] 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. For our methods we fix a single hyperparameter for all datasets. We set $\lambda_t = 0.2$ for **t-tlmdh** and $\lambda_r = 25$ for **r-tlmdh**. Similarly, we set the number of control points for depth in **s-tlmdh** to 20% of the number of images.

We measure a method's accuracy with two metrics: 3D Root Mean Square Error (RMSE), which call the 3D error and the % 3D error often used in the NRSfM literature [Agudo and Moreno-Noguer, 2015]. Both measures are almost identical and we show the 3D error in the plots. We use % 3D error when results in different sequences need to be compared in the same plot. The 3D error 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 the perspective camera (**tlmdh** and **p-isolh**) we scale their reconstructions to best align them with the ground truth. For the methods which use the affine camera (**o-sinext**, **o-lrigid** and **o-spfac**), we transform their reconstructions with a similarity transform to best align them with the ground truth. The % 3D error is defined as follows:

$$\%\ \text{3D error} = \frac{\|\mathbf{P}_{GT} - \mathbf{P}_{REC}\|_{\text{fro}}}{\|\mathbf{P}_{GT}\|_{\text{fro}}} \tag{11}$$

where $\mathbf{P}_{GT}$ represents the ground truth 3D shape ($3 \times n$ matrix) and $\mathbf{P}_{REC}$ represents the reconstructed 3D shape.

6.3 Developable Surfaces

Most non-rigid reconstruction methods focus on developable surfaces for experiments. A developable surface, such as a piece of paper or cloth, can be flattened into a planar surface without tearing or stretching. 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 7 different datasets representing such surfaces.

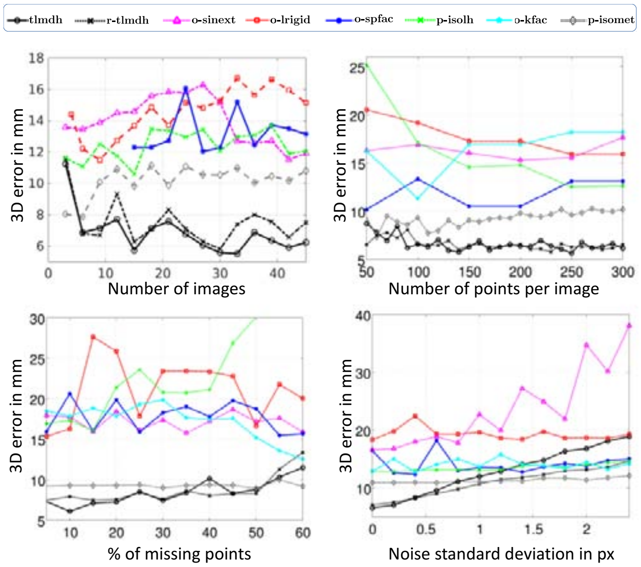

Figure 4: 3D error for the synthetic Flag dataset against the number of images and points (first row) and against the % of missing data and the amount of noise (second row). The legend is shown on the top.

The Flag dataset: We use the cloth capture data (mocap) [White et al., 2007] 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 × 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 methods and the compared 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 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 [Hu et al., 2013] 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** and **p-isomet** compute registration functions with B-splines and so we use them to fill in the missing correspondences for those methods. Figures 4 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. However, in high noise, **p-isomet** shows the best performance. Its use of the registration warps makes it robust to Gaussian noise to some extent. The same is true for a high percentage of missing data. The factorization-based method **o-spfac** and the local homography based method **p-isolh** also does better compared to the remaining methods in different conditions. We obtain a 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. We also consider the effect of noise in correspondences and use our **r-tlmdh** method to show how it performs under correspondence noise.

The KINECT Paper dataset: We use the KINECT Paper dataset [Varol et al., 2012b] as one of our real datasets for evaluation, originally used for template-based reconstruction [Ngo et al., 2016]. The dataset shows a VGA resolution sequence of a large piece of textured paper undergoing smooth deformations. Some example images were shown in figures 2 and 3. We generate correspondences by tracking points in the sequence using an optical flow-based method [Garg et al., 2013] 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 5 shows the plots of 3D error for all the images in the dataset. We obtain very accurate reconstructions that in fact compares with template-based reconstructions [Chhatkuli et al., 2014a; Ngo et al., 2016]. The best performing methods are **r-tlmdh**, **t-tlmdh**, **tlmdh** and **s-tlmdh** with mean 3D errors of 4.62 mm, 5.32 mm, 5.41 mm and 7.15 mm respectively. The local isometric method based on the metric tensor **p-isomet** is the best performing state-of-the-art method with 7.63 mm 3D error. The factorization-based methods: **o-kfac** and **o-spfac** have 3D errors of 13.93 mm and 14.66 mm respectively while **p-isolh** shows an error of 13.64 mm. The mean 3D and % 3D errors for all methods in the dataset are given in tables 2 and 3 respectively.

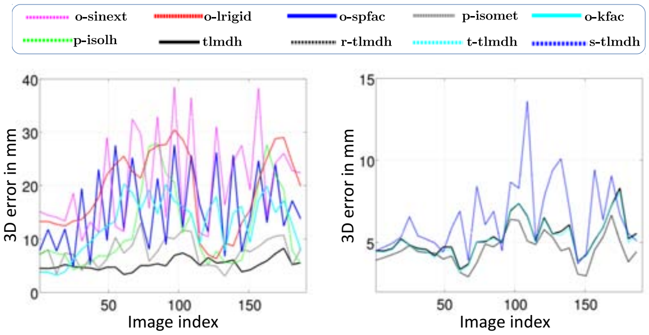

Figure 5: 3D errors for all images in the KINECT Paper dataset. The left plot shows 3D error for **tlmdh** against the compared methods and the right plot shows **tlmdh** against all other proposed methods.

The Hulk and the T-Shirt datasets: The Hulk dataset [Chhatkuli et al., 2014b] consists of a comics cover printed on a piece of paper in 21 different deformations. Similarly, the t-shirt dataset [Chhatkuli et al., 2014b] consists of a textured t-shirt with 10 different deformations. We show a few example images of the dataset in figure 6. 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. A large number of images $(m > 3/2L)$, where L is the number of shape basis, is required by **o-spfac** and a continuous video sequence is required by **o-kfac**. We give the mean error results in tables 2 and 3. The best performing methods are **tlmdh** and **r-tlmdh** with mean 3D errors of 3.51 mm and 3.45 mm for the hulk dataset; 5.41 mm and 5.39 mm for the t-shirt dataset respectively. Among the state-of-the-art methods, **p-isomet** shows the best performance with 10.76 mm and 10.60 mm error for the hulk and t-shirt datasets respectively. 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.

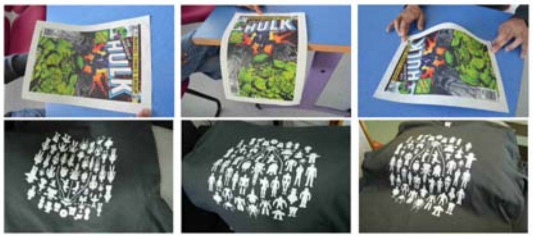

Figure 6: Example of images present in the Hulk dataset (top row) and the T-Shirt dataset (bottom row).

The Cardboard dataset: We construct a dataset using non-smooth deformations of a cardboard object. The dataset consists of 8 different deformations and images where the groundtruth 3D for each was obtained with stereo. The object used consists of repeating texture and large amount of texture-less regions. The images are taken with a focal length of about 3800 px and have a resolution of 4800 × 3200 px. We give some example images from the dataset in figure 7 below. We use a dense wide-baseline matching [Weinzaepfel et al., 2013] to compute correspondences between the images. The resulting correspondences are noisy and contains several outliers, more specifically in the textureless regions. Among our methods we test only **tlmdh** and **r-tlmdh** as we do not have a temporal continuity in the dataset images. The performance of **r-tlmdh** is particularly noteworthy with 8.35 mm 3D error in contrast to 14.86 mm for **tlmdh**. The next best performing method is **p-isolh** with 3D error of 10.02 mm. It handles the effect of outliers to some extent by the use of BBS spline-based registration. The local isometric method based on the metric tensor **p-isomet** failed to give any results for the dataset, possibly due to non-smooth surfaces and registration warps. Detailed results are provided in tables 2 and 3. We also show a comparison plot using different numbers of images in figure 8.

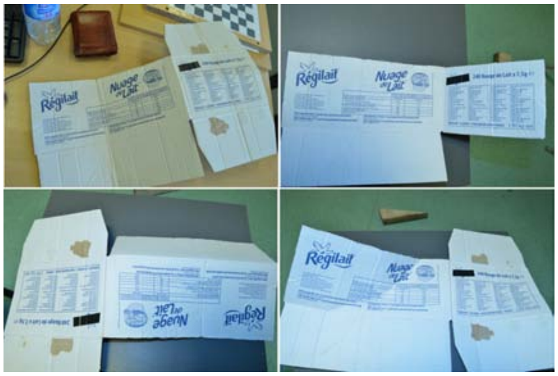

Figure 7: Example images from the Cardboard dataset.

The Rug and the Table mat datasets: We make use of existing datasets used in [Parashar et al., 2016]. The datasets

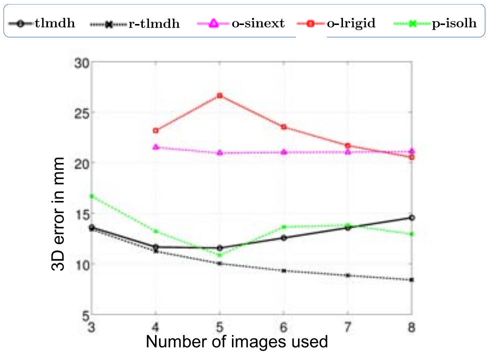

Figure 8: Mean 3D errors for different number of images in the Cardboard dataset.

are recorded with Kinect for X-box One and its images have a resolution of 1920 × 1080 px. They are taken with a focal length of 1054 px. Some example images for both datasets are shown in figure 9. The Rug dataset shows a rug being deformed smoothly in 159 images, while the Table mat dataset shows a table mat being deformed smoothly in 60 images. The correspondences are provided with the ground truth and there are no missing correspondences. However, due to the low frame-rate of the recorded sequences, the correspondences provided are not very accurate and contain outliers. We show the comparison of the proposed methods with the state-of-the-art methods for all the frames in figure 11 for the rug dataset and figure 10 for the Table mat dataset. We show the mean accuracy measures in tables 2 and 3.

Figure 9: Example images for the Table mat (top, cropped to the size of 592 × 349 px) and the Rug (bottom, original images) datasets.

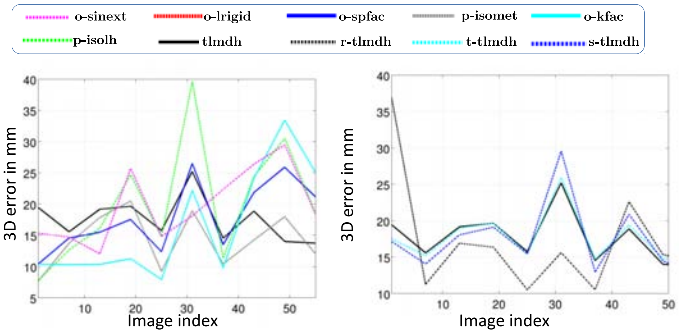

Figure 10: Mean 3D errors for all the images in the Table mat dataset. The left plot shows errors for **tlmdh** against the compared methods and the right plot shows **tlmdh** against all proposed methods.

We obtain the best results from **r-tlmdh** and **tlmdh** with 3D errors of 25.72 mm and 26.60 mm for the rug dataset; while for the Table mat dataset the compared method **p-isomet** shows the best performance with 9.6 mm compared to 14.80 mm and 16.91 mm for **r-tlmdh** and **tlmdh** respectively. We also obtain good results from **s-tlmdh** with a mean 3D error of 27.54 mm for the Rug dataset and 16.74 mm for the Table mat dataset. The compared methods **o-spfac** and **o-kfac** have a mean 3D error of 31.01 mm and 34.62 mm for the Rug dataset; 17.51 mm and 16.25 mm for the Table mat dataset. Note that the datasets are constructed with optical flow tracking on a very low frame rate sequence and thus all methods have a relatively high absolute mean error. Perhaps for the same reason, we failed to reconstruct the surfaces with **o-lrigid** using all the views. The proposed methods do not show the same level of accuracy as in the other datasets. This is also due to the relatively smaller viewpoint change and deformations present in these datasets.

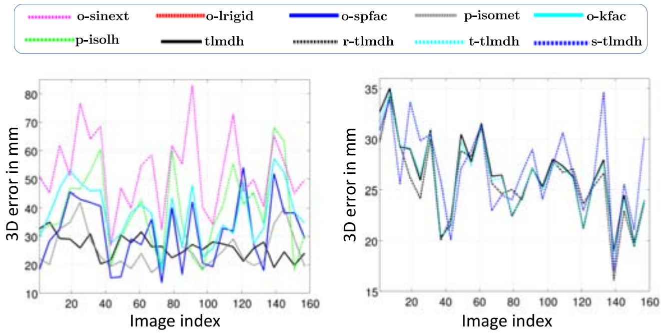

Figure 11: Mean 3D errors for all the images in the Rug dataset. The left plot shows errors for **tlmdh** against the compared methods and the right plot shows **tlmdh** against all proposed methods.

Newspaper sequence: We construct a video sequence of a tearing piece of newspaper that consists of deformation as well as articulated movement. We record the sequence using KINECT for Xbox One at full frame rate using the libfreenect2 library [Xiang et al., 2016]. The sequence has 460 images of resolution 1920 × 1080 px, taken at a focal length of about 1054 px. Some example images are shown in figure 12. We track points on the sequence again using dense point tracking [Sundaram et al., 2010]. We randomly select 900 points that are tracked in all frames. Figure 13 shows the error plots of different methods for each image in the sequence. Table 2 gives the mean accuracy measure for different methods in the sequence. The results clearly show high accuracy of the proposed methods. The mean 3D errors for **tlmdh**, **r-tlmdh** and **s-tlmdh** are 11.63 mm, 11.62 mm and 13.35 mm respectively. The closest compared method **p-isomet** has a mean 3D error of 18.40 mm. **o-spfac** shows a 3D error of 24.94 mm. There are two important reasons the proposed methods work well in this dataset: first is that the point tracking gives very good set of correspondences here due to the higher frame rate of the dataset. More importantly, the tearing of the piece of newspaper and the articulated movement tend to produce a good amount of viewpoint change. These conditions, at the same time are difficult for the compared methods to handle.

Figure 12: Example images from the Newspaper sequence.

TABLE 2: Mean 3D errors in real datasets.

	3D error measurements for different methods in mm							
Datasets	**tlmdh**	**r-tlmdh**	**p-isomet**	**p-isolh**	**o-spfac**	**o-kfac**	**o-sinext**	**o-lrigid**
KINECT Paper	5.41	**4.62**	7.63	13.64	14.66	13.93	21.45	18.65
Hulk	3.51	**3.45**	10.76	14.54	22.98	-	26.37	24.20
T-Shirt	5.41	**5.39**	10.60	8.94	-	-	18.23	-
Cardboard	14.56	**8.43**	-	12.95	-	-	35.34	20.54
Rug	26.60	**25.72**	26.15	38.26	31.01	34.62	49.14	-
Table mat	16.91	14.80	**14.21**	20.71	17.51	16.24	19.15	-
Newspaper	11.63	**11.62**	18.40	37.21	24.94	30.74	31.01	30.74

TABLE 3: Mean % 3D errors in real datasets.

	% 3D error measurements for different methods							
Datasets	**tlmdh**	**r-tlmdh**	**p-isomet**	**p-isolh**	**o-spfac**	**o-kfac**	**o-sinext**	**o-lrigid**
KINECT Paper	0.97	**0.83**	1.38	2.37	2.64	2.49	3.82	3.30
Hulk	**0.62**	**0.62**	2.81	4.17	5.10	-	5.82	5.31
T-Shirt	**1.69**	**1.69**	3.32	3.11	-	-	5.45	-
Cardboard	3.49	**2.06**	-	3.22	-	-	9.11	4.94
Rug	3.41	**3.30**	3.35	4.90	3.98	4.45	6.30	-
Table mat	1.40	1.22	**1.17**	1.71	1.45	1.34	1.58	-
Newspaper	**1.63**	**1.63**	2.63	5.20	3.50	4.24	4.34	4.31

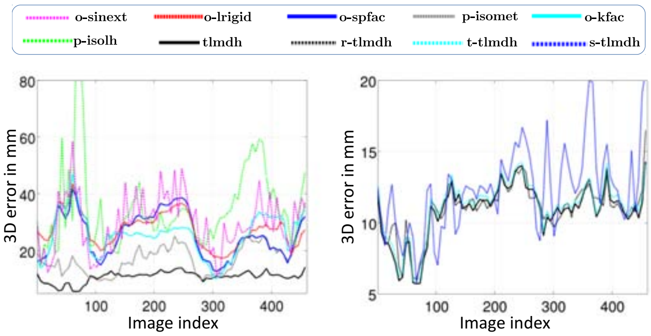

Figure 13: Mean 3D errors for all the images in the Newspaper sequence. The left plot shows errors for **tlmdh** against the compared methods and the right plot shows **tlmdh** against all proposed methods.

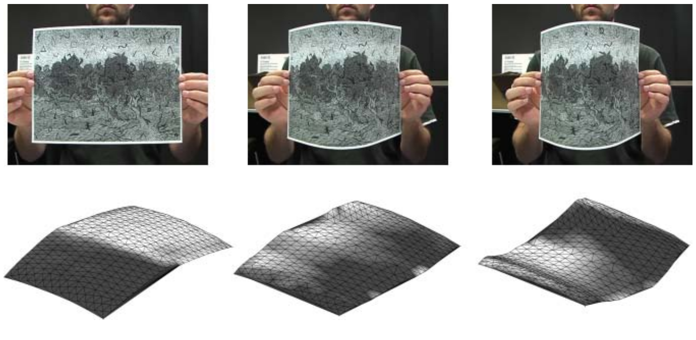

Figure 14: Failure cases: Images (top row) and their respective reconstructions (bottom row). The first two shapes appear largely incorrect.

An apparent failure case: 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 [Salzmann et al., 2007]. 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 [Sundaram et al., 2010] to obtain correspondences. The qualitative reconstructions for three frames are shown in figure 14. 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 [Vicente and Agapito, 2012], 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.

6.4 Non-developable objects

We use two different datasets to perform NRSfM on non-developable objects. They are complex objects where some of the compared methods are not even applicable, for example, both **p-isolh** and **p-isomet** 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 with our proposed **tlmdh** method. Below we describe the datasets and the experiments performed.

The Stepping Trousers dataset: The dataset [White et al., 2007] 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 × 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 [Hu et al., 2013] for all methods except ours. Figure 15 shows three reconstructed frames. From top to bottom, it shows our best reconstruction, 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 16. Because this is a large object, the 3D error can be large, yet the reconstructions can appear reasonable. We therefore also measure accuracy with a % 3D error. 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 15.

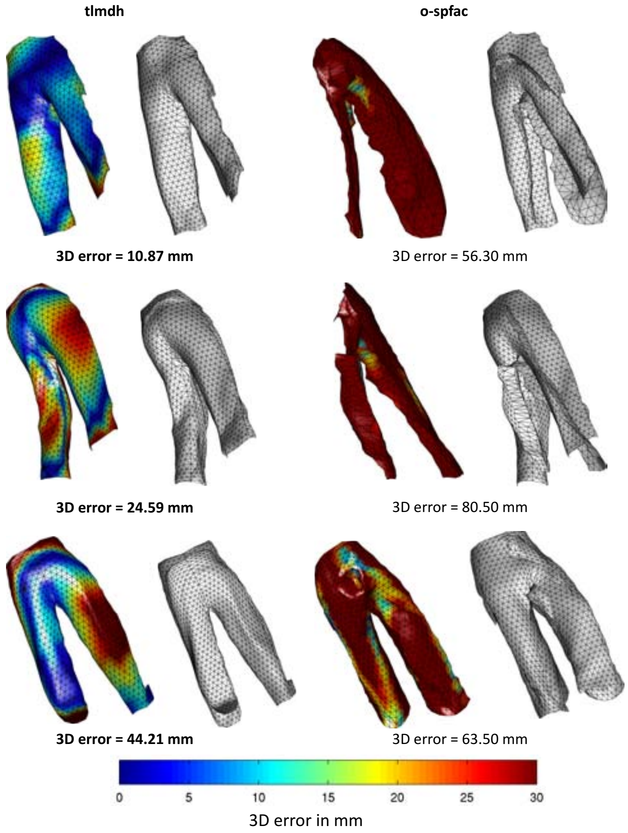

Figure 15: Reconstructions of the stepping trousers dataset for our method and **o-spfac**. Top row shows the reconstructed meshes overlaid on top of the ground truth. Bottom row shows the reconstructed mesh texture mapped with 3D error for each face in the color code shown. Note that we show our best result in the first column and the worst in the last column with a medium accuracy result in the middle.

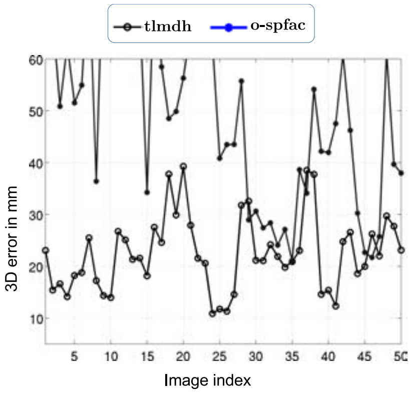

Figure 16: Plot of the depth error in trousers for uniformly sampled 50 images.

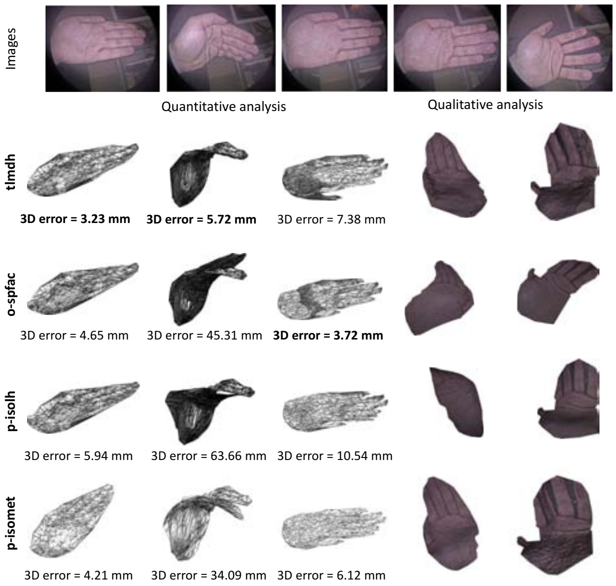

Figure 17: Results on the hand dataset. We use the best performing methods in other datasets for comparison: **o-spfac**, **p-isolh** and **p-isomet**. Ground truth is shown for three images, overlaid on top of the reconstructions. We texture map the meshes and show qualitative results for the two other images where ground truth 3D is not available.

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 modelling 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 [Sundaram et al., 2010]. 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 17 shows reconstructions of the hand compared to ground truth for our method, **o-spfac**, **p-isolh** and **p-isomet**. The results show that our method can handle complex deformations of a hand. All three compared methods were unable to capture the second deformation where they have a 3D error of over 30 mm. On the other hand we obtain a slightly higher 3D error of 7.38 mm in the third column.

6.5 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 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 [Jensen et al., 2014] on the house sequence using SIFT correspondences. We plot the 3D error for each of the 49 images for our method and compare this to a state-of-the-art rigid SfM method (VisualSfM [Wu, 2013]). We see that a reasonable error is obtained for the majority of the images.

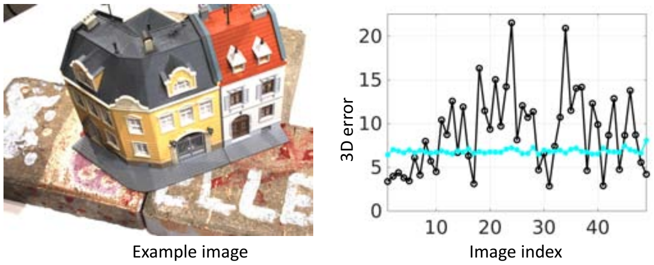

Figure 18: Results on rigid scenes. VisualSfM results are shown in cyan dots.

6.6 Sensitivity to hyperparameters

We give an analysis for the sensitivity to different hyperparameters for our methods. The common hyperparameter to all our proposed methods is K, which is the number of neighbors per point. Apart from that **r-tlmdh** and **t-tlmdh** uses an extra hyperparameter to balance two different cost terms. Finally **s-tlmdh** uses the number of control centers as a hyperparameter. We use a subset of sequences to make an analysis on these hyperparameters in figure 19 on the % 3D error. The results show that the method is not very sensitive to parameter K and λ_r as long as a high enough value is used. A higher value of K is required for scenes like Stepping Trousers due to a large number of missing correspondences and difficulty of the scene. For the plot of 3D error against λ_r we use a Gaussian noise with standard deviation of 4 pixels for the synthetic flag dataset to show that there is an optimal parameter when the noise is high. For the method with first-order smoothness **t-tlmdh**, it becomes considerably worse when a high value of λ_t is used. In **s-tlmdh**, we test the 3D percentage error against the number of control centers expressed as the percentage of the number of images m in the sequence. It is clear that the right value depends on the kind of sequence. For the flag and Kinect Paper sequence, a higher 'density' control centers are required as the frame rate is low. However, for the higher frame rate sequence of Newspaper, a lower value appears to be sufficient.

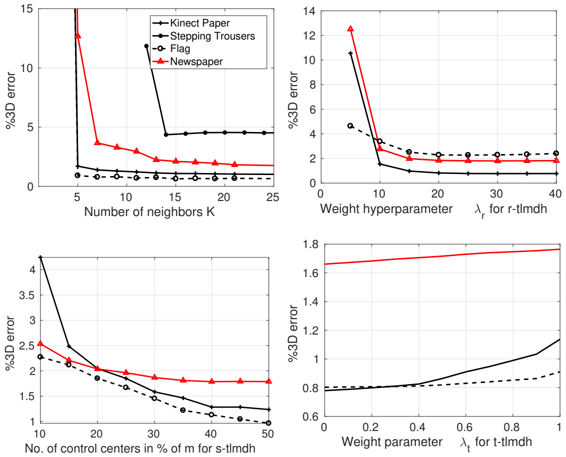

Figure 19: Results on sensitivity analysis of hyperparameters for selected sequences.

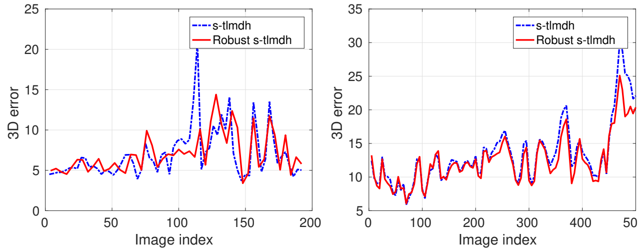

Figure 20: Comparison of **s-tlmdh** with robustness combined for KINECT Paper sequence (left) and Newspaper sequence (right) with 3D error.

7 Discussions

We presented four different convex formulations for solving NRSfM. The first formulation presented in problem (3), named **tlmdh** should be the method of choice when the point correspondences for different images have no outliers and small noise. The robust formulation **r-tlmdh**, like **tlmdh** works with wide baseline large deformations and as few as four images, albeit with an added computational cost. Both of these methods show very good performance in the experiments. However, we found that the method **t-tlmdh** of using first-order temporal smoothness as described in problem (5) provides no real improvement over the original problem. The 1D spline-based method **s-tlmdh** on the other hand, gave significant reduction in the size of the problem. It is interesting to note that enforcing temporal smoothness does not usually improve the resulting reconstruction because the original problem (3) is already well constrained. The method **s-tlmdh** can also be formulated by combining robustness as in **r-tlmdh**. Figure 20 compares the results in two sequences between the temporal smoothness only method **s-tlmdh** and the same method with robustness introduced. Here, we see an improvement in accuracy for the KINECT Paper from a 3D error of 7.15 mm to 6.96 mm while in the Newspaper sequence the 3D error improves from 13.35 mm to 12.42 mm.

Similarly, in case of no outliers, the solution of problem (10) is similar to that of problem (3). In regard to the computational complexity of solving these problems, the worst case scenario is $O(u^3)$ per iteration where u is the number of unknowns and we require about 20 to 30 iterations to solve any problem. However, the sparsity of the problem means the actual computational complexity is much lower than $O(u^3)$ per iteration.

8 Conclusion

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 that the proposed methods outperform existing ones by a large margin in many cases. In future work, we plan to study alternative relaxations of isometry apart from inextensibility. It may also be possible to formulate our approach into a sequential or incremental NRSfM so that real-time performance can be achieved.

Acknowledgements: This research has received funding from the EUs FP7 through the ERC research grant 307483 FLEXABLE. The work has also been supported by Project ARTEMISA (TIN2016-80939-R) funded by the Spanish Ministry of Economy, Industry and Competitiveness. The work is further supported by Almerys Corporation.

References

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

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.

M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).*, 2015. URL http://docs.mosek.com/7.1/toolbox/index.html.

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.

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

F. Brunet. *Contributions to Parametric Image Registration and 3D Surface Reconstruction*. PhD thesis, Université d'Auvergne, 2010.

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

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

A. Chhatkuli, D. Pizarro, T. Collins, and A. Bartoli. Inextensible non-rigid shape-from-motion by second-order cone programming. In *CVPR*, 2016.

A. Chhatkuli, D. Pizarro, A. Bartoli, and T. Collins. A stable analytical framework for isometric shape-from-template by surface integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):833–850, 2017.

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

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

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

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

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.

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

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

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.

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

H. Li. Multi-view structure computation without explicitly estimating motion. In *CVPR*, 2010.

J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, 2004.

T. D. Ngo, J. O. Östlund, and P. Fua. Template-based monocular 3D shape recovery using laplacian meshes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (1):172–187, 2016.

S. Parashar, D. Pizarro, and A. Bartoli. Isometric non-rigid shape-from-motion in linear time. In *CVPR*, 2016.

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

M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *International journal of computer vision*, 95(2):124–137, 2011.

J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration and realistic augmentation. *International Journal of Computer Vision*, 76(2):109–122, 2008.

D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. *International*

Journal of Computer Vision, 97(1):54–70, 2012.

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

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

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

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.

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

N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

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

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

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.

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

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

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

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

P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *ICCV*, 2013.

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

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

L. Xiang, F. Echtler, C. Kerl, T. Wiedemeyer, Lars, hanyazou, R. Gordon, F. Facioni, laborer2008, R. Wareham, M. Goldhoorn, alberth, gaborpapp, S. Fuchs, jmtatsch, J. Blake, Federico, H. Jungkurth, Y. Mingze, vinouz, D. Coleman, B. Burns, R. Rawat, S. Mokhov, P. Reynolds, P. Viau, M. Fraissinet-Tachet, Ludique, J. Billingham, and Alistair. libfreenect2: Release 0.2, 2016.

Biographies

Ajad Chhatkuli received his Msc degree in Computer Vision from the University of Burgundy in 2013. He recently completed his PhD in Computer Vision at Université Clermont Auvergne under the supervision of Prof. Adrien Bartoli and Dr. Daniel Pizarro. He is currently a PostDoc researcher supervised by Prof. Luc Van Gool at ETH Zürich. His research interests include template-based and template-free non-rigid 3D reconstruction.

Daniel Pizarro Pérez received the PhD degree in Electrical Engineering in 2008 from the University of Alcala. In 2005-2012 he was an Assistant Professor and member of the GEINTRA group at the University of Alcala. Since 2013 he is an Associate Professor at Université d'Auvergne and member of ALCoV. His research interests are in optimization and Computer Vision, including image registration and deformable reconstruction, and their application to Minimally Invasive Surgery.

Toby Collins received the MSc degree in Artificial Intelligence at the University of Edinburgh (first in class) in 2005. In 2006 he began his PhD in Computer Vision at the University of Edinburgh. Since 2009 he has been a full-time research fellow in ALCoV. His research interests include nonrigid shape analysis, registration and reconstruction, AR for deformable surfaces and computer assisted intervention.

Adrien Bartoli has held the position of Professor of Computer Science at Université d'Auvergne since fall 2009. He leads the ALCoV (Advanced Laparoscopy and Computer Vision) research group, member of CNRS and Université d'Auvergne, at ISIT. His main research interests include image registration and Shape-from-X for rigid and non-rigid environments, with applications to computer-aided endoscopy.