Dense Non-Rigid Structure-from-Motion and Shading with Unknown Albedos

Mathias Gallardo[1], Toby Collins[2] and Adrien Bartoli[1]
[1] EnCoV, IP, UMR 6602 CNRS, Université Clermont Auvergne, SIGMA, France
[2] IRCAD, Strasbourg, France
Mathias.Gallardo@gmail.com, Toby.Collins@gmail.com, Adrien.Bartoli@gmail.com

Abstract

Significant progress has been recently made in Non-Rigid Structure-from-Motion (NRSfM). However, existing methods do not handle poorly-textured surfaces that deform non-smoothly. These are nonetheless common occurrence in real-world applications. An important unanswered question is whether shading can be used to robustly handle these cases. Shading is complementary to motion because it constrains reconstruction densely at textureless regions, and has been used in several other reconstruction problems. The challenge we face is to simultaneously and densely estimate non-smooth, non-rigid shape from each image together with non-smooth, spatially-varying surface albedo (which is required to use shading). We tackle this using an energy-based formulation that combines a physical, discontinuity-preserving deformation prior with motion, shading and contour information. This is a large-scale, highly non-convex optimization problem, and we propose a cascaded optimization that converges well without an initial estimate. Our approach works on both unorganized and organized small-sized image sets, and has been empirically validated on four real-world datasets for which all state-of-the-art approaches fail.

1. Introduction

NRSfM aims to recover the 3D shapes of an object under deformation from apparant motion in a set of 2D images, and is a fundamental and open computer vision problem. NRSfM is challenging because we do not assume any knowledge of the object's 3D shape *a priori*. We also do not assume the object is rigid in some of the images, which prevents initializing the reconstruction using rigid Structure-from-Motion (SfM). NRSfM differs from the related and easier problem of template-based 3D reconstruction (also called *Shape-from-Template* (SfT)), where the object's 3D shape is known in a fixed reference pose. NRSfM is often called *template-free reconstruction* to make a clear distinction, with all input data coming from 2D images.

NRSfM methods have progressed significantly [30, 32, 21, 33, 7, 26], however none of these can handle poorly-textured surfaces and non-smooth deformations such as folds or creases. Firstly, most methods use motion from feature correspondences, which work well only for densely-textured objects with many discriminative features. This is not common in most real practical applications, particularly with man-made objects that usually have very weak texture. These feature-based methods can be divided into two types: those which reconstruct only the features (usually called *sparse reconstructions*) [30, 32, 21, 33, 7, 8, 26], and those which reconstruct the object's surface densely (usually called *dense reconstructions*) [19]. In most applications a dense reconstruction is required. However it is difficult to achieve high-accuracy using feature-based matching, because they only give sparse motion information. Typically non-smooth deformations such as creases and folds can never be accurately reconstructed. For this reason, most of NRSfM methods that densely reconstruct surfaces from feature matches have only been demonstrated on very smooth and well-textured objects such as bending sheets of paper. A direct approach to NRSfM has been recently proposed [34], where dense reconstruction is performed using motion information directly at the pixel level. This works by jointly reconstructing the surface and registering it to each image through intensity-based matching. It was only shown to work on weakly textured surfaces that deform very smoothly. Similarly to feature-based reconstruction, the reason is because *motion information is fundamentally insufficient to reconstruct textureless surface regions undergoing non-smooth deformations*.

Our goal is to solve NRSfM densely by combining motion, shading and a generic physical deformation model that can accurately represent non-smooth deformations. We refer to this problem as NRSfMS (Non-Rigid Structure-from-Motion and Shading). We are specifically interested in solving this problem for objects with unknown spatially-varying albedo, which is the situation in most practical cases. This is because albedos cannot be inferred directly from 2D images of a deforming object. However we must know albedos

in order to apply shading constraints. Therefore our problem is to simultaneously and densely estimate non-smooth, non-rigid shape from each image together with non-smooth, spatially-varying surface albedo. This problem has not been tackled before, and is a crucial missing component for densely reconstructing images in unconstrained settings.

This is very challenging to solve for three main reasons. Firstly, we deal with very high-dimensional deformation spaces, which are needed to model non-smooth deformations such as surface creases that can form in arbitrary places. One cannot therefore approximate the problem using a globally smooth surface representation (as is common in previous dense NRSfM methods), which both increases the search space dramatically, and leads to a highly non-convex energy landscape. Secondly, we cannot apply shading to constrain non-rigid shape without knowing surface albedo. Similarly, through shading surface albedo constrains non-rigid shape. We must therefore simultaneously and densely estimate both, which is a highly non-convex problem. Thirdly, the fact that albedo is spatially varying significantly complicates the problem. Fundamentally, it requires us to densely register the image data, which is challenging, particularly at weakly-textured regions.

We use a dense triangulated mesh model to represent the object's surface, and the objective is to estimate the mesh's vertex positions in camera coordinates for all input images, together with surface albedos. We approach the problem with an energy-based formulation that combines the physical, discontinuity-preserving deformation prior with motion, shading and contour information. This is a large-scale, highly non-convex optimization problem, and we propose a cascaded optimization that converges well without an initial estimate. Because this is the first approach to solve NRSfMS, we also include an empirical analysis of the problem's stability using perturbation analysis. We provide real experiments with ground truth data and show that our method can accurately reconstruct dense shape where existing state-of-the-art NRSfM methods fail.

2. State-of-the-Art

2.1. Non-Rigid Structure-from-Motion

There is a substantial number of approaches to NRSfM and we do not attempt a detailed review here. The methods can be broken down along a number of different dimensions, and the main ones are as follows: *(i)* the type of camera model used, such as perspective or affine; *(ii)* if the approach uses features or direct pixel-level matching; *(iii)* if it operates on a batch of images or sequentially on video frames; *(iv)* what surface representation is used, such as a pointcloud, a spline surface or a mesh surface; *(v)* if the problem formulation is convex or non-convex; *(vi)* what deformation prior is used. There is no general consensus on the best way to formulate NRSfM according to the above dimensions. The current trend focuses on handling videos (because temporal continuity can be exploited), perspective cameras (because these are generally the most accurate) and mesh surface representations (because they are simple and can model arbitrary topology).

Feature-based approaches are the most common way to tackle NRSfM. These are popular because the problem of motion estimation (through feature matching) can be strongly decoupled from 3D reconstruction. Recently, some featureless NRSfM methods have been proposed [19, 34]. These either decouple motion estimation from 3D reconstruction, by estimating motion with multi-view optic flow [19], or by jointly reconstructing the surface and motion estimation [34]. The advantage of decoupling motion estimation is to simplify the reconstruction problem. However, the disadvantage is that the results can be sub-optimal and mistakes in the estimated motion cannot be corrected.

To overcome measurement noise and ambiguities in NRSfM, two classes of deformation priors have emerged: statistical [11, 30, 15, 21, 1] and physics-based [7, 26, 34, 32, 33, 8] priors. The first class is based on the assumption that the space of object shapes or object deformations lies on a low-dimensional manifold which can be learned jointly with reconstruction. In all such approaches, the manifold is modeled by a linear combination of shape bases [5], which must be estimated during reconstruction. These approaches give good results for objects with a strong rigid component, such as human faces. However, they often require a large number of images and are not suitable for objects with high deformation spaces such as deforming fabric, or objects that can crease or fold in unexpected ways. Physics-based deformation models operate very differently to statistical models, and restrict the space of possible deformations according to physical properties of the object's material. The most common physics-based model is isometry or quasi-isometry [7, 26, 34, 32, 33, 8]. These assume the object's surface does not stretch or shrink significantly as it deforms. Quasi-isometry means that there is non-negligible stretching or shearing, but the model prefers solutions that minimize stretching or shrinking. These models have been used extensively because they dramatically restrict the solution space, and are applicable for many object classes such as those made of thick rubber, tightly-woven fabrics, paper, cardboard and plastics. It appears that NRSfM with the isometric model is well-posed if motion can be estimated densely across the object's surface [32, 7, 26].

2.2. Shading in Other 3D Reconstruction Problems

Shading has been used previously in several other 3D reconstruction problems. These include Shape-from-Shading (SfS) [28, 13, 14] and photometric stereo [3, 35, 6], rigid SfS [22, 23, 24] and SfT [20, 25, 24, 18]. In SfS, shading is

used to reconstruct a depth-map from a single image. However, it has had very limited success because it is a weakly constrained problem with one constraint at each pixel, and is only solvable if accurate models of surface reflectance (including albedos) and scene illumination are known *a priori*. SfS also has tremendous difficulty with external and self-occlusions. If albedos are unknown, the problem is ill-posed. Thus, SfS cannot be used to reconstruct the datasets in this paper. The usual way forward is to assume that albedo is constant, which makes the problem solvable up to scale and the bas relief ambiguity. Photometric stereo is the extension of SfS using multiple light sources and has shown great success for reconstructing high-accuracy surface details with unknown albedo *e.g.* [3, 35]. However requires a special hardware setup where the scene is illuminated by a sequence of lights placed at different points in the scene, during which time the scene is assumed to be rigid. This setup is not applicable in many situations. Shading has also been used in rigid SfM to achieve high-accuracy at both textured and textureless regions. Unlike SfS, this works using multiple images and combines motion with shading information. However deformable objects are not handled.

In SfT, the problem is to register a deformable 3D model (also called an *object template*) in 3D camera coordinates using visual information present in a 2D image. In practice the object template can be built from a CAD model or built from images of object in an undeformed state using dense multiview SfM [10, 2, 25, 24]. Most SfT methods used only motion information, though shading information has been recently introduced to handle weakly-textured objects [20, 25, 24, 18]. Compared to NRSfMS, SfT is a considerably easier problem because the object template is determined *a priori*. By contrast in NRSfMS there is no *a priori* object template.

3. Problem Modeling and Optimization

3.1. Modeling Assumptions and Inputs

To solve NRSfMS, modeling assumptions are required for the following: shape deformation, albedo, surface reflectance, scene illumination, camera response and scene geometry. We now list our assumptions. *Shape deformation*: we use a quasi-isometric prior and discontinuity-preserving smoother that favours piecewise-smooth deformations. This has been shown previously to be a good model for handling non-smooth, creasing surfaces [18]. We assume the surface does not tear over time. *Albedo*: we assume albedo is constant over time, and piecewise-constant over the surface. This is applicable for many objects and particularly man-made ones. The piecewise-constant assumption is used to reduce ambiguity between smooth intensity variation caused by albedo variation versus surface gradient variation. We do not assume albedo is pre-segmented. *Surface reflectance*: we use a Lambertian model, which gives a good approximation of many surfaces, and we handle modeling errors due to *e.g.* specular reflections with robustification (see §3.3). *Illumination*: we assume it is constant, pre-calibrated and defined in camera coordinates. In practice this can be done if we have a camera-light rig setup such as an endoscope or camera with flash, or a non-rig where the light and camera are not physically connected but do not move relative to each other during image acquisition. The model we use in our experiments is the second-order spherical harmonic model, which are very common in SfS with 9 parameters. *Camera response function*: we assume it is known *a priori*, or linear and constant over time. *Scene geometry*: we assume no self or external occlusions, which is a typical assumption in state-of-the-art dense NRSfM and there can be background clutter. Our model and algorithm may in principle use a reference view where the object's surface may be smooth or creased. In practice however, we have obtained better reconstruction accuracy for a smooth reference view. Our investigation of why it happens so has not revealed a clear reason so far and we chose to leave this point for future research.

We use $t = 1, \ldots, N$ as the image index. Our inputs are as follows. *(i)* a set of N RGB images $\{I_t\}$, $I_t : \mathbb{R}^2 \to [0, 255]^3$ and the corresponding luminance images $\{L_t\}$, $L_t : \mathbb{R}^2 \to \mathbb{R}^+$. The luminance image stores the light intensity striking the image plane at each pixel coordinate, and if the camera response function is known it can be built from I_t. If it is unknown, camera response is assumed to be constant and linear. In this case we set L_t as the pixel intensity, which gives luminance up to a global scale factor. This global scale factor can be absorbed into the surface albedos and has no effect on the reconstruction problem. *(ii)* the camera intrinsics, denoted by the functions $\pi_t : \mathbb{R}^3 \to \mathbb{R}^2$ which project from 3D camera coordinates to pixel coordinates. *(iii)* a segmentation of the object of interest in one of the images, which we call the *reference image*, denoted by the region $\Omega \subset \mathbb{R}^2$. Without loss of generality let this be the first image. *(iv)* the scene illumination coefficients which we denote by $\mathbf{l} \in \mathbb{R}^{4 \text{ or } 9}$. *(v)* a set of putative 2D correspondences from Ω to all other images. These are assumed to be mostly correct with some outliers, and can be computed using existing methods such as SURF or SIFT. We denote this by the sets $\{\mathcal{S}_t\}$.

3.2. Shape and Albedo Modeling

We use a regular triangular 3D mesh to model the object's 3D surface, which we build by meshing Ω using a regular 2D triangular mesh, with M vertices (we use a default mesh grid of 100×100 vertices that encloses Ω). We denote $\mathcal{E}$ as the mesh's edges, where N_E is the number of edges. Because we assume the surface does not tear the mesh edges are fixed over time. Our task is to determine, for each mesh

vertex i, its position $\mathbf{v}_t^i \in \mathbb{R}^3$ in 3D camera coordinates for each image $t \in [1, N]$. We use $\mathcal{V}_t = \{\mathbf{v}_t^i\}_{i\in[1,M]}$ to denote the vertices in 3D camera coordinates for image t.

We parameterize $\mathcal{V}_1$ along lines-of-sight. Specifically, let $\mathbf{u}_i \in \mathbb{R}^2$ denote the 2D position of the i^{th} vertex in the first image, defined in normalized pixel coordinates. Its corresponding position in 3D camera coordinates at $t = 1$ is $\mathbf{v}_i^1 = d_i[\mathbf{u}_i^\top, 1]^\top$, where d_i is its unknown depth. We collect these unknown depths into the set $\mathcal{D} = \{d_1, \ldots, d_N\}$. The full set of unknowns that specify the object's shape in all images is therefore $\{\mathcal{D}, \mathcal{V}_2, \ldots, \mathcal{V}_N\}$, which corresponds to $3M(N-1)+M$ real-valued unknowns. We use the mesh to transform any 2D point $\mathbf{u} \in \Omega$ to 3D camera coordinates using $\mathcal{V}_t$, which is done using a barycentric interpolation (a linear interpolation of the positions of the three vertices surrounding $\mathbf{u}$). We denote this by $f(\mathbf{u}; \mathcal{V}_t) : \Omega \to \mathbb{R}^3$. The surface normal at $\mathbf{u}$ is computed from the enclosing triangle, denoted by $n(\mathbf{u}; \mathcal{V}_t) : \Omega \to \mathcal{S}_3$. Unlike f, n is non-linear. We define an *albedo-map* $A(\mathbf{u}) : \Omega \to \mathbb{R}^+$ as the function that gives the unknown albedo for a pixel $\mathbf{u} \in \Omega$. From the piecewise-constant assumption we can write this as $A(\mathbf{u}) : \Omega \to \mathcal{A}$ where $\mathcal{A} = \{\alpha_1, \ldots \alpha_K\}$ denotes a discrete set of K unknown albedos with $\alpha_k \in \mathbb{R}^+$. We discuss how $\mathcal{A}$ is built in §3.4.

3.3. The Energy-based Objective Function

We construct the energy-based objective function by combining physical priors with shading, point correspondence and boundary contour information extracted from all images. The objective function C has the following form:

$$C(\mathcal{V}_1, \ldots, \mathcal{V}_N, \mathcal{A}) \triangleq \sum_{t=1}^{N} C_{shade}(\mathcal{V}_t, \mathcal{A}; L_t, \mathbf{l}) + \quad (1)$$
$$\lambda_{corresp} C_{corresp}(\mathcal{V}_t; \mathcal{S}_t) + \lambda_{contour} C_{contour}(\mathcal{V}_t; \mathcal{I}_t) +$$
$$\lambda_{iso} C_{iso}(\mathcal{V}_1, \mathcal{V}_t) + \lambda_{smooth} C_{smooth}(\mathcal{V}_t).$$

The terms C_{shade}, $C_{corresp}$ and C_{bound} are shading, motion and boundary data terms respectively. The terms C_{smooth} and C_{iso} are physical deformation prior terms. The terms $\lambda_{corresp}$, λ_{bound}, λ_{iso} and λ_{smooth} are positive weights and are the method's tuning parameters.

The shading term. The shading term robustly encodes the Lambertian relationship between albedo, surface irradience and pixel luminance. We evaluate it as:

$$C_{shade}(\mathcal{V}_t, \mathcal{A}; L_t, \mathbf{l}) \triangleq \quad (2)$$
$$\frac{1}{|\Omega|} \sum_{\mathbf{u}\in\Omega} \rho\left(\mathcal{A}(\mathbf{u})\, r\left(n(\mathbf{u}; \mathcal{V}_t); \mathbf{l}\right) - L_t\left(\pi_t \circ f(\mathbf{u}; \mathcal{V}_t)\right)\right).$$

The function $r(\mathbf{n}; \mathbf{l})$ evaluates the Lambertian irradiance for a normal vector $\mathbf{n}$ according to the spherical harmonics model with light coefficients $\mathbf{l}$. The function $\rho : \mathbb{R} \to \mathbb{R}$ is an *M-estimator* which is used to enforce similarity between the modeled and measured luminance, while also allowing for some points to violate the model (caused by specular reflection, small shadows and other unmodeled factors). We use the Huber M-estimator with free parameter set to 0.005.

The correspondence term. Recall that the set $\mathcal{S}_t$ holds putative correspondences between Ω and image $t \in [2, N]$. We denote this by $\mathcal{S}_t = \{(\mathbf{u}_j, \mathbf{q}_j)\}_{j=1}^{s(t)}$, where $\mathbf{u}_j$ denotes the correspondence position in Ω and $\mathbf{q}_j$ denotes the corresponding position in image t. The number of correspondences are denoted by $s(t)$ which varies in general between images. The term robustly encourages each point $\mathbf{u}_j$ to transform to its corresponding point $\mathbf{p}_j$, and is given by:

$$C_{corresp}(\mathcal{V}_t; \mathcal{S}_t) \triangleq \sum_{(\mathbf{u}_j, \mathbf{p}_j)\in\mathcal{S}_t} \rho\left(\|\pi_t \circ f(\mathbf{u}_j; \mathcal{V}_t) - \mathbf{q}_j\|\right). \quad (3)$$

The boundary contour term. This constraint works for surfaces with disc topology. It encourages the surface's boundary contour to lie close to image edges, and was shown to significantly help register surfaces with weak texture [17, 24]. We discretize the boundary of Ω to obtain a set of boundary pixels $\mathcal{B} \triangleq \{\mathbf{u}_{k\in[1,B]}\}$. We then compute a 'boundariness map' for each image $E_t : \mathbb{R}^2 \to \mathbb{R}^+$ where high values of $E_t(\mathbf{p})$ correspond to a high likelihood of pixel $\mathbf{p}$ being on the boundary contour. The term is evaluated as:

$$C_{bound}(\mathcal{V}_t; E_t) \triangleq \frac{1}{|B|} \sum_{\mathbf{u}_k\in\mathcal{B}} \rho\left(E_t(\pi_t \circ f(\mathbf{u}_k; \mathcal{V}_t))\right). \quad (4)$$

We found that E_t cannot be built naively using for instance an edge response filter, because of many false positives, particularly with background clutter and strong object texture. Instead we build it using an edge response filter that is modulated to suppress false positives according to one or more segmentation cues. The right cue depends on the particular dataset, for example if the background is constant over the image set, or if the object has a distinct color distribution to the background. We give the exact formula for computing E_t for each tested dataset in the supplementary material.

The smoothing term. The quasi-isometry term penalizes within-plane stretching or shearing but not curvature change. Thus it is insufficient to use as a regularizer for mitigating noise. We deal with this using a discontinuity-preserving smoother which automatically deactivates smoothing where needed at creased regions. This

is based on [17] where it was used in the SfT problem:

$$C_{smooth}(\mathcal{V}_t) \triangleq \frac{1}{|\Omega|} \sum_{\mathbf{u}_j \in \Omega} \rho \left(\frac{\partial^2 f}{\partial \mathbf{u}^2}(\mathbf{u}_j; \mathcal{V}_t) \right). \quad (5)$$

The quasi-isometry term. We enforce quasi-isometry using mesh edge-length constancy. Specifically, we measure constancy with respect to the mesh edges in the reference image. This is defined as follows:

$$C_{iso}(\mathcal{V}_1, \mathcal{V}_t) \triangleq \frac{1}{|E|} \sum_{(i,j) \in E} \left(1 - \|\mathbf{v}_1^i - \mathbf{v}_1^j\|_2^{-2} \|\mathbf{v}_t^i - \mathbf{v}_t^j\|_2^2 \right)^2. \quad (6)$$

This penalizes a change in edge length *relative* to the reference mesh, and unlike many other ways to impose isometry, is invariant to a global scaling of the reconstruction.

3.4. Optimization strategy

Overview. Optimizing equation (1) is a non-trivial task because it is large-scale (typically $O(10^5)$ unknowns), is highly non-convex, and the shading term requires dense, pixel-level registration. Recall that we do not assume the images come from an uninterrupted video sequences, which makes dense registration much harder to achieve. Our strategy is to first achieve a rough initial estimate for the shape terms $(\mathcal{D}, \mathcal{V}_2, \dots \mathcal{V}_N)$ (and hence an initial estimate for registration) using only motion constraints from the point correspondences. We then introduce the contour boundary constraints and refine these estimates by optimizing equation (1) using iterative numerical minimization. Next we estimate albedos by fixing the shape terms, and finally optimize equation (1) over all unknowns using all information (point correspondences, boundary contours and shading) using iterative numerical minimization. We propose this strategy because point correspondences can be used to provide a rough, smooth solution to non-rigid shape without requiring an initial estimate. By contrast we find that boundary contour and shading terms require a good initialization to prevent incorrect convergence in a local minimum. Concretely, our optimization strategy is divided into 4 stages which we now describe in detail.

Stage 1: Correspondence-based template initialization. We take the point correspondences $\{\mathcal{S}_t\}$ between the reference image and the other images and input them to an existing surface-based, initialization-free NRSfM method. The method we currently used is [7] which has publicly available code. This provides us with a rough estimate of the reference image's vertex depths $\mathcal{D}$. Note that all existing initialization-free surface-based methods assume the object's surface is smooth in all views, thus the initial estimate will not normally be highly accurate.

Stage 2: Motion and contour-based Shape-from-Template. We back-project the mesh vertices in the reference view using their initial depth estimates $\mathcal{D}$. This gives a rough estimate of the object's 3D shape in a reference position (corresponding to the reference image). We then use this mesh as an object template, and call an existing SfT method to initialize, for each image, the vertex positions $\mathcal{V}_t$ using the correspondence set $\mathcal{S}_t$. The current method we use is [17]. We then optimize equation (1) without shading by setting $\lambda_{shade} = 0$, over the shape unknowns $\{\mathcal{V}_2, \dots \mathcal{V}_N\}$ with $\mathcal{D}$ kept fixed. This can be done efficiently because the unknowns are now decoupled between images, so each $\mathcal{V}_t$ can be minimized independently. Finally we optimize equation (1) over all shape unknowns $\{\mathcal{V}_2, \dots \mathcal{V}_N\}$ with $\lambda_{shade} = 0$. To achieve good convergence we compute the boundary distance map (equation (4)) with an image pyramid, using 3 levels with one octave per level.

Stage 3: Albedo initialization. We now use our current shape estimates to infer albedos using the shading term. For this we segment the reference image into local superpixel-like clusters, where within each cluster we assume the albedo is constant. Such segmentation will never be perfect, so to handle this we aim for an *oversegmentation*, where neighboring segments can share the same albedo but within the same segment we want the albedo to be constant. We achieve this by performing an intrinsic images decomposition [4] on the reference view's intensity image and cluster the resulting 'reflectance image' using [16] with a low cluster tolerance (we use a default of 10). For each cluster k, we assign a corresponding albedo α_k. This is done by taking each pixel $\mathbf{u}_j$ in the cluster, estimating its albedo by inverting the shading equation: $\alpha \approx L_t(\pi_t \circ f(\mathbf{u}; \mathcal{V}_t))\, r(n(\mathbf{u}; \mathcal{V}_t); \mathbf{l})^{-1}$. We then initialize α_k as the median over all estimates within the cluster.

Stage 4: Full refinement. We refine our estimates by minimizing equation (1) using all terms and over all unknowns, which is achieved with Gauss-Newton iterative optimization and backtracking line-search. Because of the very large number of unknowns, at each iteration we solve the normal equations using an iterative solver (diagonally-preconditioned conjugate gradient), with a default iteration limit of 200. Recall that there is a scale ambiguity (as in all NRSfM problems), because we cannot differentiate a smaller surface viewed close to the camera from a large surface viewed far away. We fix the scale ambiguity by scaling all vertices to have a mean depth of 1 after each iteration. To achieve good convergence we blur each L_t with a Gaussian blur pyramid, with a default of three octaves. For the first two pyramid levels we run Gauss-Newton until either convergence is reached or a fixed number of iterations have passed (we use 20 iterations). For the final pyramid level

we run it until convergence. Processing time is typically several minutes for small-sized image sets (<10 images), with a sub-optimal Matlab implementation on the CPU.

4. Experimental Validation

4.1. Overview

We divide the experimental validation into two parts. In the first part we analyze the convergence basin of our energy function through perturbation analysis. This is to understand both how sensitive our formulation is to the initial solution, and fundamentally, whether NRSfMS can be cast as an energy-based minimization with a strong local minimum near the true solution. In the second part we compare performance to state-of-the-art NRSfM methods. Our evaluation has been performed using public datasets and a new dataset, all with ground truth.

4.2. Method Comparison and Accuracy Metrics

We compare with the following competitive NRSfM methods [30, 32, 29, 33, 7, 9, 26], denoted respectively with **EM08**, **PP09**, **LRG10**, **SI12**, **IP14**, **MDH16** and **LT16**. **EM08**, **LRG10**, **SI12** and **MDH16** are methods which reconstruct only point correspondences, whereas **PP09**, **IP14** and **LT16** are methods which reconstruct dense surfaces. To see the contribution of some terms of (1), we compare with two versions of our method, **NoS**, where shading is not used, and **NoB**, where the boundary constraint is not used in stages 2 and 4. We have evaluated on 4 datasets (three public and one new). Each dataset consists of a disc-topology surface in 5 different deformed states, with one state per image. We show these in figure 3. From top down we have 'floral paper' from [18], 'paper fortune teller' from [18], 'creased paper' (new) and 'Kinect paper' from [31]. 'Kinect paper' is a video dataset and has no accompanying illumination parameters and no camera response function. We approximated camera response with a constant linear model, and estimated the illumination parameters using the image data and the accompanying depthmaps. This was performed by selecting in a small rectangular region on the surface with both constant albedo and non-saturated pixels, then measuring the average pixel intensity within the region and fitting a local plane to the region using the depthmap. This was repeated using 30 images in the sequence, and we then estimated the spherical harmonics illumination vector by inverting the Lambertian shading model using linear least squares. The 5 images we used for evaluation were uniformly sampled from the video. We followed the same procedure as described in [18] to make the 'creased paper' dataset, with sub-millimetre accuracy depth-maps computed by a structured light system [12]. Images were captured with standard PointGrey RGB camera [27] with a linear camera response. Each dataset has a set of point correspondences between the first and all other images. As the three first datasets are poorly-textured, the correspondences are sparse. We plot them in each input image in figure 3. For 'creased paper', the texture is repetitive and wide-baseline feature matches fail, so we found them manually at approximately 20 corner locations. We note that manual correspondences are commonly used to evaluate NRSfM methods. The tuning-parameters of all methods were manually adjusted to obtain the most visually pleasing results.

We measure accuracy by comparing 3D distances with respect to ground truth. Because reconstruction is up to scale, we compute for each method the best-fitting scale factor that aligns the predicted point correspondences with their true locations in the $L2$ sense, then measure accuracy with the scale-corrected reconstruction. This was done at three locations: *(i)* at point correspondences, *(ii)* densely across the ground truth surface, and *(iii)* densely at 'creased regions', which are any points on the ground truth surfaces that are within 5mm of a surface crease. Note that a dense ground truth registration is not available on the datasets, however we do have ground truth registration at the true point correspondences. Thus we measure *(i)* by comparing predicted surface normals and 3D positions at each correspondence with the ground truth values. For *(ii)* and *(iii)* we compare normals and 3D positions for each ground truth surface point with the nearest reconstructed surface point. For methods which only reconstruct the correspondences, we can only measure the 3D point correspondence error.

4.3. Quantitative and Qualitative Results

We show in figure 3 the test datasets and a reconstruction of one of the images per dataset from our method and the best performing previous method (the one with lowest Root Mean Squared Error (RMSE) with respect to *(ii)* above). Visually we can see that considerable surface detail is accurately reconstructed by our method as well as the global shape.

In figure 2 we give the RMSE across all test datasets and all compared methods. The first row gives from left to right the distance RMSE at point correspondences *(i)*, over the whole GT surfaces *(ii)* and over creased regions in the GT surfaces *(ii)*. The second row gives the respective surface normal RMSEs. 'Kinect paper' has no creases and the deformation is very smooth in all images. We observe that, for all datasets other than 'Kinect paper', there is a good improvement with respect to all error metrics compared to the other methods. This is strongest in the second and third columns, which show our method successfully exploits shading information in textureless and creased regions. For 'Kinect paper' we see that our method does not obtain the highest accuracy across all error metrics. The reason is that it is a very smooth, densely textured surface, and shading is not needed to achieve an accurate reconstruction.

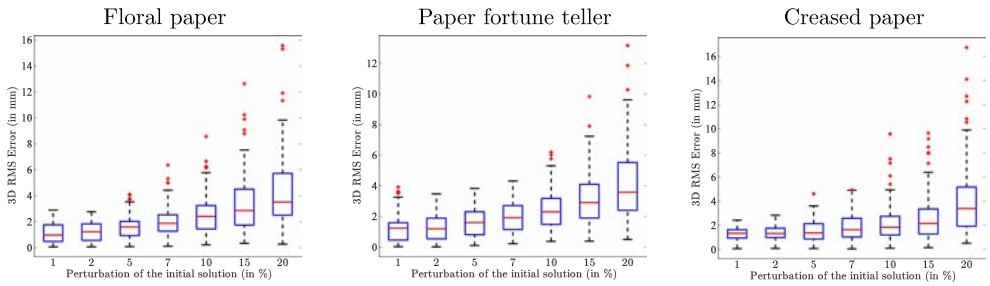

Figure 1. Numerical results of the convergence basin analysis.

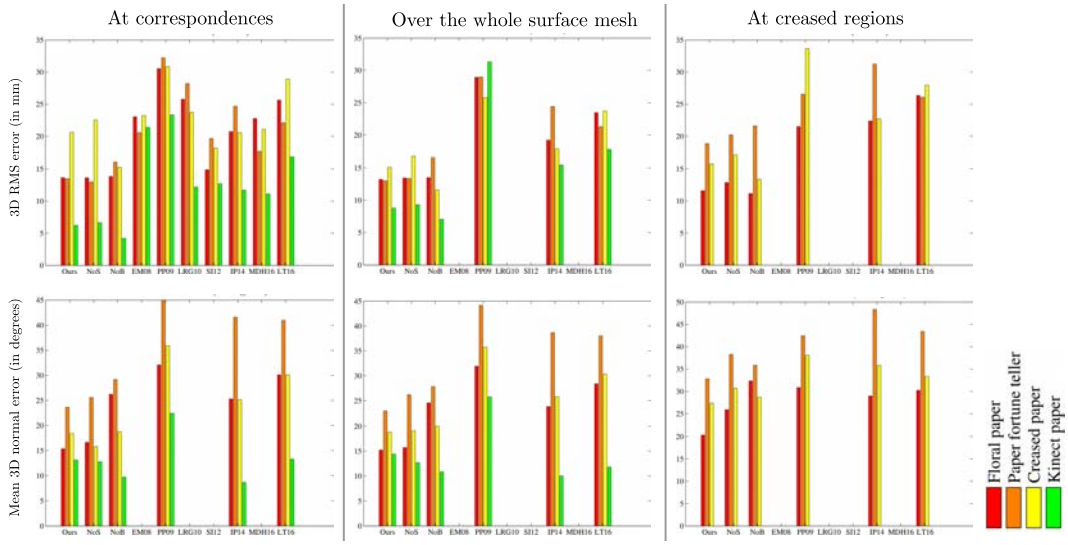

Figure 2. Reconstruction accuracy statistics across all test datasets and all compared methods. We recall that **EM08**, **LRG10**, **SI12** and **MDH16** reconstruct only point correspondences, whereas **PP09**, **IP14** and **LT16** reconstruct dense surfaces. Also, the 'Kinect paper' dataset does not present any crease.

However, our method still obtain competitive results on this dataset. We observe that the use of shading improves globally the shape of the reconstructions and that the boundary constraint allows a better use of shading.

4.4. Convergence Basin Analysis

We performed perturbation analysis as follows. We started with an initial reconstruction close to the ground truth, then applied a low-pass filter (to smooth out creases, because we do not expect these to be present in the initial solution), then randomly perturbed the vertex positions using smooth deformation functions. For each perturbation we optimized equation (1) by performing stages 3 and 4 in §3.4. The initial reconstruction was carefully done by hand, using the ground truth surfaces, point correspondences, and a quasi-isometric nonrigid ICP registration. The perturbations were designed to globally deform the initial solutions, which is more realistic than a local perturbation of each vertex. This was implemented using a $4 \times 4 \times 4$ b-spline enclosing the reconstructed surfaces and randomly perturbing the spline's control points at 7 different noise levels, with 30 random perturbations per noise level. We report results as box-plots for the 'floral paper', 'paper fortune teller' and 'creased paper' datasets in figure 1. The x axis gives the average perturbation in mm for each noise level from the initial solution. The y axis gives the dense surface RMSE as defined in for each random sample. For small noise levels ($< 5\%$), the box-plots are very similar, which tells us our energy landscape has a strong local minimum close to the ground truth, which supports our claim that NRSfMS can be cast as an energy-based minimization (via equation (1)), with a strong local minimum near the true solution. For larger noise levels ($> 5\%$) we can see a significant increase in error, indicating that the optimization now becomes trapped more frequently in local minima.

4.5. Failures Modes

The main failure mode is if a good initial solution cannot be obtained after stages 1 and 2. Typically this occurs if there are very few, poorly-distributed point correspondences. In these cases it is difficult to initialize dense shape with any current NRSfM method. For unorganized

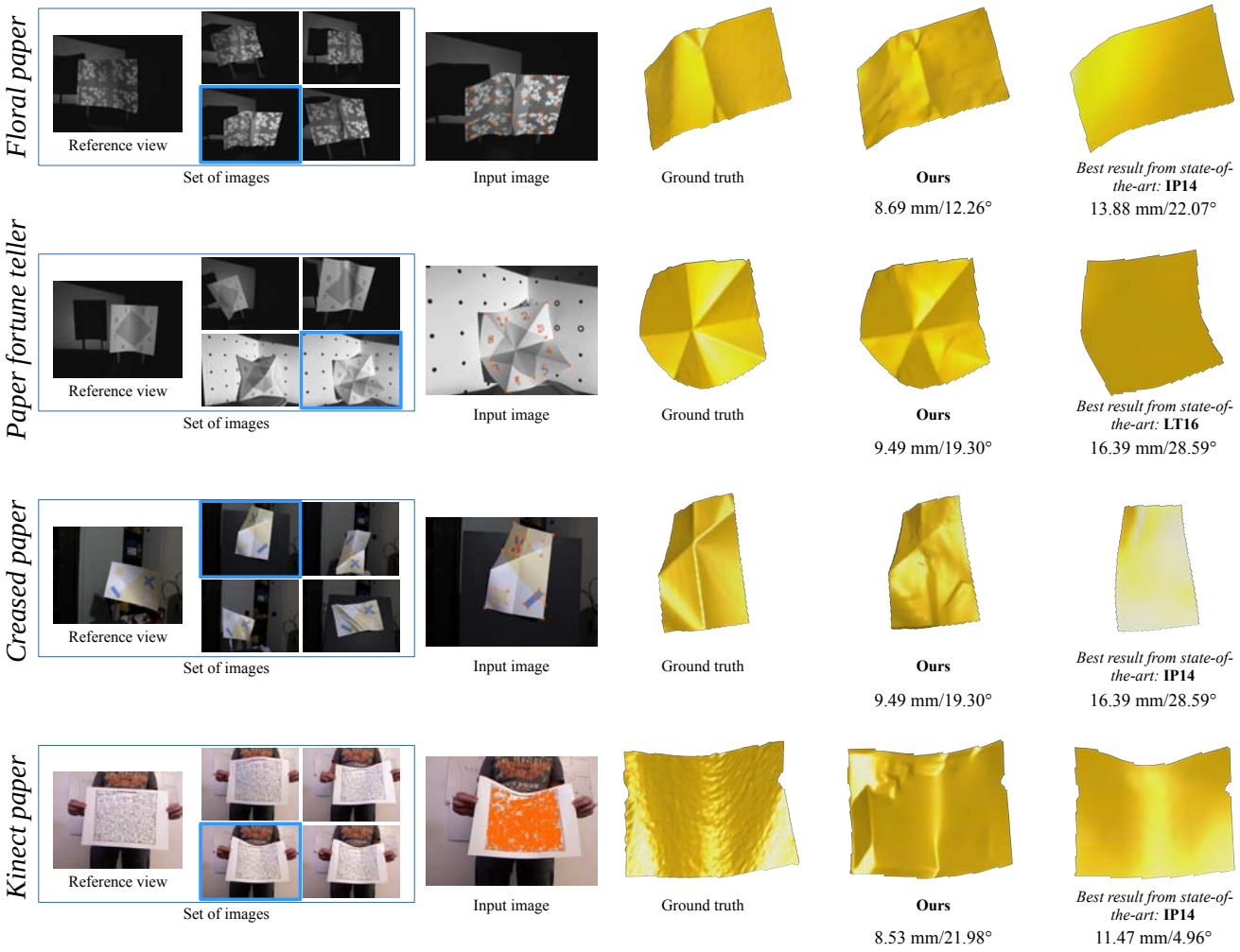

Figure 3. The four real-world test datasets and results visualizations. Here we show the images from each dataset, and sample reconstructions from one of the images using our method and the best performing NRSfM method. Under each reconstruction, we give the 3D RMS and mean 3D normal errors over the whole surface mesh. In each input image, we show the correspondences with an orange cross sign.

image sets this is a difficult problem to overcome. For video sequences, dense point correspondences can usually be obtained by exploiting temporal continuity and dense frame-to-frame tracking [19]. Another failure mode is the presence of some false positive creases. They may be caused by the robust estimator applied in the shading term. Such artefacts can be reduced by interleaving intensity edge/shape-edge aware filtering with the optimization framework, which we let for future works.

5. Conclusion

We have studied the problem of NRSfMS with unknown, spatially varying albedos. This is a hard and important vision problem, needed for high-accuracy dense reconstruction of weakly-textured surfaces undergoing non-smooth deformation from 2D images. We have proposed an energy-based solution and a cascaded numerical optimization strategy, and have demonstrated encouraging results on four real-world datasets, for which all competitive NRSfM methods fail. This marks the first time that strongly creased, deformable, low-textured surfaces with unknown albedos have been densely reconstructed and registered from 2D image sets without a 3D template. Our work is the basis for many future directions, including handling non-smooth reference views and unknown light, modeling occlusions and shadows, developing an incremental version to handle large image sets, and a theoretical study of well-posedness.

Acknowledgments. This research has received funding from the EU's FP7 through the ERC research grant 307483 FLEXABLE and from Almerys Corporation.

References

[1] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid Structure from Motion in Trajectory Space. In *Advances in Neural Information Processing Systems 21*, pages 41–48. 2009.

[2] A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro. Shape-from-Template. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2099–2118, 2015.

[3] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-Quality Single-Shot Capture of Facial Geometry. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 2010.

[4] S. Bell, K. Bala, and N. Snavely. Intrinsic Images in the Wild. *ACM Trans. on Graphics (SIGGRAPH)*, 2014.

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering Non-Rigid 3D Shape from Image Streams. In *CVPR*, 2000.

[6] G. J. Brostow, C. Hernndez, G. Vogiatzis, B. Stenger, and R. Cipolla. Video Normals from Colored Lights. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2104–2114, 2011.

[7] A. Chhatkuli, D. Pizarro, and A. Bartoli. Non-Rigid Shape-from-Motion for Isometric Surfaces using Infinitesimal Planarity. In *BMVC*, 2014.

[8] A. Chhatkuli, D. Pizarro, 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*, 2016.

[9] A. Chhatkuli, D. Pizarro, T. Collins, and A. Bartoli. Inextensible Non-Rigid Shape-from-Motion by Second-Order Cone Programming. In *CVPR*, June 2016.

[10] T. Collins and A. Bartoli. Realtime Shape-from-Template: System and Applications. In *International Symposium on Mixed and Augmented Reality*, 2015.

[11] 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.

[12] David 3D Scanner. `http://www.david-3d.com/en/products/david4`, 2014.

[13] J.-D. Durou, M. Falcone, and M. Sagona. Numerical Methods for Shape-from-Shading: A New Survey with Benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, Jan. 2008.

[14] A. Ecker and A. D. Jepson. Polynomial shape from shading. In *CVPR*, June 2010.

[15] J. Fayad, A. D. Bue, L. Agapito, and P. Aguiar. Non-Rigid Structure from Motion using Quadratic Deformation Models. In *BMVC*, 2009.

[16] K. Fukunaga and L. Hostetler. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. IEEE *Transactions on Information Theory*, 21(1):32–40, Jan 1975.

[17] M. Gallardo, T. Collins, and A. Bartoli. Can we Jointly Register and Reconstruct Creased Surfaces by Shape-from-Template Accurately? In *ECCV*, 2016.

[18] M. Gallardo, T. Collins, and A. Bartoli. Using Shading and a 3D Template to Reconstruct Complex Surface Deformations. In *BMVC*, 2016.

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

[20] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, volume 32, pages 158:1–158:10, November 2013.

[21] P. F. U. Gotardo and A. M. Martinez. Kernel Non-Rigid Structure from Motion. In *ICCV*, November 2011.

[22] H. Jin, D. Cremers, D. Wang, E. Prados, A. Yezzi, and S. Soatto. 3-D Reconstruction of Shaded Objects from Multiple Images Under Unknown Illumination. *International Journal of Computer Vision*, 76(3):245–256, 2008.

[23] K. Kim, A. Torii, and M. Okutomi. *Multi-view Inverse Rendering Under Arbitrary Illumination and Albedo*. 2016.

[24] Q. Liu-Yin, R. Yu, L. Agapito, A. Fitzgibbon, and C. Russell. Better Together: Joint Reasoning for Non-rigid 3D Reconstruction with Specularities and Shading. In *BMVC*, 2016.

[25] A. Malti and A. Bartoli. Combining Conformal Deformation and Cook-Torrance Shading for 3D Reconstruction in Laparoscopy. IEEE *Transactions on Biomedical Engineering*, 61(6):1684–1692, June 2014.

[26] S. Parashar, D. Pizarro, and A. Bartoli. Isometric Non-rigid Shape-from-Motion in Linear Time. In *CVPR*, June 2016.

[27] Point Grey. Flea2G 1.3 MP Color Firewire 1394b (Sony ICX445). `https://www.ptgrey.com/`.

[28] E. Prados and O. Faugeras. Perspective shape from shading and viscosity solutions. In *ICCV*, volume 2, pages 826–831, Nice, France, Oct 2003. IEEE Computer Society.

[29] J. Taylor, A. D. Jepson, and K. Kutulakos. Non-Rigid Structure from Locally-Rigid Motion. In *CVPR*, 2010.

[30] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid Structure-from-Motion: Estimating Shape and Motion with Hierarchical Priors. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 30(5):878–892, 2008.

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

[32] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-Free Monocular Reconstruction of Deformable Surfaces. In *ICCV*, 2009.

[33] S. Vicente and L. Agapito. Soft Inextensibility Constraints for Template-Free Non-rigid Reconstruction. In *ECCV*, June 2012.

[34] X. Wang, M. Salzmann, F. Wang, and J. Zhao. Template-Free 3D Reconstruction of Poorly-Textured Nonrigid Surfaces. In *ECCV*, 2016.

[35] Z. Zhou, Z. Wu, and P. Tan. Multi-view Photometric Stereo with Spatially Varying Isotropic Materials. In *CVPR*, June 2013.