"Maximizing Rigidity" Revisited: a Convex Programming Approach for Generic 3D Shape Reconstruction from Multiple Perspective Views

Pan Ji
University of Adelaide

Hongdong Li
ANU

Yuchao Dai
ANU & NPU, China

Ian Reid
University of Adelaide

Abstract

Rigid structure-from-motion (RSfM) and non-rigid structure-from-motion (NRSfM) have long been treated in the literature as separate (different) problems. Inspired by a previous work which solved directly for 3D scene structure by factoring the relative camera poses out, we revisit the principle of "maximizing rigidity" in structure-from-motion literature, and develop a unified theory which is applicable to both rigid and non-rigid structure reconstruction in a rigidity-agnostic way. We formulate these problems as a convex semi-definite program, imposing constraints that seek to apply the principle of minimizing non-rigidity. Our results demonstrate the efficacy of the approach, with state-of-the-art accuracy on various 3D reconstruction problems.

1. Introduction

Structure-from-motion (SfM) is the problem of recovering the 3D structure of a scene from multiple images taken by a camera at different viewpoints. When the scene structure is rigid the problem is generally well defined and has been much studied [25, 12, 24, 8], with rigidity at the heart of almost all vision-based 3D reconstruction theories and methods. When the scene structure is non-rigid (deforming surface, articulated motion, and *etc*.), the problem is underconstrained, and constraints such as low dimensionality [4] or local rigidity [23] have been exploited to limit the set of solutions.

Currently, non-rigid structure from motion (NRSfM) lags far behind its rigid counterpart, and is often treated entirely separately from rigid SfM. Part of the reason for this separate treatment lies in the usual formulation of the SfM problem, which approaches the task in two stages: first the relative camera motions w.r.t. the scene are estimated; then the 3D structure is computed afterwards. In each stage, different methods and implementations have to be developed for rigid and non-rigid scenarios separately because of the different structure priors that are exploited. This has a further disadvantage in that it can be difficult to determine *a priori* whether the scene is rigid or nonrigid (and if the latter, in what way).

Therefore, it is highly desirable to have a generic SfM framework that can deal with both rigid and non-rigid motion, which leads to the main theme of this paper. In fact, as early as in the year 1983, Ullman [26] proposed a "maximizing rigidity" principle that relies on a non-convex rigidity measure to reconstruct 3D structure from both rigid and non-rigid (rubbery) motion. This idea has resurfaced in various work under the ARAP ("as rigid as possible") moniker [15, 21]. However, it has not been further developed under the modern view of 3D reconstruction, mainly due to the difficulty in its optimization. In this paper we revisit Ullman's "maximizing rigidity" principle and propose a novel convex rigidity measure that can be incorporated into a modern SfM framework for both rigid and non-rigid shape reconstruction.

Our proposed formulation yields reconstructions that are more accurate than current state-of-the-art for non-rigid shape reconstruction, and which enforces rigidity when this is present in the scene. This is because our framework aims at maximizing the rigidity while still satisfying the image measurements. We thus achieve a unified theory and paradigm for 3D vision reconstruction tasks for both rigid and non-rigid surfaces. Our method does not need to specify which case (out of the above scenarios) is the target to be reconstructed. The method will automatically output the optimal solution that best explains the observations.

2. Related Work

Traditionally, under a perspective camera, the pipeline of RSfM consists of two steps, *i.e*., a camera motion estimation step and a following structure computation step [12, 10, 9, 8]; or the camera motion and 3D structure can also be estimated simultaneously through measurement matrix factorization [24, 22]. In RSfM literature, most related to our work was by Li [11], who proposed an unusual approach to handle SfM which bypasses the motion-estimation step. This method does not require any explicit motion estimation and was called the "structure-without-motion" method.

In contrast to RSfM, NRSfM remains an open active research topic [4, 3, 30] in computer vision. One of the commonly used constraints in NRSfM is the local rigidity constraint [23, 28], or in some literature the inextensibility constraint [29].

Taylor *et al*. [23] formulated a NRSfM framework in terms of a set of linear length recovery equations using local three-point triangles under an orthographic camera, and grouped these "loosely coupled" rigid triangles into non-rigid bodies. Varol *et al*. [28] estimated homographies between corresponding local planar patches in both images. These yield approximate 3D reconstructions of points within each patch up to a scale factor, where the relative scales are resolved by considering the consistency in the overlapping patches. Both methods form part of a recent trend of piece-wise reconstruction methods in NRSfM. Instead of relying on a single model for the full surface, these approaches model small patches of the surface independently. Vicente and Agapito [29] exploited a soft inextensibility prior for template-free non-rigid reconstruction. They formulated an energy function that incorporates the inextensibility prior and minimized it via the QPBO algorithm. Very recently, Chhatkuli *et al*. [3] presented a global and convex formulation for template-less 3D reconstruction of a deforming object by using perspective cameras, where the 3D reconstruction problem is recast as a Second-Order Cone Programming (SOCP) using the Maximum-Depth Heuristic (MDH) [17, 18, 20].

The literature of RSfM and NRSfM advance in relatively independent directions. To the best of our knowledge, the first attempt to unify the two fields was by Ullman [26], who proposed to use the principle of "maximizing rigidity" to recover 3D structure from both rigid and non-rigid motion. Ullman's original formulation maintained and updated an internal rigid model of the scene across a temporal sequence. A rigid metric was defined in terms of point distance to measure the deviation from the estimated structure to the internal model. The 3D structure was recovered by minimizing the overall deviation from rigidity (internal model) to a local optimum via a gradient method. Compared to Ullman's method, our method unifies rigid and non-rigid SfM within a convex program, from which we obtain a global optimal solution.

3. Maximizing Rigidity Revisited

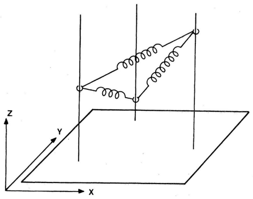

Figure 1. A spring model illustration for Ullman's maximizing rigidity principle. Each of the viewed points (three in this example) is constrained to move along one of the rigid rods, and its position along the rod represents its depth. The connecting springs represent the distances between points in the current internal model. The points would slide along the rods until a minimum energy configuration is reached. The final configuration represents the modified internal model. Image and caption are modified from Figure 2 of [26].

In this section, we discuss Ullman's "maximizing rigidity" principle in more detail. Ullman assumed that there is an internal model of the scene and the internal model should be as rigid as possible [26]. Let $\bar{d}_{ij}$ be the Euclidean distance between points i and j of the internal model, and d_{ij} the Euclidean distance between points i and j of the estimated structure. A measure of the difference between $\bar{d}_{ij}$ and d_{ij} was defined as

$$\Delta_{ij} = \frac{(\bar{d}_{ij} - d_{ij})^2}{\bar{d}_{ij}^3} \, . \tag{1}$$

Under an orthographic camera model, the pairwise distance d_{ij} was directly parameterized as $(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2$, where $\{(x_i, y_i)\}$ are the known image (coordinate) measurements and $\{z_i\}$ are the unknown depths. Unfortunately, no principled way was provided to handle the general perspective camera model. Intuitively, this measure is the least square difference reweighted by the inter-point distance of the internal model. The reweighting indicates that a point is more likely to move rigidly with its nearest neighbors [26]. However, the reweighting also makes Ullman's rigidity measure non-convex in terms of $\bar{d}_{ij}$.

Then the problem of determining the most rigid structure can be formulated as minimizing the overall deviation from rigidity $\sum_{ij} \Delta_{ij}$. Since the internal model of the scene is often unknown, the remedy that Ullman proposed was to start from a flat plane (as the initial internal model) and incrementally estimate the internal model and 3D structure. This internal model was shown to converge to a local optimum for both rigid and non-rigid motions [26, 7]. See Figure 1 for a spring model illustration of Ullman's method.

From the analysis above, we can identify three major drawbacks of Ullman's method: (**i**) it cannot handle the perspective camera in a principled way; (**ii**) the rigidity measure used is non-convex, which leads to local optimum; (**iii**) it relies on building fully connected graphs (for every pair of points), which, in practice, is often redundant and unnecessary. In the following section, we'll show how these drawbacks can be circumvented by introducing a novel convex rigidity measure which can be further incorporated into an edge-based 3D reconstruction framework for perspective projection.

4. Our 3D Shape Reconstruction Model

In this section, we present our unified model for both rigid and nonrigid 3D shape reconstruction. The core component of our model lies in a novel convex rigidity measure as introduced below. For notation, points are indexed with a subscript $i \in \{1, \cdots, n\}$, and image views (or frames) are indexed with a superscript $k \in \{1, \cdots, m\}$. We assume that the world frame is centered at the camera center and aligned with camera coordinate system.

4.1. Our Rigidity Measure

Ullman's rigidity measure requires to build a fully connected graph within each time frame and penalize distant edges (as distant point pairs are more likely to move non-rigidly). Instead of using a fully connected graph, we build a K-nearest-neighbor graph (K-NNG), which connects each point i to a set of its K nearest neighbors, denoted as $\mathcal{N}(i)$, based on the Euclidean distance on 2D images [3]. We also use a different internal model than Ullman's. Specifically, we define a rigid internal model with inter-point distance $g_{ij} = \max_k\{d_{ij}^k\}_{k=1,\cdots,m}$, *i.e.*, g_{ij} is the maximal distance between points $\mathbf{Q}_i$ and $\mathbf{Q}_j$ over all frames. So our internal model can be thought of as a "maximum distance" model. For rigid shapes, g_{ij} corresponds to the Euclidean distance between $\mathbf{Q}_i$ and $\mathbf{Q}_j$, which is invariant over all frames; for non-rigid inextensible shapes, g_{ij} corresponds to the maximal Euclidean distance between $\mathbf{Q}_i$ and $\mathbf{Q}_j$ over all frames, which generally equals the geodesic distance between $\mathbf{Q}_i$ and $\mathbf{Q}_j$. For example, for a non-rigidly deforming paper, its internal model corresponds to the flat paper. To enforce rigidity, we define a measure of the total difference between g_{ij} and d_{ij}^k for all (i, j, k) as

$$\Delta' = \sum_{i,j\in\mathcal{N}(i),k} |g_{ij} - d_{ij}^k| \, . \tag{2}$$

Compared to Ullman's rigidity measure, our rigidity measure has three merits: (**i**) we significantly reduce the number of edges by using a K-NNG instead of a fully connected graph; (**ii**) our measure is convex, which is crucial for optimization; (**iii**) we use a robust L-1 norm instead of the L-2 norm in the rigidity measures. To make the reconstructed scene as rigid as possible, we need to minimize Δ'. As we will see in the following subsections, our rigidity measure can be naturally incorporated into an edge-based 3D reconstruction framework under perspective projection.

4.2. Edge-Based Reconstruction

Given a set of n 2D point correspondences across m images $\{\mathbf{q}_i^k\}$, our target is to find their 3D coordinates $\mathbf{Q}_i^k$ in the same global coordinate system. We denote the edge (distance) between the camera center $\mathbf{O}$ and $\mathbf{Q}_i^k$, which we call a "leg", as ℓ_i^k. Define the angle between legs ℓ_i^k and ℓ_j^k as θ_{ij}^k. Clearly, we have $\theta_{ij}^k = \theta_{ji}^k$. We assume that the camera is intrinsically calibrated, so the angles θ_{ij}^k can be trivially computed. Denote the Euclidean distance between two points $\mathbf{Q}_i^k$ and $\mathbf{Q}_j^k$ in the k^{th} frame as d_{ij}^k. For rigid motion, d_{ij}^k is constant over frames for the same pair of point correspondences. In the case of non-rigid motions, d_{ij}^k may change over frames, but is bounded by a maximal value g_{ij}, *i.e.*, $g_{ij} = \max_k\{d_{ij}^k\}_{k=1,\cdots,m}$.

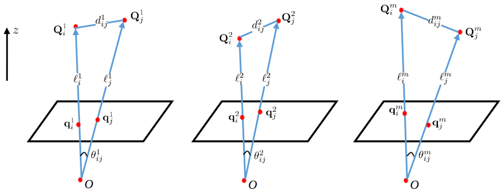

Figure 2. Viewing triangles of a pair of 3D points $\mathbf{Q}_i$ and $\mathbf{Q}_j$ in different views. We use O to denote the camera center from which we draw viewing ray $\mathrm{O}\mathbf{Q}_i^k$ intersecting with the image plane at $\mathbf{q}_i^k$. For example, $\triangle\mathrm{O}\mathbf{Q}_i^1\mathbf{Q}_j^1$ forms a viewing triangle in the first view for $\mathbf{Q}_i$ and $\mathbf{Q}_j$.

Motivated by [11], we build our model based on viewing triangles formed by each pair of points to compute the 3D structure. See Figure 2 for an illustration. Note that the viewing triangles can only be formed with points of the same frame.

Within each viewing triangle, we have a basic equation following the cosine law

$${\ell_i^k}^2 + {\ell_j^k}^2 - 2\ell_i^k\ell_j^k\cos\theta_{ij}^k = {d_{ij}^k}^2 \, . \tag{3}$$

We can rewrite this equation in a matrix form as

$$\begin{bmatrix}\ell_i^k & \ell_j^k\end{bmatrix}\begin{bmatrix}1 & -\cos\theta_{ij}^k \\ -\cos\theta_{ij}^k & 1\end{bmatrix}\begin{bmatrix}\ell_i^k \\ \ell_j^k\end{bmatrix} = {d_{ij}^k}^2 \, . \tag{4}$$

With all viewing triangles, we can construct a system of quadratic equations of the above form in terms of the unknowns ℓ_i^k, ℓ_i^k and d_{ij}^k.

Stack all the legs ℓ_i^k into a vector $\boldsymbol{\ell} = [{\boldsymbol{\ell}^1}^T \; \cdots \; {\boldsymbol{\ell}^m}^T]^T$, with the leg vector for the k^{th} frame $\boldsymbol{\ell}^k = [\ell_1^k \; \cdots \; \ell_n^k]^T$. Define the cosine-matrix as $\mathbf{C}^k \in \mathbb{R}^{n\times n}$ with the diagonal elements as one and off-diagonal elements as $c_{ij}^k = -\cos\theta_{ij}^k$. Let $\mathbf{e}_i$ be the unit basis vector (*i.e.*, all 0 but 1 at the i^{th} entry), and define the diagonal matrix $\mathbf{E}_{ij} = \mathrm{diag}(\mathbf{e}_i + \mathbf{e}_j)$. Eq. (4) can then be rewritten as

$${\boldsymbol{\ell}^k}^T \mathbf{E}_{ij}^T \mathbf{C}^k \mathbf{E}_{ij} \boldsymbol{\ell}^k = {d_{ij}^k}^2 \, . \tag{5}$$

Note that the matrix $\mathbf{A}_{ij}^k \doteq \mathbf{E}_{ij}^T\mathbf{C}^k\mathbf{E}_{ij}$ is highly sparse with only four non-zero elements.

The edge-based 3D reconstruction problem becomes

$$\text{Find} \;\; \boldsymbol{\ell}, \mathbf{d} \tag{6a}$$

$$\text{s.t.} \;\; {\boldsymbol{\ell}^k}^T \mathbf{A}_{ij}^k \boldsymbol{\ell}^k = {d_{ij}^k}^2 \, , \tag{6b}$$

$$\ell_i^k \geq 0, \; d_{ij}^k \geq 0, \; \forall(i, j, k) \, , \tag{6c}$$

where $\mathbf{d}$ is a vector containing all d_{ij}^k.

However, the above formulation is not well constrained because: (i) the scale of the solutions cannot be uniquely determined due to the homogeneous equation in (6b); (ii) a trivial all-zero solution always exists; (iii) for non-rigid motions, we only have one equality constraint for each d_{ij}^k, which is insufficient to get deterministic solutions[1]. The scale ambiguity is intrinsic to 3D reconstruction under the perspective camera model [8]. In practice, we can fix a global scale of the scene by normalizing $\boldsymbol{\ell}$ or $\mathbf{d}$.

Maximum-Leg Heuristic (MLH). To get the desired solutions, we apply a so-called Maximum-Leg Heuristic (MLH). After fixing the global scale, we want to maximize the sum of all legs ℓ_i^k under the non-negative constraint, or equivalently

$$\min_{\boldsymbol{\ell},\mathbf{d}} \quad -\sum_{i,k} \ell_i^k \tag{7a}$$

$$\text{s.t.} \quad {\boldsymbol{\ell}^k}^T \mathbf{A}_{ij}^k \boldsymbol{\ell}^k = {d_{ij}^k}^2 \,, \tag{7b}$$

$$\ell_i^k \geq 0,\ d_{ij}^k \geq 0,\ \forall (i,j,k)\ . \tag{7c}$$

In this way, trivial solutions are avoided. Note that, in the NRSfM literature, there is a commonly used heuristic called Maximum Depth Heuristic (MDH) [17, 18, 20], which maximizes the sum of all depths under the condition that each depth and distance are positive. Our MLH virtually plays the same role as MDH because under a perspective camera, we have $\ell_i^k = z_i^k \|\hat{\mathbf{q}}_i^k\|_2$, where z_i^k represents the depth of the i^{th} point in the k^{th} frame (*i.e.*, $\mathbf{Q}_i^k$), and $\hat{\mathbf{q}}_i^k = \mathbf{K}^{-1}[{\mathbf{q}_i^k}^T \ 1]^T$.

5. Convex Program for 3D Shape Reconstruction

Incorporating our rigidity measure in (2) into (7), we get our overall formulation for 3D shape reconstruction as follows

$$\min_{\boldsymbol{\ell},\mathbf{d},\mathbf{g}} \quad -\sum_{i,k} \ell_i^k + \lambda \sum_{i,j,k} |g_{ij} - d_{ij}^k| \tag{8a}$$

$$\text{s.t.} \quad {\boldsymbol{\ell}^k}^T \mathbf{A}_{ij}^k \boldsymbol{\ell}^k = {d_{ij}^k}^2 \,, \tag{8b}$$

$$d_{ij}^k \leq g_{ij} \,, \tag{8c}$$

$$\sum_{ij} g_{ij}^2 = 1 \,, \tag{8d}$$

$$\ell_i^k \geq 0,\ d_{ij}^k \geq 0\,, \forall \big(i, j \in \mathcal{N}(i), k\big)\,, \tag{8e}$$

where $\lambda > 0$ is a trade-off parameter, and the equality constraint (8d) fixes the global scale of the reconstructed shape.

However, the above formulation is still non-convex due to the quadratic terms in the both sides of Eq. (8b) and in the left-hand-side of Eq. (8d). To make it convex, we first define $\hat{g}_{ij} = {g_{ij}}^2$ and $\hat{d}_{ij}^k = {d_{ij}^k}^2$. We then change our formulation to the following form

$$\min_{\boldsymbol{\ell},\hat{\mathbf{d}},\hat{\mathbf{g}}} \quad -\sum_{i,k} \ell_i^k + \lambda \sum_{i,j,k} (\hat{g}_{ij} - \hat{d}_{ij}^k) \tag{9a}$$

$$\text{s.t.} \quad {\boldsymbol{\ell}^k}^T \mathbf{A}_{ij}^k \boldsymbol{\ell}^k = \hat{d}_{ij}^k \,, \tag{9b}$$

$$\hat{d}_{ij}^k \leq \hat{g}_{ij} \,, \tag{9c}$$

$$\sum_{ij} \hat{g}_{ij} = 1 \,, \tag{9d}$$

$$\ell_i^k \geq 0,\ \hat{d}_{ij}^k \geq 0\,, \forall \big(i, j \in \mathcal{N}(i), k\big)\,, \tag{9e}$$

where we approximate $|g_{ij} - d_{ij}^k|$ with $|\hat{g}_{ij} - \hat{d}_{ij}^k|$, and drop the absolute value operator as we have an inequality constraint (9c) to make sure $\hat{g}_{ij} - \hat{d}_{ij}^k$ is always non-negative. Due to (9b), our formulation turns out to be a quadratically constrained quadratic program (QCQP), which is unfortunately still a non-convex and even NP-hard problem for indefinite $\mathbf{A}_{ij}^k$ [16, 19].

5.1. Semi-Definite Programming (SDP) Relaxation

We now show how our formulation can be converted to a convex program using SDP relaxation. Note that we have ${\boldsymbol{\ell}^k}^T \mathbf{A}_{ij}^k \boldsymbol{\ell}^k = \mathrm{tr}(\mathbf{A}_{ij}^k \boldsymbol{\ell}^k {\boldsymbol{\ell}^k}^T)$. We can introduce auxiliary variables $\mathbf{Y}^k = \boldsymbol{\ell}^k {\boldsymbol{\ell}^k}^T$ for $k = 1, \cdots, m$. Then Eq. (9b) equivalently becomes two equality constraints $\mathrm{tr}(\mathbf{A}_{ij}^k \mathbf{Y}^k) = \hat{d}_{ij}^k$, $\mathbf{Y}^k = \boldsymbol{\ell}^k {\boldsymbol{\ell}^k}^T$. We can directly relax the last non-convex equality constraint $\mathbf{Y}^k = \boldsymbol{\ell}^k {\boldsymbol{\ell}^k}^T$ into a convex positive semi-definiteness constraint $\mathbf{Y}^k \succeq \boldsymbol{\ell}^k {\boldsymbol{\ell}^k}^T$ [5]. Using a Schur complement, $\mathbf{Y}^k \succeq \boldsymbol{\ell}^k {\boldsymbol{\ell}^k}^T$ can be reformulated [1] as $\begin{bmatrix} 1 & {\boldsymbol{\ell}^k}^T \\ \boldsymbol{\ell}^k & \mathbf{Y}^k \end{bmatrix} \succeq 0$. Ideally, $\mathbf{Y}^k$ should be a rank-one matrix. But, after the relaxation, the rank constraint for $\mathbf{Y}^k$ may not be maintained. We can minimize $\mathrm{tr}(\mathbf{Y}^k)$ as the convex surrogate of $\mathrm{rank}(\mathbf{Y}^k)$.[2]

Our formulation becomes an SDP written as:

$$\min_{\boldsymbol{\ell},\hat{\mathbf{d}},\hat{\mathbf{g}},\mathbf{Y}^k} \quad \sum_k \mathrm{tr}(\mathbf{Y}^k) - \lambda_1 \mathbf{1}^T \boldsymbol{\ell} - \lambda_2 \mathbf{1}^T \hat{\mathbf{d}} \tag{10a}$$

$$\text{s.t.} \quad \mathrm{tr}(\mathbf{A}_{ij}^k \mathbf{Y}^k) = \hat{d}_{ij}^k \,, \tag{10b}$$

$$\begin{bmatrix} 1 & {\boldsymbol{\ell}^k}^T \\ \boldsymbol{\ell}^k & \mathbf{Y}^k \end{bmatrix} \succeq 0 \,, \tag{10c}$$

$$\hat{d}_{ij}^k \leq \hat{g}_{ij} \,,\ \mathbf{1}^T \hat{\mathbf{g}} = 1 \,, \tag{10d}$$

$$\ell_i^k \geq 0,\ \hat{d}_{ij}^k \geq 0\,, \forall \big(i, j \in \mathcal{N}(i), k\big)\,, \tag{10e}$$

[1] For rigid motions, $d_{ij}^k = d_{ij}, \forall k \in \{1, \cdots, m\}$, and we have sufficient constraints for d_{ij}. But in many cases, we don't know whether the scene is rigid or non-rigid *a priori*.

[2] For positive semi-definite $\mathbf{Y}^k$, $\mathrm{tr}(\mathbf{Y}^k) = \|\mathbf{Y}^k\|_*$, and the nuclear norm is a well-known convex surrogate for the rank.

where λ_1, λ_2 are two positive trade-off parameters, and $\mathbf{1}$ is an all-one column vector with appropriate dimensions. Our model incorporates our maximizing rigidity principle with the MLH under the constraints of viewing-triangle cosine-law and the internal model. Note that we remove a term of $\hat{\mathbf{g}}$ in the objective (10a) because we have $\sum_{i,j,k} \hat{g}_{ij} = m$, which is a constant. Our formulation consists of a linear objective subject to linear constraints and SDP constraints, which is known as a convex problem. This convex SDP problem can be solved effectively by any modern SDP solver to a global optimum.

Incomplete Data. Incomplete measurements are quite common due to occlusions. To handle incomplete measurements, we can introduce a set of visibility masks $\mathbf{W} \doteq \{w_i^k\}$, where $w_i^k = 1$ if the i^{th} point is visible in frame k, otherwise $w_i^k = 0$. With the visibility masks, the terms related to ℓ_i^k become $w_i^k \ell_i^k$ and the terms related to d_{ij}^k become $w_i^k w_j^k d_{ij}^k$. The problem is still convex and solvable with any SDP solver. Here we assume that the number of visible points in one frame is greater than the neighborhood size; otherwise, we remove that frame.

5.2. 3D Reconstruction from Legs

Under the perspective camera model, we can relate $\mathbf{Q}_i^k$ and $\mathbf{q}_i^k$ with

$$\mathbf{Q}_i^k = z_i^k \hat{\mathbf{q}}_i^k = l_i^k \hat{\mathbf{q}}_i^k / \|\hat{\mathbf{q}}_i^k\|_2 \ , \tag{11}$$

where $\hat{\mathbf{q}}_i^k = \mathbf{K}^{-1}[\mathbf{q}_i^{k^T} \ \ 1]^T$. After we get the solutions for all legs ℓ_i^k, we can then substitute them back to the above equation to compute the 3D coordinates for all points.

Degenerate Cases. Our system becomes degenerate if there is only pure rotation (around the camera center) in the scene. In fact, pure rotation over the camera center do not change the angles between two vectors, *e.g.*, $\mathbf{v}_1$ and $\mathbf{v}_2$,

$$\cos(\theta) = \frac{\mathbf{v}_1^T \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} = \frac{(\mathbf{R}\mathbf{v}_1)^T (\mathbf{R}\mathbf{v})_2}{\|\mathbf{R}\mathbf{v}_1\| \|\mathbf{R}\mathbf{v}_2\|} \ , \tag{12}$$

where the equations hold because $\mathbf{R}^T\mathbf{R} = \mathbf{I}$ and rotation on vectors does not change their length. So if there is only pure rotation in the scene, our system will become under-constrained. This also corresponds to the fact in epipolar geometry that pure rotation cannot be explained by the essential/fundamental matrix (but homography instead). Another degenerate case is when the camera model is close to orthographic. In this case, the viewing angles are all close to zero, which makes our formulation unsolvable.

6. Experiments

We compare our method with four baselines for rigid and non-rigid 3D shape reconstruction. These baselines include: the rigid "structure-without-motion" method for a perspective camera in [11], the non-convex soft-inextensibility based NRSfM method for an orthographic camera in [29], the prior-free low-rank factorization based NRSfM method for an orthographic camera in [4], and the second-order cone programming based NRSfM method for a perspective camera [3]. For the baselines, we use the source codes provided by the authors. We implement our method in Matlab and use the MOSEK [14] SDP solver to solve our formulation. We fix all the parameters of the baseline methods to the optimal values. We find that our method is not sensitive to the parameters λ_1 and λ_2, and set $\lambda_1 = 1$ and $\lambda_2 = 20$ for all our experiments, which are obtained by validating on a separate dataset. To give a fair comparison, we always use the same K-NNG for [3] and our method. Due the limit of space, our qualitative reconstruction results on all synthetic datasets are provided in the supplementary videos.

The metrics we use to evaluate the performance are the 3D Root Mean Square Error (RMSE) (in mm) and the relative 3D error (denoted as R-Err) (in %), which are respectively defined as

$$\text{RMSE} = \frac{1}{m} \sum_k \sqrt{\frac{1}{n} \sum_i \|\bar{\mathbf{Q}}_i^k - \mathbf{Q}_i^k\|_2^2} \ ,$$

$$\text{R-Err} = \frac{1}{m} \sum_k \frac{\|\bar{\mathbf{Q}}^k - \mathbf{Q}^k\|_F}{\|\bar{\mathbf{Q}}^k\|_F} \times 100\% \ ,$$

where $\bar{\mathbf{Q}}_i^k$ is the ground truth coordinates of point i in frame k. We always have a scale ambiguity for all structure-from-motion methods. For methods that use a perspective camera model, we re-scale their reconstructions to best align them with the ground truth before computing the errors. For methods that use an orthographic camera model, we do Procrustes analysis to solve for a similarity transformation that best aligns the reconstructions with the ground truth.

6.1. Non-rigid Structure from Motion

Our method and [3] rely on constructing a K-NNG. For both methods, we use the same K-NNG and fix the neighborhood size K as 20 for this set of experiments.

The Flag (Semi-Synthetic) Dataset. This flag dataset [31] consists of an image sequence of a fabric flag waving in the wind. The ground truth 3D points are provided in the dataset, but neither 2D projection trajectories nor camera calibrations are available. We subsample the 3D points in each frame and generate the input data from a virtual perspective camera with the field-of-view angle as $81.69°$. The final sequence contains 90 points (on each frame) and 50 frames. We report the 3D RMSE and mean relative 3D error in Table 1. Note that our method achieves the lowest 3D reconstruction error among all the competing methods.

The KINECT Paper, Hulk, and T-Shirt Datasets. The KINECT paper dataset [27] contains an image sequence of smoothly deforming well-textured paper captured by a

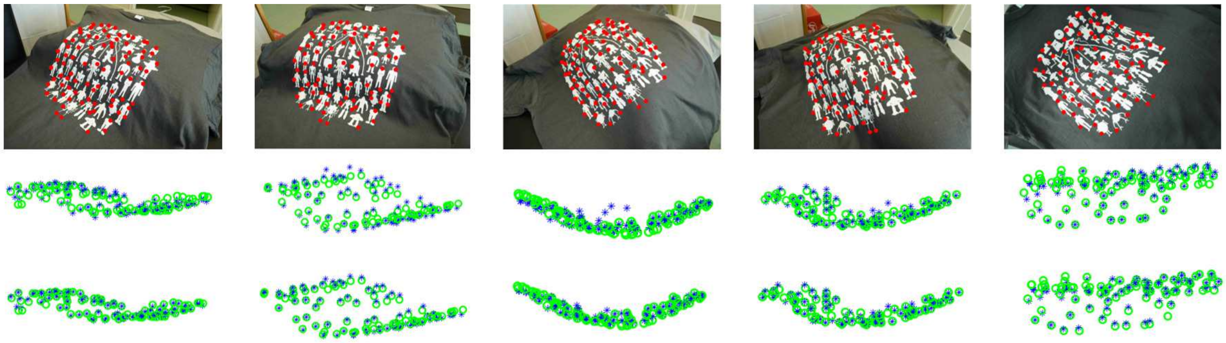

Figure 3. Qualitative comparison of the 3D reconstruction results on the T-shirt dataset. The green circles plot the ground truth 3D points, and the blue stars show the reconstructed 3D points. **Top row**: 2D images with feature points in red dots. **Middle row**: results of [3]. **Bottom row**: results of our method. Best viewed on screen with zoom-in.

Table 1. Mean 3D errors for the Flag Paper dataset.

	[29]	[4]	[3]	Ours
RMSE	41.92	26.23	21.08	**16.75**
R-Err	12.76%	7.51%	6.38%	**5.07%**

KINECT camera. The camera calibration and ground truth 3D are provided. We use the trajectories provided by [3], which was obtained by tracking interest points in this sequence using a flow-based method of [6]. The trajectories are complete, semi-dense and outlier-free. Due to the large number of points and frames, we subsample the points and frames in this dataset and get a sequence with 151 points (on each frame) and 23 frames.

The Hulk dataset [2] consists of 21 images taken at different unrelated smooth deformations. The deforming scene is a well-textured paper cover of a comics. The intrinsic camera calibration matrix, 3D ground truth shape and 2D feature trajectories are provided in this dataset. This dataset contains 122 trajectories in 21 views.

The T-Shirt dataset [2] consists of 10 images taken for a deforming T-shirt. As in the Hulk dataset, the intrinsic camera calibration matrix, 3D ground truth shape and 2D feature trajectories are all provided in this dataset. This dataset contains 85 point trajectories in 10 frames.

We show the mean 3D errors of our method and the baselines in Figure 4. We can see that our method achieves the lowest 3D reconstruction error on all the three datasets. We also give a qualitative comparison with the best-performing baseline [3] in Figure 3.

The Jumping Trousers Dataset with missing data. This dataset [31] contains 3D ground truth points for jumping trousers obtained from cloth motion capture. The complete 2D trajectories are generated by projecting the 3D points through a virtual perspective camera. However, due to self-occlusions, the 2D trajectories would have a considerable amount of missing entries, and the visibility masks are provided in the original data. We subsample the points and frames, and get a sequence of 97 points and 29 frames. Since the first two baselines [29, 4] cannot handle incomplete data, we input complete trajectories for them. We use the incomplete trajectories for [3] and our method as the two methods can handle incomplete data. [3] The results are reported in Table 2. Our method, with incomplete data as input, outperforms all the other baselines.

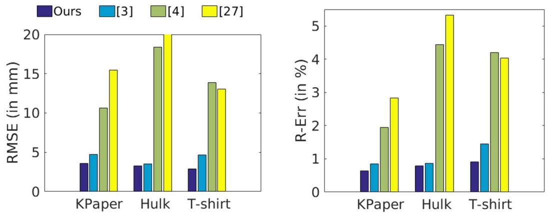

Figure 4. Mean 3D reconstruction errors for the KINECT Paper (denoted as "KPaper" in the figure), Hulk, and T-shirt Datasets. Our method (plotted in blue) achieves the lowest 3D reconstruction errors on all three datasets.

Table 2. Mean 3D errors for the Jumping Trousers dataset.

	[29]	[4]	[3]	Ours
RMSE	190.17	49.97	44.05	**37.70**
R-Err	55.10%	12.67%	13.57%	**11.65%**

From this set of experiments, we have shown that our method consistently outperforms all the baselines. We note that on those datasets there is always a significant performance gap between those orthographic camera model based methods ([29, 4]) and those perspective camera model based methods ([3] and ours). In the following experiments, we will only compare with the perspective camera model based methods ([11, 3]).

Robustness to various numbers of points/views, different levels of missing data and noise. In Figure 5,

[3] Note, for incomplete data, we only compute average 3D reconstruction error for the visible points. And also note that this comparison is unfair for [3] and our method as the other two use complete data.

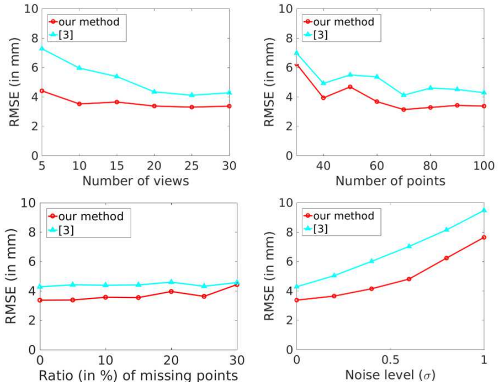

Figure 5. Performance evaluation on the KINECT Paper dataset with increasing number of points/views, increasing ratios of missing data, and increasing levels of synthetic Gaussian noise.

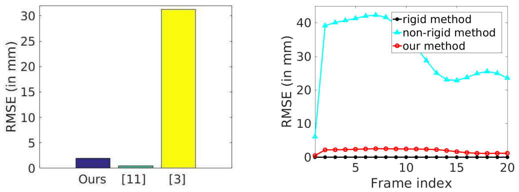

Figure 6. **Left:** mean 3D errors for the synthetic rigid dataset. **Right:** the RMSE (in mm) for each frame of the synthetic sequence by the rigid method [11] (in black dots), the non-rigid method [3] (in cyan dots), and our method (in red dots).

we show the performance of our method and the best-performing baseline [3] on the KINECT paper dataset with increasing number of points/views, increasing ratios of missing data, and increasing levels of synthetic zero-mean Gaussian noise (with various standard deviations σ). The default experimental setting is with 100 points and 30 views, and the parameters are fixed as $\lambda_1 = 1$, $\lambda_2 = 20$, and $K = 20$. We can see that our method consistently outperforms the baseline method in all scenarios, which verifies the robustness of our method. We believe that our superior performance comes from the novel maximizing rigidity regularization, which better explains the image observations.

6.2. Rigid Structure from Motion

In this set of experiments, we test our method for rigid structure reconstruction. Since our method does not utilize the rigidity prior of the scene, we can well expect that our method performs worse than the specifically designed rigid method. The main goal of this set of experiments is thus to show that our method can achieve comparable rigid structure reconstruction to the rigid method. We compare our method with the best-performing baseline [3] for non-rigid structure from motion, and another method [11] specifically designed for rigid structure from motion. The neighborhood size is set as 20 for all methods.

Rigid Synthetic Dataset. We verify our method for rigid structure computation on a synthetic dataset. To generate the data, we subsample the ground truth 3D points of one frame of the KINECT paper dataset [27], and apply a transformation (rotation and translation) to these points over time. After a perspective projection, we get a sequence for rigid motion with 61 points and 20 frames. The mean 3D reconstruction errors for all competing methods are reported in Figure 6. We also plot the RMSE (in mm) for each frame of the sequence in Figure 6 and compare our method with the state-of-the-art non-rigid SfM method [3] and the rigid SfM method [11]. It's no surprising that [11] achieves the lowest reconstruction errors in this rigid dataset as it utilizes the prior knowledge that the scene is rigid. Our method, without inputting any prior knowledge of the scene rigidity, gets close results to [11] and significantly outperforms the NRSfM method [3].

The Model House Dataset. We use the VGG model house dataset [4]as the real-world dataset for rigid SfM. The camera projection matrices, 2D feature coordinates and 3D ground truth points are provided in this dataset, and the 2D measurements contain moderate amount of noise. The camera intrinsic matrices are computed from camera projection matrices using R-Q decomposition [8]. We generate a sequence with complete feature point trajectories of 95 points and 7 frames. We report the 3D reconstruction errors of all methods in Table 3. Again, our method obtains comparable results to [11] and lower reconstruction error than the NRSfM method [3].

Table 3. Mean 3D errors for the Model House dataset.

	[11]	[3]	Ours
RMSE	0.158	0.200	0.162
R-Err	2.95%	3.73%	3.02%

6.3. Articulated Motion Reconstruction

In this set of experiments, we evaluate our method for the 3D reconstruction of articulated motions, and compare our method with the best-performing baseline [3].

Synthetic Articulated Dataset. We first test our method on two synthetic sequences where the objects undergo articulated motions. To generate the synthetic data, we take a subset of the ground truth 3D points in the first image of the KINECT paper dataset [27] and divide them into two groups. We synthesize two kinds of articulated motions: (i) the point-articulated motion (denoted as "point-articulated" in Table 4), *i.e.*, the two groups of points rotate around a common point in the dataset and meanwhile undertake the same translations through time; (ii) the axis-articulated motion (denoted as "axis-articulated" in Table 4), *i.e.*, the two

[4] http://www.robots.ox.ac.uk/~vgg/data/data-mview.html

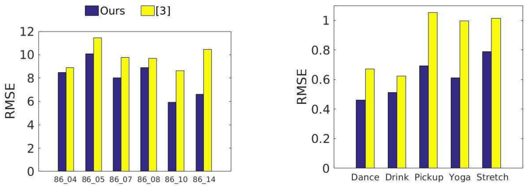

Figure 7. **Left**: mean 3D reconstruction errors on six sampled sequences (86_04, 86_05, 86_07, 86_08, 86_10, and 86_14) of CMU Mocap Database. **Right**: mean 3D reconstruction errors on Dance, Drink, Pickup, Yoga, and Stretch sequences.

groups of points rotate around a common axis in the dataset and also undertake the same translations. The 2D feature points are generated by projecting these 3D points with a virtual perspective camera. We finally get two synthetic sequences with 61 points and 19 frames. We report the RMSE (in millimeter) and the mean relative 3D error in Table 4. Our method achieves much lower 3D reconstruction error than the baseline method [3].

Table 4. RMSE (in mm) and mean relative 3D error (shown in brackets) in percentage (%) for the synthetic articulated data.

sequence	[3]	Ours
point-articulated	17.48 (2.45%)	**7.70 (1.11%)**
axis-articulated	9.13 (1.36%)	**3.07 (0.45%)**

Human Motion Capture Database. We sample six sequences in the CMU Mocap Database [5] and five sequences (Dance, Drink, Pickup, Yoga, and Stretch sequences) used in [4] to form the human motion capture database. For the latter five sequences, the data are centered to fit the factorization-based methods, so we further add random translations to each frame. Each sequence of this database consists of 28 (for CMU Mocap), 41 (for Drink, Pickup, Yoga, Stretch) or 75 (for Dance) points with 3D ground truth coordinates. The input data are generated from a virtual camera with perspective projection. We uniformly subsample the frames of each sequence with a sample rate 10 (*i.e.*, $1 : 10 : \text{end}$) for CMU Mocap and a sample rate 5 for other sequences, producing sequences with 52 to 335 frames. For CMU Mocap, we set the neighborhood size K as 28 for all competing methods, which lets us to use all available points to build the edges; for other sequences, we set K as 20. We show the quantitative results of our method and the baseline method in Figure 7, and also give a qualitative comparison of the 3D reconstruction on this dataset in Figure 8. We can see that our method consistently outperforms the baseline [3]. However, we notice our reconstruction is still far from perfect. We conjecture that this is because the distance based measure cannot resolve the two-fold ambiguity of human poses along the viewing rays[6].

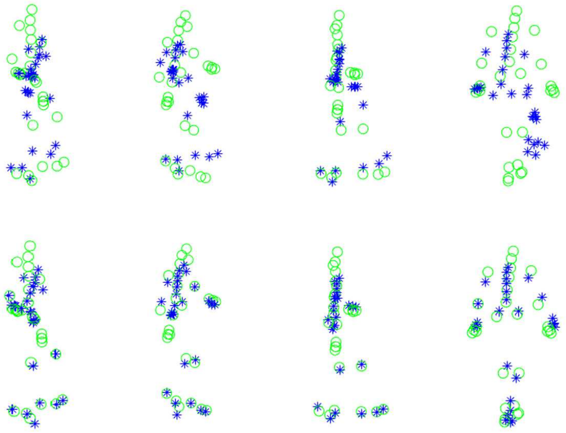

Figure 8. Qualitative comparison of the 3D reconstruction results on the CMU Mocap Database. The green circles plot the ground truth 3D points, and the blue stars show the reconstructed 3D points. **Top row**: results of [3]. **Bottom row**: results of our method.

7. Concluding Remarks

In this paper, we have revisited Ullman's principle of maximizing rigidity and proposed a novel convex rigidity measure that can be incorporated into a modern structure reconstruction framework to unify both rigid and non-rigid SfM from multiple perspective images. Our reconstruction method relies on directly building viewing triangles, thus not requiring to estimate camera poses. Importantly, our formulation (after SDP relaxations) is convex such that a global optimal solution is guaranteed. We have verified the efficacy of our method by extensive experiments on multiple rigid, non-rigid and articulated datasets.

Limitation and Future Work. The computational bottleneck of our method lies in solving the SDPs. For a sequence of m views and n points (for each view), we need to solve m SDPs of size $(n + 1) \times (n + 1)$. Using an interior-point method, one SDP has a worst-case complexity of $O(n^{4.5} \log(1/\epsilon))$ given a solution accuracy $\epsilon > 0$ [13], which remains the limiting factor preventing us from testing on modern large-scale datasets. In the future, we aim to explore the possibility of applying modern large-scale SDP solver, such as [33, 32], to solve our problem more efficiently. Furthermore, we also plan to investigate how to address the degenerate cases as discussed in Sec. 5.2.

Acknowledgement

This research was supported by the Australian Research Council (ARC) through the Centre of Excellence in Robotic Vision, CE140100016, and through Laureate Fellowship FL130100102 to IDR. Y. Dai was supported in part by ARC Grant (DE140100180) and National Natural Science Foundation of China (61420106007).

[5] http://mocap.cs.cmu.edu/

[6] We thank one of the anonymous reviewers for pointing this out.

References

[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 4

[2] A. Chhatkuli, D. Pizarro, and A. Bartoli. Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In *BMVC*, 2014. 6

[3] A. Chhatkuli, D. Pizarro, T. Collins, and A. Bartoli. Inextensible non-rigid shape-from-motion by second-order cone programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1719–1727, 2016. 2, 3, 5, 6, 7, 8

[4] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 1, 2, 5, 6, 8

[5] A. dAspremont and S. Boyd. Relaxations and randomized methods for nonconvex QCQPs. *EE392o Class Notes, Stanford University*, 2003. 4

[6] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013. 6

[7] N. M. Grzywacz and E. C. Hildreth. Incremental rigidity scheme for recovering structure from motion: Position-based versus velocity-based formulations. *JOSA A*, 4(3):503–518, 1987. 2

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 4, 7

[9] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997. 1

[10] R. I. Hartley and P. Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997. 1

[11] H. Li. Multi-view structure computation without explicitly estimating motion. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2777–2784. IEEE, 2010. 1, 3, 5, 6, 7

[12] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, MA Fischler and O. Firschein, eds*, pages 61–62, 1987. 1

[13] Z.-Q. Luo, W.-k. Ma, A. M.-C. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3):20, 2010. 8

[14] A. Mosek. The MOSEK optimization software. *Online at http://www. mosek. com*, 54:2–1, 2010. 5

[15] S. Parashar, D. Pizarro, A. Bartoli, and T. Collins. As-rigid-as-possible volumetric shape-from-template. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 891–899, 2015. 1

[16] P. M. Pardalos and S. A. Vavasis. Quadratic programming with one negative eigenvalue is NP-hard. *Journal of Global Optimization*, 1(1):15–22, 1991. 4

[17] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. In *British Machine Vision Conference*, 2008. 2, 4

[18] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *International Journal of Computer Vision*, 2010. 2, 4

[19] S. Sahni. Computationally related problems. *SIAM Journal on Computing*, 3(4):262–279, 1974. 4

[20] 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. 2, 4

[21] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, 2007. 1

[22] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *European conference on computer vision*, pages 709–720, 1996. 1

[23] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2761–2768, June 2010. 1, 2

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

[25] S. Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London B: Biological Sciences*, 203(1153):405–426, 1979. 1

[26] S. Ullman. Maximizing rigidity: The incremental recovery of 3-d structure from rigid and nonrigid motion. *Perception*, 13(3):255–274, 1984. 1, 2

[27] A. Varol, M. Salzmann, P. Fua, and R. Urtasun. A constrained latent variable model. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2248–2255, 2012. 5, 7

[28] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-free monocular reconstruction of deformable surfaces. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1811–1818, Sept 2009. 2

[29] S. Vicente and L. Agapito. Soft inextensibility constraints for template-free non-rigid reconstruction. In *European Conference on Computer Vision*, pages 426–440, 2012. 2, 5, 6

[30] X. Wang, M. Salzmann, F. Wang, and J. Zhao. Template-free 3d reconstruction of poorly-textured nonrigid surfaces. In *European Conference on Computer Vision*, pages 648–663. Springer, 2016. 2

[31] R. White, K. Crane, and D. A. Forsyth. Capturing and animating occluded cloth. In *ACM Transactions on Graphics (TOG)*, volume 26, page 34. ACM, 2007. 5, 6

[32] L. Yang, D. Sun, and K.-C. Toh. SDPNAL+: a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Mathematical Programming Computation*, 7(3):331–366, 2015. 8

[33] X.-Y. Zhao, D. Sun, and K.-C. Toh. A newton-cg augmented lagrangian method for semidefinite programming. *SIAM Journal on Optimization*, 20(4):1737–1765, 2010. 8