

Chapter 4

Non-Rigid Structure-from-Motion and Shading



Mathias Gallardo, Toby Collins, and Adrien Bartoli

Abstract We show how photometric and motion-based approaches can be combined to reconstruct the 3D shape of deformable objects from monocular images. We start by motivating the problem using real-world applications. We give a comprehensive overview of the state-of-the-art approaches and discuss their limitations for practical use in these applications. We then introduce the problem of Non-Rigid Structure-from-Motion and Shading (NRSfMS), where photometric and geometric information is used for reconstruction, without prior knowledge about the shape of the deformable object. We present in detail the first technical solution to NRSfMS and close the chapter with the main remaining open problems.

4.1 Introduction

Deformable 3D reconstruction aims to recover the 3D shape of deformable objects from one or more 2D images. While 3D reconstruction of rigid objects is well understood with robust methods and commercial products [1, 2], deformable 3D reconstruction is still an open challenge. Taking up these challenges is important because many objects of interest are deformable, including faces, bodies, organs, clothes, and fabrics. Furthermore, 2D cameras are by far the most common types of imaging sensors in use today, yielding a broad range of useful applications for passive methods, as discussed in the next section.

M. Gallardo (✉)

EnCoV, Institut Pascal, UMR 6602, CNRS/UBP/SIGMA, 63000 Clermont-Ferrand, France
e-mail: Mathias.Gallardo@gmail.com

T. Collins

IRCAD and IHU-Strasbourg, 1 Place de l'Hôpital, 67000 Strasbourg, France
e-mail: toby.collins@ircad.fr

A. Bartoli

EnCoV, Institut Pascal, UMR 6602, CNRS/UBP/SIGMA, 63000 Clermont-Ferrand, France
e-mail: Adrien.Bartoli@gmail.com

© Springer Nature Switzerland AG 2020

J.-D. Durou et al. (eds.), *Advances in Photometric 3D-Reconstruction*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-51866-0_4

Along with passive methods which only use 2D images, active methods with depth sensors have also tackled this problem. This has been done for instance using stereo-cameras [79], infrared projectors or time-of-light systems with Kinect [40, 55] and more elaborated systems such as color photometric stereo techniques [12, 31]. Despite their impressive results, active depth sensors suffer from inherent limitations: some have a restricted range (they cannot sense depth when the object is too far from or too close to the sensors), others have a significantly higher power consumption than RGB cameras, and some others are often strongly affected by outdoor illumination conditions. There may also be physical restrictions, such as in endoscopic applications, where it is not possible to use bulky active vision sensors. Finally, there are billions of monocular cameras used every day on mobile devices, which yields a huge potential for usage and commercialization and underlines the need for solving the problem of monocular deformable 3D reconstruction.

Four main paradigms have emerged to tackle deformable reconstruction with monocular cameras: Shape-from-Template (SfT), Non-Rigid Structure-from-Motion (NRSfM), Shape-from-Shading (SfS) and neural network-based reconstruction (NNR). We now briefly summarize them. SfT uses a known *template* of the object and at least one image of the object being deformed. It works by registering the object to the input image and deforming the template of the object accordingly in 3D [6, 71]. NRSfM uses multiple monocular images and recovers the 3D shape of the deforming object in each image [11]. This paradigm is much harder to solve than SfT because no template is available and consequently, the physical structure of the object is unknown a priori. Figure 4.2 illustrates both paradigms. SfS only uses a single image and recovers the depth or surface normal at each pixel. SfS works exclusively with shading information which links surface geometry, surface reflectance, scene illumination, pixel intensity, and camera response function [37]. SfS is often very difficult to use in practice, mainly because it is generally ill-posed, requires a complete *photometric calibration* of the scene *a priori* and suffers discrete convex/concave ambiguities. NNR approaches predict the 3D shape of an object or the depth-map of a scene from a single image using a trained neural network. Most of these methods pose the problem as a supervised learning task. It works well for common object classes with very large datasets available, such as man-made objects [65] and faces [67] using specific low-dimensional deformation models. This category has not shown to enable 3D reconstruction of objects under very high dimensional deformations, which notably limits its applicability. Two other shortcomings are the need for large amounts of annotated training data, comprising images and known 3D deformation pairs, which are hard to obtain with real data, and the need for a training phase which may not be practical in several real applications, when the template is acquired at run-time. For these reasons, the most practical paradigms are currently SfT and NRSfM. Nevertheless, neural networks can be used in conjunction with SfT and NRSfM to provide state-of-the-art solutions to intermediate problems, such as motion estimation and feature extraction.

Most SfT and NRSfM methods use the apparent motion of the object's surface, also called motion cue. That is, by knowing the relative movement between the surface template and the image, or between images, they infer the 3D deformation.

However, the motion cue is often insufficient to infer the 3D shape of a deforming object, because motion can be explained by possibly infinitely many 3D deformations (the so-called depth ambiguity). To fix this, SfT and NRSfM methods, as with other deformable 3D reconstruction approaches, use deformation priors, which we discuss in detail in Sect. 4.3.1.1. Despite the inclusion of deformation priors, SfT and NRSfM methods generally fail in two cases: when the object is poorly-textured or when it deforms non-smoothly. Figures 4.6 and 4.7 illustrate this with some reconstructions from state-of-the-art NRSfM methods [14, 60]. At poorly-textured regions, motion information is sparse and accurate reconstruction becomes difficult. In the last years, to overcome these limitations, some methods proposed to complement the motion cue with the shading cue. Unlike motion, shading can be used to reconstruct textureless surfaces, as it is considered the most important visual cue for inferring shape details at textureless regions [61].

Contributions

We propose to combine shading with NRSfM in order to reconstruct densely-textured and poorly-textured surfaces under 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. However, we must know albedos in order to apply shading constraints. Therefore, our problem is to simultaneously and densely estimate non-rigid 3D deformation from each image together with spatially-varying surface albedo. We assume deformation is either piecewise or globally smooth, which allows us to reconstruct creasing or wrinkled surfaces. We assume that the albedo is piecewise smooth, which is very common for man-made objects. Furthermore, this assumption on albedo reduces the potential ambiguity between smooth albedo changes and smooth surface orientation changes. This problem has not been tackled before. It is a crucial missing component for densely reconstructing images in unconstrained settings and may then enlarge the spectrum of deformations and surfaces for more real-world applications.

4.2 Applications of Monocular Deformable 3D Reconstruction

Research on monocular deformable 3D reconstruction has raised considerable interest in its applications in many domains including medical image processing, special effects, data-driven simulation, Augmented Reality (AR) games, and soft body mechanics. Some examples are shown in Fig. 4.1.

There are many important applications in medical AR with Minimally Invasive Surgery (MIS) and more precisely with laparoscopic surgery. This advanced surgery technique is performed by inserting, through small incisions, small surgical instruments and a laparoscope, which is a thin, tube-like instrument with a light and a digital camera. The surgeon uses the live video from the camera to perform the surgery. This

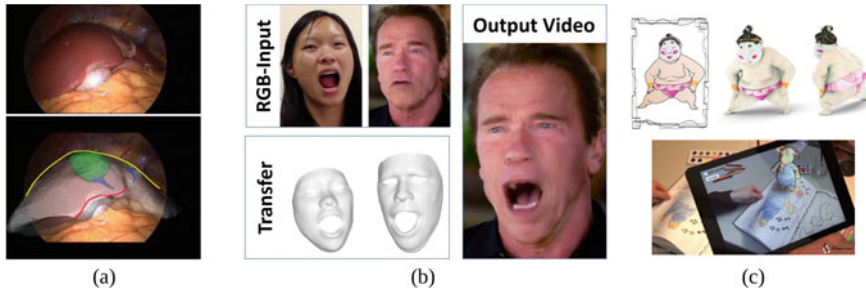


Fig. 4.1 Applications of monocular deformable 3D reconstruction: **a** medical imaging [44], **b** post-production movie editing [77] and **c** AR gaming [50]

reduces the patient’s trauma and shortens the recovery time, compared to open surgeries. However, during MIS, surgeons face three main problems: the viewpoint is limited, the localization in 3D and the perception of depth become harder, and they cannot see the locations of important subsurface structures such as a tumor or major vessels. AR appears to be a very suitable way to give a real-time feedback during MIS. This is done by augmenting the live video with a deformable model of the organ, including its surface and internal structures. The deformable model can be constructed from a preoperative medical image, such as MRI or CT, and the task of registering the organ model to the laparoscopic video is an SfT problem. Using a monocular laparoscope, a deformable registration of a preoperative template of a liver (obtained from CT) was presented in [44]. This permits one to register at the same time the tumor (in green) and the internal structures of the liver such as veins (in blue).

Another application area is video post-production. Video editors are often required to modify videos after the recording, by removing, introducing or modifying content. When the content is deformable, this can be highly labor intensive. Most videos are not recorded with depth sensors, which makes monocular methods extremely valuable. A real-time technique of facial performance capture and editing on 2D videos was proposed in [77]. It works by reconstructing the 3D faces of a source actor and of a target actor, and transfer the facial expression of the source actor to the target actor.

Another large application domain is AR gaming. The idea is to offer players new gameplay experiences and a different game environment that combine virtual content with the real-world environment. Nearly all AR games assume the scene to be rigid. Recently new games have been presented, enabled by SfT. For instance, an AR coloring book application is presented in [50]. The idea is for a player to interactively color a virtual 3D character by coloring a 2D sketch of the character printed in a book, using color pencils. An SfT algorithm is used to register a template of each book page and estimate the deformation of the visible page. This allows registration of the colored page with the virtual character, which then allows the transfer of pencil colors to the virtual character and visualization in real-time.

4.3 Background on Deformable Monocular 3D Reconstruction and Shading

Since the first works on deformable monocular 3D reconstruction [11, 32], many technical and theoretical aspects have been explored in both NRSfM and SfT. The main ones are (i) shape and deformation modeling, (ii) data constraints extracted from the input images, (iii) 3D shape inference and (iv) the use of temporal coherence. As our main contribution relates to NRSfM, we thoroughly review NRSfM and each of the above directions. We then propose an overview of 3D reconstruction using shading and especially focus on some works which integrate shading with SfT. This is motivated by the fact that SfT and NRSfM are closely related.

4.3.1 Non-Rigid Structure-from-Motion

The goal of NRSfM is to recover a deformable object's shape from a set of unorganized images or a video, as depicted in Fig. 4.2.

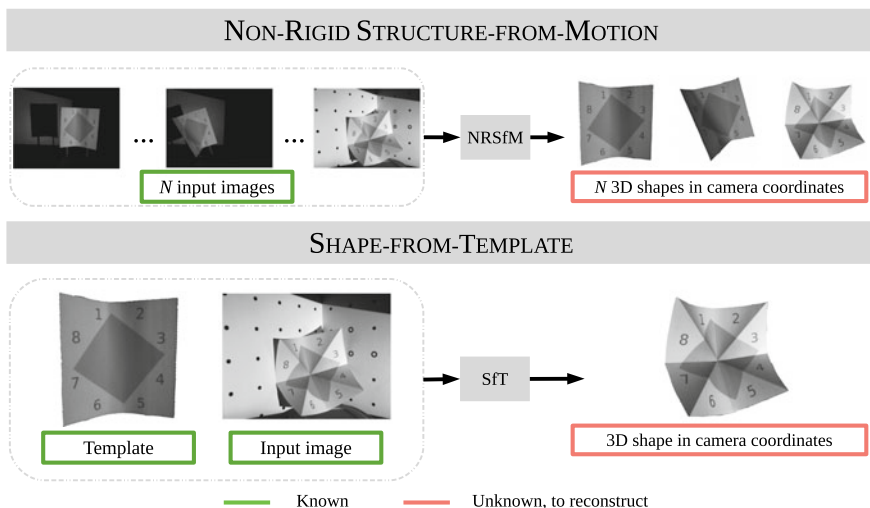


Fig. 4.2 Illustration of the problems of NRSfM and SfT

4.3.1.1 Deformation Priors in NRSfM

Deformation priors are required to make NSRfM well-posed. Three classes of deformation priors have emerged: *statistical* priors, *physics-based* priors and *temporal smoothness* priors.

Statistical priors have been formulated in two main ways: low-rank shape bases and low-rank trajectory bases. Both ways use a reduced space of modes to model the shapes, which are learned during the reconstruction, i.e., which is thus the joint estimation of the model modes, weights, and the camera poses. These modes are usually constrained to lie in a linear space spanned by a small number of an unknown 3D shape bases [11], or of unknown 3D trajectory bases [4, 18]. Both approaches reduce the problem dimensionality, however, they present three limitations. They tend to require a large number of images and short-baseline data to achieve good results, and they lose the ability to model high-frequency deformations, i.e., discontinuities, such as creases.

Physics-based deformation priors 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* priors is isometry or quasi-isometry [14, 15, 60, 81, 83, 85]. It enforces the geodesic distance between two points on the surface to be preserved by deformation. When imposed exactly, no stretching or shrinking is permitted. When imposed inexactly, it is called quasi-isometry and penalizes solutions with increased stretching or shrinking using a penalty function. Isometry and quasi-isometry 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 such as the ones shown in Figs. 4.6 and 4.7.

It appears that NRSfM with the isometric prior can be solved up to discrete, spatially localized two-fold ambiguities if motion can be estimated densely across the object's surface [14, 60, 81]. The main difficulty with isometry is that it is a non-convex constraint. The *inextensibility* prior relaxes isometry in order to form a convex constraint. It prevents the Euclidean distance between two neighboring surface points from exceeding their geodesic distance. However, it is too weak to reconstruct geometry accurately and must be combined with additional constraints. This has been done using the so-called Maximum Depth Heuristic (MDH), where a depth maximization constraint is imposed to prevent the reconstructed surface from shrinking arbitrarily. The inextensibility prior has been first proposed for SfT [13, 63, 70], but it has been adapted for NRSfM [16]. The MDH has been shown to produce very good reconstructions when the perspective effects of the camera are strong.

Temporal smoothness priors assume that the object deforms smoothly over time. These priors have been mainly used through two approaches: (i) using temporal smoothing constraint [4, 29] and (ii) initializing the shape of an input image using the one of the previous input image [85]. One important advantage of the approach (ii) is that the problem is more constrained. This may then provide more accurate reconstructions since it optimizes an initial solution. However, this can turn as a shortcoming. The solution can be stuck in a local minimum if the solution from

the previous frame is wrong, because of tracking loss which may happen in case of sudden illumination changes or occlusions. Temporal smoothness can also be used to assist correspondence estimation generally using optical flow approaches to obtain dense correspondences [29, 85].

4.3.1.2 Data Constraints in NRSfM

NRSfM methods rely fundamentally on motion constraints, and these can be divided into two types: *correspondence constraints*, which assume the correspondences are computed a priori [14, 15, 29, 30, 60, 81, 83], and *direct constraints* those which compute correspondence jointly to deformation inference using brightness constancy [85]. By far, the most common constraints are motion constraints, however, *contour constraints* have been also used in NRSfM [85].

Correspondence constraints force 3D points on the surface to project at their corresponding 2D points in each input image. The points used by these constraints are usually obtained by matching features [49], or tracking points [29, 74], in the images. Correspondence constraints have four main limitations. First, feature-based matching methods may fail to establish correspondences without errors. Second, the computational time to extract features, compute descriptors, and perform the matching can be long without high performance GPUs. Third, tracking-based methods require short-baseline input images. Fourth, they work well only for densely-textured objects with discriminative texture, which are not common in most real practical applications, particularly with man-made objects and many natural objects.

Direct constraints work by maximizing the photometric agreement, i.e., brightness constancy, between the input images [85]. Their main advantage is to provide denser motion constraints than correspondence constraints. They have, however, three main limitations. First, they are highly non-convex and they require iterative optimization. Because of non-convexity, they are usually applied in a frame-to-frame tracking setup. Second, direct constraints are sensitive to strong photometric changes which may be induced by complex deformations or complex illuminations. Third, direct constraints may require reasoning about surface visibility (they should be deactivated at surface regions that are occluded).

Contour constraints force the object's occluding contours to align with the corresponding contours in the images [85]. Contour constraints are interesting because they do not depend on the surface's texture, and therefore, are applicable for poorly-textured and even non-textured surfaces. There exist two types of contour constraints: silhouette contour constraints and boundary contour constraints. Silhouette contour constraints work by forcing the object's silhouette to align with silhouette contours detected in the input image and can be used for surfaces and volumes. These constraints have not been used in NRSfM yet, mainly because they are very difficult to use without any prior on the object's shape. Boundary contour constraints are applicable for open surface templates such as a piece of paper. They work by enforcing the boundary contour projects to image edges [85]. Similarly, to direct constraints, contour constraints are highly non-convex, usually enforced iteratively, and require

a good initial estimate. However, they are also difficult to apply robustly, particularly with strong background clutter.

4.3.1.3 Local and Global Methods to NRSfM

Another important way to characterize NRSfM methods is whether they reconstruct a surface using local surface regions (usually called local methods), or whether they reconstruct the whole surface at once (usually called global methods). *Local methods* work by dividing the surface into local regions, reconstructing each region individually, and then reconstructing the full surface using surface continuity. Most local methods assume isometric deformations. They mainly differ by the way they locally model the surface: piecewise planes [76, 81], quadrics [23], or Partial Differential Equations (PDEs) [14, 60]. Their advantages are that they can be fast and can provide closed-form solutions. However, they also produce sub-optimal results, because they do not enforce the physical prior globally over the whole surface, and they may be unstable and present ambiguities. *Global methods* use instead constraints acting over the whole surface. These methods produce large, non-convex optimization problems that cannot lead to closed-form solutions. They generally use energy minimization frameworks for optimization. This allows them to handle more complex deformations and to use more complex constraints, leading to potentially more accurate reconstructions. However, they generally require high computation time and a good initial solution, and they are often difficult to optimize and not easily parallelizable.

4.3.1.4 Unorganized Image Sets Versus Video Inputs

A final way to categorize NRSfM methods is if they operate with unorganized image sets [14, 60, 81] or videos [4, 16, 23, 29, 76, 83] as inputs. A fundamental difference between these two settings is that temporal continuity can be exploited in the latter setting but not in the former setting. Typically for video inputs, methods work in an incremental style where new frames are added to the optimization process while fixing unknowns in past frames [4, 23, 29]. This strategy is used to manage the growing number of unknowns with video inputs, allowing it to scale well for long videos.

4.3.2 3D Reconstruction Using Shading

Shading relies on the photometric relationship between surface geometry, surface reflectance, illumination, the camera response, and pixel intensity. This relationship, also called the shading equation, provides one constraint on the surface normal at any given pixel. Shading is a powerful visual cue because, unlike motion, it can constrain

3D shape at poorly-textured surface regions and recover complex deformations in such surfaces [62]. Shading has been first used alone in the paradigm of SfS and then in other 3D reconstruction problems and in SfT.

4.3.2.1 Shape-from-Shading

SfS consists in using shading to recover the 3D shape of an object from a single image. Precisely, it recovers the surface normal at each pixel of the image. SfS has been intensively studied in the last decades and the SfS literature can be explored through four main components: (a) the camera projection model, (b) the illumination model, (c) the surface reflectance model, and (d) the 3D shape inference algorithm. For (a), SfS has been first studied with the orthographic camera [37, 61] and then the perspective camera model [64, 75]. For (b), most of the existing methods assume a distant light source, but more complex illumination models are also used, such as the near-point lighting with fall-off [58, 64]. Most SfS methods also assume known illumination. For (c), a very common assumption of reflectance is the Lambertian model [21, 37, 39, 43, 47, 58, 61, 64, 68, 69, 78, 88]. The Lambertian model assumes a diffuse reflection of the surface, i.e., the surface luminance is independent from the viewing angle. Non-Lambertian reflectance models are also studied [3, 46], such as the Oren-Nayar and Ward models which, respectively, take into account the micro-facets reflections and specular reflections. Nearly all methods assume either constant and fixed albedo or known albedo. This is because SfS is fundamentally an ill-posed problem with unknown varying albedo. Some works, however, propose solutions to handle multi-albedo surfaces. To handle multiple albedos, [5] forms a complex energy function which simultaneously solves several problems related to the photometric formation of the image, namely SfS, intrinsic images decomposition, color constancy and illumination estimation. For (d), SfS methods can be divided in six subcategories: (i) propagation approaches [64], (ii) local approaches [61], (iii) linear approaches [78], (iv) convex minimization approaches [21], (v) non-convex minimization approaches [5], and (vi) learning-based approaches [68].

Despite the great interest drawn by SfS and the diversity of the proposed approaches, almost all of the existing SfS methods share the same shortcomings. First, they assume the albedo values and the scene illumination to be known, i.e., they require a complete photometric calibration, as the survey [20] shows. Second, they suffer from convex/concave ambiguities [9, 38]. Third, they cannot handle depth discontinuities and provide a surface solution up to a global scale factor.

4.3.2.2 Extending SfS to Multiple Images

Shading has been used previously in several other 3D reconstruction problems. These include photometric stereo [12, 31, 91], multi-view SfS [41, 86], multi-view reconstruction [8, 42, 45, 79, 87], SfM and SfS [42]. Their main limitations are that they work for rigid objects or/and use impractical setups.

Photometric stereo is the extension of SfS using multiple light sources. The images taken under different illumination contain no motion. This is one big difference between photometric stereo and the other extensions of SfS. It has shown great success for reconstructing high-accuracy surface details with unknown albedo such as [12, 31, 91]. However, it requires a special hardware setup where the scene is illuminated by a sequence of lights placed at different points in the scene. This setup is not applicable in many situations. Another limitation is that the scene is assumed to be rigid during the acquisition [12, 91]. [31] proposes, however, a photometric stereo technique which works for deforming surfaces.

Multi-image SfS methods, such as [41, 86], have shown that using shading and a collection of images, from monocular [41] or several tracked cameras [86], provide reasonably good reconstructions of poorly-textured surfaces such as statues or bones. Multi-view reconstruction methods, such as [8, 45, 79, 87], have shown that shading reveals fine details for e.g., clothes or faces. However, these methods assume rigid objects, use two or more cameras and require a special design of the scene, which may not be practical.

Shading has also been used in rigid SfM [42], which uses multiple images showing a rigid object. This approach initializes the surface using motion through SfM and MVS and then refines it by combining motion with shading information. Unlike the other extensions of SfS, this approach requires to solve a registration problem, to link pixel information across different images. One limitation may come from the difficulty of establishing correspondences accurately. However, because of the MVS constraints, this approach may achieve higher accuracy than photometric stereo at both textured and textureless regions.

4.3.2.3 Existing Methods to Solve SfT with Shading

We briefly present the principle of SfT regarding the directions *(i)* shape and deformation model, *(ii)* image data constraint, and *(iii)* 3D shape inference. We then describe how shading has been used in SfT. We refer the readers to [25], for more details on SfT.

Shape-from-Template

The goal of SfT is to register and reconstruct the 3D shape of a deforming object from a single input image, using a template of the object. This is illustrated in Fig. 4.2. The template is a textured 3D model of the surface in a rest position and the problem is solved by determining the 3D deformation that transforms the template into camera coordinates. The main difference between SfT and NRSfM is that in NRSfM the object's template is not provided a priori, and this makes it a considerably harder problem.

The template brings strong object-specific prior knowledge to the problem. It comprises a *shape model*, an *appearance model* and a *deformation model*. The *shape model* represents the object's 3D shape in a fixed reference position. The *appearance model* is used to describe the photometric appearance of the object. The *deformation*

model is used to define the transformation of the template’s reference shape and the space of possible deformations. For this, most methods use dimensionality reduction through *smooth parameterizations*, *explicit smoothing* priors or *physics-based* priors. *Smooth parameterizations* have included thin-plate splines and B-splines, and reduce dimensionality by modeling deformation with a discrete set of control points [56, 71, 73]. Smooth parameterizations reduce in general the cost of optimization, however, they lose the ability to model high-frequency deformations, i.e., discontinuities, such as creases. *Explicit smoothing* priors penalize non-smooth deformations explicitly. They use a smoothing term within an energy-based optimization, usually using the ℓ_2 norm [7, 13]. This norm strongly penalizes non-smooth deformations and provides strong problem regularization, but can prevent the formation of discontinuities such as creases. *Physics-based* priors in SfT work in a very similar manner than in NRSfM. The most commonly used is the *isometry* prior [6, 15, 48, 71], however, other priors have been studied: inextensibility [13, 63, 70], conformal (angle preservation) [6, 51] and elasticity [35, 53, 59].

Data constraints must be extracted from the input image in order to match the template’s shape with the object’s true shape. Similar to NRSfM, the most common data constraints are motion constraints [6, 13, 17, 52, 56, 57, 71, 89], but some methods also rely on contour [72, 84] and shading constraints.

Combining Motion and Shading in SfT

Shading has been also used as a complementary visual cue in SfT [27, 48, 51, 54, 82]. These methods differ in the way the problem is modeled and optimized. [54, 82] use motion and shading information sequentially, and not in an integrated manner. We refer to these as *non-integrated* approaches. The proposed approaches are difficult to use in practice because of several significant drawbacks. [82] requires a full photometric calibration a priori and the illumination to be the same at training and test time. [54] requires to know the reflectance of the template and only works for smooth deformations.

By contrast, in [27, 48, 51], shading, motion, and deformation priors are integrated together into a single non-convex energy function which is minimized through iterative refinement. We refer to these as *integrated* approaches. Their advantage is that they combine constraints from multiple cues simultaneously, to improve reconstruction accuracy, which is not possible with *non-integrated* approaches. [48, 51] simplify the problem by assuming a rigid observation video is available prior to deformable reconstruction. This is a video of the object taken from different view-points before any deformation occurs and is used for reconstructing the template. Reconstruction then proceeds with an SfT-based approach. The main limitation of this is that it requires control of the environment to ensure the object does not deform during the observation video. This is not often possible in real applications. Furthermore, such methods assume the scene illumination is constant and fixed during the rigid observation video, which is not always possible to achieve.

Current Limitations

As Figs. 4.6 and 4.7 illustrate, poorly-textured surfaces present an important limitations of nearly all state-of-the-art NRSfM and SfT methods, particularly with creases.

The reason is that motion information, which is the most used data constraint, is fundamentally insufficient to reconstruct textureless surface regions undergoing non-smooth deformations. As mentioned in Sect. 4.3.2.1, shading works on textureless regions and can be used also to infer fine surface details. The *integrated* methods in SfT [27, 48, 51] show that it is possible to combine motion (from textured regions) and shading (from poorly-textured regions) constraints with the physical constraints from the template in order to reconstruct densely at textured and poorly-textured regions. However, since NRSfM does not assume the object’s template to be given, combining shading and NRSfM appears to be a much more difficult problem.

4.4 Proposed Solution to NRSfMS

We now focus on the problem of Non-Rigid Structure-from-Motion and Shading (NRSfMS) and we present the first integrated solution. We show in detail how combining motion and shading allows reconstructing creasable, poorly, and well-textured surfaces. The challenge we face is to simultaneously and densely estimate non-smooth, nonrigid shape from each image together with unknown and spatially-varying surface albedo. We solve this with a cascaded initialization and a non-convex refinement that combines a physical, discontinuity-preserving deformation prior with motion, shading, and boundary contour information. Our approach works on both unorganized and organized small-sized image sets, and has been empirically validated on six real-world datasets for which all state-of-the-art approaches fail.

In Sect. 4.4.1, we present our modeling of the problem and our motion and shading-based cost function. In Sect. 4.4.2, we present our optimization framework and in Sect. 4.4.3, we study the basin of convergence of our method and validate it with high-accuracy ground-truth datasets. In Sect. 4.4.4, we provide our conclusions and some research axes of future work.

4.4.1 Problem Modeling

4.4.1.1 Overview

There are many possible ways to define an NRSfMS problem, and many potential choices that must be made regarding scene assumptions, models, unknown, and known terms, *etc.* We present a rigorous definition of NRSfMS through eight fundamental components. To define an NRSfMS problem, we must define or *instantiate* each component. In the following, we describe each component and an instantiation justified by practical considerations for real-world application.

(a) *Models (shape, reflectance, illumination, camera response, and camera projection)*. We use a high-resolution thin-shell 3D mesh to model the object’s 3D shape, and a barycentric interpolation to describe deformations across the surface.

This allows us to model complex deformations using high-resolution meshes. Deformation is modeled quasi-isometrically and creases are modeled with a novel implicit energy term as described in Sect. 4.4.1.4. Surface reflectance is modeled using the Lambertian model with piecewise-constant albedo, which has been also used by [5]. This gives a good approximation of many man-made objects such as clothes, fabrics, and cardboards. Scene illumination is assumed to be constant over time and fixed in camera coordinates. In practice, this can be assumed if we have a camera-light rig setup such as an endoscope or camera with flash, or a non-rig where the light and a camera are not physically connected but do not move relative to each other during image acquisition. We use the first-order spherical harmonic model, however, the second-order model can be also used in our proposed solution to increase modeling accuracy. Spherical harmonics are very commonly used in SfS [5, 66] and photometric stereo [34]. We assume the perspective camera model [36], which handles well most real-world cameras, and we assume the camera intrinsics are known a priori using a standard calibration process with, e.g., OpenCV. The camera response model maps the *irradiance image*, i.e., the image which stores the light striking the image plane at each pixel, to the *intensity image*, which is the grayscale image outputted by the camera, before that the camera nonlinearities, such as gamma mapping, vignetting, and digitization, are introduced. For the camera response, we assume that it is linear, which is a valid assumption for many CCD cameras. We assume that it can change over time in order to handle changes due to camera shutter speed and/or exposure. (b) *Known and unknown model parameters*. Most of the above mentioned models have parameters that must be set. The NRSfMS problem changes dramatically according to which model parameters are known *a priori*. We consider the unknowns are as follows: The surface albedo, which, due to the assumption of piecewise-constant albedo, corresponds to an albedo-map segmentation and segment values. The mesh vertex positions in camera coordinates, which provide the 3D shape of the surface in each image. The illumination, the camera responses, and the camera intrinsics are assumed known. These assumptions are reasonable for two reasons. First, the illumination and the camera can be calibrated a priori using standard techniques and the camera responses can be obtained from the camera or computed using, e.g., the background. Second, it is unrealistic to know a priori the reflectance model of a surface as the object is a priori unknown, contrary to SfT. Secondly, as this is the first solution to NRSfMS, our goal is to show that it can be solved in simplified conditions, and in the future, we can investigate releasing the assumption of known illumination, camera response, and intrinsics. (c) *Visual cues*. The visual cues determine which visual information is used to constrain the problem. We use motion, boundary contour, and shading. Motion is used to constrain textured regions of the surface and boundary contours to constrain the perimeter of the surface. We use shading constraint to densely reconstruct surfaces and reveal creases in poorly-textured regions. (d) *Number of required images*. We require at least 5 images. We discuss the implications of using smaller numbers of images in the conclusion Sect. 4.4.3.5. (e) *Expected types of deformations*. We assume quasi-isometric and piecewise-smooth deformations and no tearing. Tearing implies the surface mesh topology must adapt during reconstruction and this adds considerable complexity to the problem. Here it

is sufficient to show that non-torn surfaces can be reconstructed. (f) *Scene geometry*. We assume the surface to be reconstructed has no self or external occlusions, but there can be background clutter. These are typical assumptions in NRSfM state-of-the-art, and the assumption of no occlusions is used to simplify data association (i.e., knowing which regions of the images correspond to which regions of the surface). We also assume there is a *reference image* within the image set. The reference image is one of the input images that we use to construct the surface’s mesh model. We can use any image as the reference image, however, in practice, we obtain better reconstructions using a reference image where the surface is smooth. (g) *Requirement for putative correspondences*. Putative correspondences are points in the reference image whose positions are known in the other images. We assume to know a priori a set of putative 2D correspondences computed using standard methods such as SIFT. We assume there may be a small proportion of mismatches, e.g., <20%, which is the case in real applications. (h) *Surface texture characteristics*. We assume the surface presents a combination of both well and poorly-textured regions.

4.4.1.2 Shape, Deformation, Reflectance, Illumination, and Camera Modeling

We define Ω as the segmented region of the object of interest in the reference image. We build the *shape model* by meshing Ω using a regular 2D triangular mesh, with M vertices and M on the order of 10^4 . We denote the mesh’s edges as E , where N_E is the number of edges. 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 . Without loss of generality we assume the reference image is the first image. We then 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 image, defined in normalized pixel coordinates. Its corresponding position in 3D camera coordinates at $t = 1$ is $\mathbf{v}_1^i = d_i[\mathbf{u}_i^\top, 1]^\top$, where d_i is its unknown depth. We collect these unknown depths into the set $\mathcal{D} = \{d_1, \dots, d_M\}$. The full set of unknowns that specify the object’s shape in all images is, therefore, $\{\mathcal{D}, \mathcal{V}_2, \dots, \mathcal{V}_N\}$, which corresponds to $3M(N - 1) + M$ real-valued unknowns.

The *deformation model* transforms each vertex to 3D camera coordinates: we model the position of each vertex $i \in \{1, \dots, M\}$ in camera coordinates by $\mathbf{v}_t^i \in \mathbb{R}^3$, where t denotes time. We transform a point $\mathbf{u} \in \Omega$ to camera coordinates according to \mathcal{V}_t with a barycentric interpolation, which is a linear interpolation of the positions of the three vertices surrounding \mathbf{u} . This barycentric interpolation, therefore, defines a piecewise-linear embedding function from Ω to 3D, parameterized by the vertex positions \mathcal{V}_t . We denote φ this barycentric interpolation and $n(\mathbf{u}; \mathcal{V}_t) : \mathbb{R}^{3 \times M} \rightarrow \mathbb{S}_3$ its unit surface normal.

For the *surface reflectance model*, we define an *albedo-map* $A(\mathbf{u}) : \Omega \rightarrow \mathbb{R}^+$ 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 $A(\mathbf{u}) : \Omega \rightarrow \mathcal{A}$ where $\mathcal{A} = \{\alpha_1, \dots, \alpha_K\}$

denotes a discrete set of K unknown albedos with $\alpha_k \in \mathbb{R}^+$. We discuss how A is built in Sect. 4.4.2.

The *illumination model* gives the power and spatial distribution of light. We denote the unknown illumination coefficients by \mathbf{I} . This shading equation predicts the intensity of a pixel given the models of illumination, surface shape, surface reflectance, camera projection, and camera response. This starts with the *surface irradiance* which is the amount of light received by the surface. We use the function r to denote the surface irradiance for a normal vector \mathbf{n} according to \mathbf{I} . Then, the amount of light reflected by the surface and striking the camera forms the irradiance image. This image contains the photometric variations caused by shading in particular. At any time t , we denote the irradiance image by R_t and the intensity image by L_t . We denote the *camera response function* by $g_t : \mathbb{R} \rightarrow \mathbb{R}$ which transforms the irradiance image R_t into the intensity image L_t .

As we use the first-order spherical harmonic model, the illumination model is a combination of a light source at infinity and an ambient term. Note that, as \mathbf{I} is represented by spherical harmonics, the surface irradiance r is linear in \mathbf{I} . As we assume the Lambertian reflectance model, we have $r(\mathbf{n}, \mathbf{I}) = (\mathbf{n}^\top, \mathbf{1}) \mathbf{I}$. As we assume g_t is linear, we have $L_t = \beta_t R_t$ with $\beta_t \in \mathbb{R}^+$.

4.4.1.3 Inputs and Outputs

Our inputs are as follows. (i) a set of N input RGB images $\{I_t\}_{t \in [1, N]}$, $I_t : \mathbb{R}^2 \rightarrow [0, 255]^3$ with a deforming object and the corresponding intensity images $\{L_t\}_{t \in [1, N]}$, $L_t : \mathbb{R}^2 \rightarrow \mathbb{R}^+$. In practice, the intensity image L_t is obtained by calibrating radiometrically the camera or by selecting the second component of the projection of the input RGB image I_t in the CIE XYZ color space, which is done for our experiments. (ii) the camera intrinsics of all perspective projection functions Π_t . (iii) a segmentation of the object of interest in the reference image, denoted by the region $\Omega \subset \mathbb{R}^2$. (iv) the scene illumination coefficients $\mathbf{I} \in \mathbb{R}^4$. (v) the camera response functions g_t . (vi) N sets \mathcal{S}_t of matched putative 2D correspondences from Ω to each input image I_t . We denote it by $\mathcal{S}_t = \{(\mathbf{u}_j, \mathbf{p}_t^j)\}$ where \mathbf{u}_j denotes the j th 2D point in Ω and \mathbf{p}_t^j denotes its corresponding position in the t th input image I_t . The number of correspondences for each image t is denoted by s_t . Details for how this is done for our experimental datasets are given in Sect. 4.4.3.1.

The outputs of our solution to NRSfMS are: (i) the vertices \mathcal{V}_t of the shape model in the camera coordinates for all input images and (ii) the segmented albedo-map A with its K segments and values $\{\alpha_1, \dots, \alpha_K\}$.

4.4.1.4 Problem Modeling with an Integrated Cost Function

The cost function combines *physical deformation priors* (quasi-isometry and smoothing constraints) with shading, motion and boundary constraints extracted from all images. The objective function C_{total} has the following form:

$$\begin{aligned}
C_{total}(\mathcal{V}_1, \dots, \mathcal{V}_N, \alpha_1, \dots, \alpha_K) \triangleq & \sum_{t=1}^N \left(C_{shade}(\mathcal{V}_t, \alpha_1, \dots, \alpha_K) + \right. \\
& \lambda_{motion} C_{motion}(\mathcal{V}_t) + \lambda_{contour} C_{contour}(\mathcal{V}_t) + \\
& \left. \lambda_{iso} C_{iso}(\mathcal{V}_1, \mathcal{V}_t) + \lambda_{smooth} C_{smooth}(\mathcal{V}_t) \right). \quad (4.1)
\end{aligned}$$

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

The shading constraint. This robustly encodes the Lambertian relationship between surface, albedo, surface irradiance, pixel intensity, and camera response. We use the piecewise-constant albedo model given earlier, and we decide to not optimize all albedo segments. There are two reasons. First, there is a potential difficulty with using shading at textured regions. This comes from the fact that the mis-registration errors at textured regions may imply mis-registration of the albedo-map over the surface. This then may lead to large errors in albedo estimation and surface reconstruction because of the linear dependency of the shading constraint in albedo values. Second, textured regions are very informative for motion constraints. The shading constraint is then less useful or even not useful at textured regions. Therefore, we propose to not use shading in textured regions. For this, we use the fact that textured regions can be detected as small albedo segments. We propose to exclude from the optimization albedo segments which are smaller in area than the threshold T_A (in % of the number of pixels contained in the image). In practice, we found that using $T_A = 0.022\%$ allows to reduce reconstruction errors at textured regions. We give details about how this is integrated to our proposed algorithm in stage 3 in Sect. 4.4.2. We remind that φ is the piecewise-linear embedding function from Ω to 3D, parameterized by the vertex positions, and we form every constraint using this function. We evaluate the shading constraint at each pixel of albedo segments larger than T_A , which gives

$$C_{shade}(\mathcal{V}_t, \alpha_1, \dots, \alpha_K) \triangleq \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \rho_0 \left(A(\mathbf{u}) r(n(\mathbf{u}; \mathcal{V}_t); \mathbf{l}) - L_t \Pi_t(\varphi(\mathbf{u}; \mathcal{V}_t)) \right), \quad (4.2)$$

$$\text{with } \rho_0(x) = \begin{cases} \frac{x^2}{2}, & \text{if } |x| \leq k \\ k \left(|x| - \frac{k}{2} \right), & \text{if } |x| \geq k, \end{cases} \quad (4.3)$$

which is the Huber M-estimator. For the experiments, we found that the Huber hyperparameter set to $k = 0.005$ gives the best results. The function ρ is used to enforce similarity between the modeled and measured intensity, while also allowing for some points to violate the model (caused by specular reflection, small shadows, and other unmodeled factors). When the residual of such points is not too high, we find that a robust estimator based on an M-estimator is very effective to handle them. In order to reduce computation time, pixels from Ω are downsampled by a factor

of X , by taking one pixel every X pixels from Ω . In practice, we found that using $X = 2$ gives good reconstructions.

The motion constraint. We recall that the set \mathcal{S}_t holds s_t putative correspondences between Ω and image $t \in [1, N]$. The constraint robustly enforces each point \mathbf{u}_j to transform to its corresponding point \mathbf{p}_t^j , and is given by

$$C_{motion}(\mathcal{V}_t) \triangleq \sum_{(\mathbf{u}_j, \mathbf{p}_t^j) \in \mathcal{S}_t} \rho_1 \left(\left\| \Pi_t \circ \varphi(\mathbf{u}_j; \mathcal{V}_t) - \mathbf{p}_t^j \right\|_2 \right), \quad (4.4)$$

where ρ_1 is the parameter-free (ℓ_1 - ℓ_2) M-estimator

$$\rho_1(x) = 2 \left(\sqrt{1 + \frac{x^2}{2}} - 1 \right). \quad (4.5)$$

This constraint encourages the function φ to project each point \mathbf{u}_j onto the input image at the correspondence position \mathbf{p}_t^j .

The boundary contour constraint. We discretize the boundary of Ω to obtain a set of boundary pixels $\mathcal{B} \triangleq \{\mathbf{u}_{k \in [1, N_{\mathcal{B}]}}\}$, with $N_{\mathcal{B}}$ the number of boundary pixels. We then compute a boundariness-map for each image $B_t : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ where high values of $B_t(\mathbf{p})$ correspond to a high likelihood of pixel \mathbf{p} being on the boundary contour. The constraint is evaluated as

$$C_{contour}(\mathcal{V}_t) \triangleq \frac{1}{N_{\mathcal{B}}} \sum_{\mathbf{u}_k \in \mathcal{B}} \rho_1 \left(B_t(\Pi_t \circ \varphi(\mathbf{u}_k; \mathcal{V}_t); I_t) \right). \quad (4.6)$$

From the input image I_t , we build B_t using an edge response filter that is modulated to suppress false positives according to one or more segmentation cues. We use two different segmentation cues: the projection-based and the color-distribution segmentation cues. An illustration of a boundariness-map is given in Fig. 4.4b. The exact choice for computing B_t for each tested dataset is reported in [25].

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

$$C_{iso}(\mathcal{V}_1, \mathcal{V}_t) \triangleq \frac{1}{|E|} \sum_{(i,j) \in E} \left(1 - \|\mathbf{v}_1^i - \mathbf{v}_1^j\|_2^{-2} \|\mathbf{v}_t^i - \mathbf{v}_t^j\|_2^2 \right)^2. \quad (4.7)$$

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

The crease-preserving smoothing constraint. We propose to use from [26], the smoothing constraint based on M-estimators [90]. This will lead to a discontinuity-preserving smoother which automatically deactivates smoothing, where needed at

creased regions. Precisely, this constraint penalizes the surface curvature change using a robust bending energy as follows:

$$C_{smooth}(\mathcal{V}_t) \triangleq \frac{1}{|\Omega|} \sum_{\mathbf{u}_j \in \Omega} \rho_1 \left(\frac{\partial^2 \varphi}{\partial \mathbf{u}^2}(\mathbf{u}_j; \mathcal{V}_t) \right). \quad (4.8)$$

In practice, for this constraint, we compute the curvature change in a discrete way. This can be done analytically because position and gradient can be computed using the barycentric coordinates, which is a linear operation in the unknowns, i.e., the vertices. The ability of this constraint to allow creases formation comes from the behavior of the M-estimator for high residuals. Regarding our problem, high residuals in the regularizer correspond to high changes of curvature, which occur at creased regions. Observing the behavior of several M-estimator functions reveal that they grow sub-quadratically at high residuals. Therefore, the impact of high residuals on the optimization of the regularizer will be much smaller when using an M-estimator rather than the ℓ_2 norm, which is used by most of the current methods for the smoothing constraint. It is, however, important to consider that the creases formation is encouraged by the data terms and allowed by the smoothing constraint.

Handling scale. In the cost function (4.1), the shading, the motion, the boundary contour, and the quasi-isometry constraints are invariant to the scale of the reconstruction, however, the smoothing constraint is not invariant. This is because a trivial solution for the smoothing constraint is to put all vertices at the origin. Therefore, to rule out the dependency on scale, we constrain the mean depth of the reconstruction to a fixed positive value. Details are given in Sect. 4.4.3.2.

4.4.2 Optimization Strategy

Optimizing Eq. (4.1) is a nontrivial task because it is large-scale (typically $O(10^5)$ unknowns), is highly non-convex, and the shading constraint requires dense, pixel-level registration. Recall that we do not assume that the images come from an uninterrupted video sequences, which makes dense registration much harder to achieve. We propose a strategy in four stages, illustrated in Fig. 4.3.

Stage 1: We first achieve a rough initial estimate for the shape parameters $(\mathcal{D}, \mathcal{V}_2, \dots, \mathcal{V}_N)$ (and hence an initial estimate for registration) using only motion constraints from the point correspondences. We do this using the initialization-free NRSfM method [14] which has publicly available code.¹ Note that all existing initialization-free surface-based methods assume that the object’s surface is smooth in all views, thus the initial estimate will not normally be highly accurate. This provides a rough estimate of the reference image’s vertex depths \mathcal{D} , which we use to back-project the mesh vertices in the reference image to obtain \mathcal{V}_1 .

¹The code is available at http://igt.ip.uca.fr/~ab/code_and_datasets/index.php (Matlab SFT Toolbox).

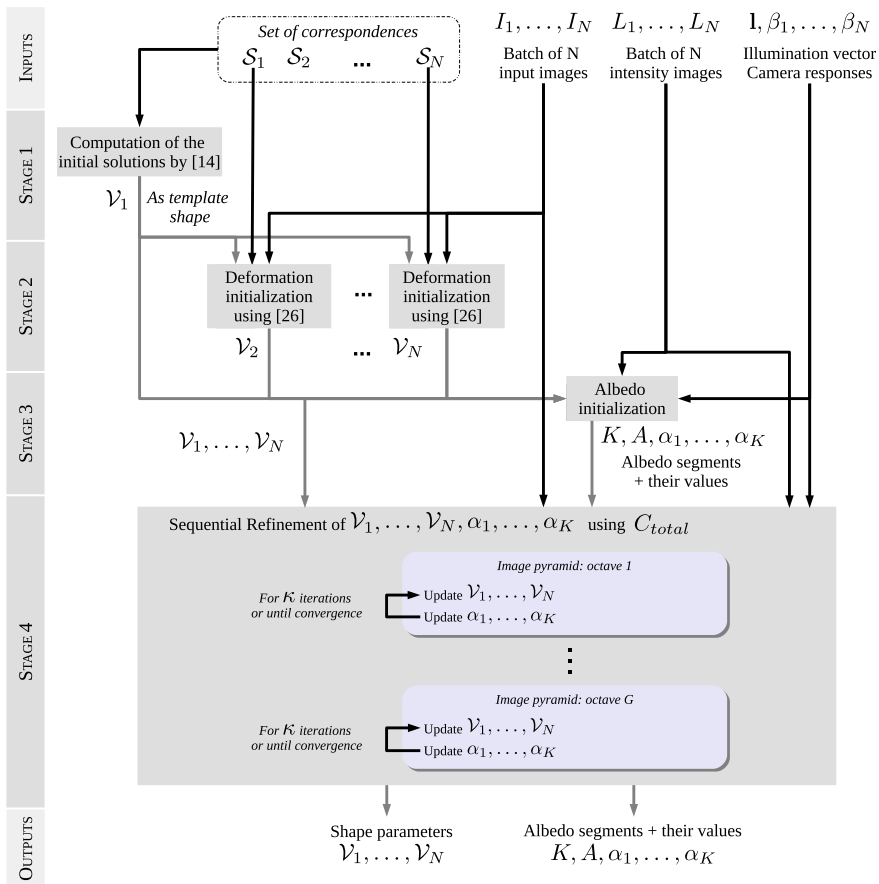


Fig. 4.3 Schematic of our proposed solution to solve NRSfMS

Stage 2: We then use \mathcal{V}_1 as a template and perform the SfT method [26] independently on each \mathcal{V}_t . It introduces the boundary contour constraints and refines the shape parameters by optimizing Eq. (4.1), with $\lambda_{shade} = 0$, using iterative numerical minimization. [26] also uses two strategies to improve the convergence of the refinement. The first is to refine only with the motion constraint as image data constraint, then we add the boundary contour constraint. The second is to construct from each input image I_t the boundariness-map B_t using an image pyramid, and sequentially optimize with each pyramid level. We found that three octaves for the pyramid level provide good convergence.

Stage 3: This consists of the segmentation of the reference image I_1 in regions of constant albedos and in the estimation of the albedos by inverting the shading equation. For this, we use an intrinsic image decomposition [10], on the reference image's intensity image and cluster the resulting "reflectance image" using [24], with a low

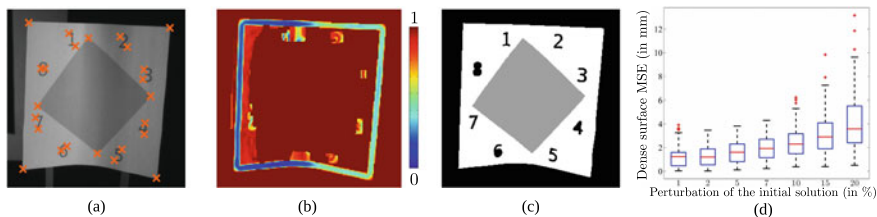


Fig. 4.4 **a** Visualization of the correspondences of the input image n°1 (zoom) of the *paper fortune teller* dataset. **b** Boundariness-map (zoom) for the input image n°1 of the *paper fortune teller* dataset. **c** Albedo-map (zoom) estimated for the *paper fortune teller* dataset. **d** Numerical results of convergence basin analysis for the *paper fortune teller* dataset

cluster tolerance (we use a default of 10). For each cluster k , we assign a corresponding albedo α_k : for each pixel \mathbf{u}_j in the cluster, we estimate its albedo by inverting the shading equation: $\alpha \approx L_t (\Pi_t \circ \varphi(\mathbf{u}; \mathcal{V}_t)) r(n(\mathbf{u}; \mathcal{V}_t); \mathbf{I})^{-1}$. We then initialize α_k as the median overall estimates within the cluster. This can be done because, at this stage, we have estimated the scene illumination, the camera response for each image and the shape parameters. We aim for an oversegmentation: neighboring segments can share the same albedo but within each segment we assume the albedo constant. The reason is that our method is not designed to recover from an under segmentation. Even if oversegmentation requires more unknowns, under segmentation is a more difficult problem since it may strongly impact the estimation of surface orientation and illumination and requires then an automatic process to re-segment the albedo-map when needed. The last step of the clustering is the thresholding of the pixels number of each albedo segment to remove the ones which correspond to the textured regions.

Figure 4.4c shows an illustration of a segmented albedo-map: the black holes visible on the surface corresponds to the textured regions whose area is smaller than T_A . The black holes visible on the surface in Fig. 4.4c corresponds to these textured regions. If there are K segments, then the albedo set $\{\alpha_1, \dots, \alpha_K\}$ has size K .

Stage 4: We refine alternately the shape parameters and the albedo values by minimizing Eq. (4.1) using all constraints. This 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 use the two strategies of [27]. First, we use only the motion constraint as image data constraint to refine, then we add the boundary contour constraint and end by refining the three image data constraints. Second, we construct from each input image I_t the boundariness-map B_t using an image pyramid, blur each L_t with a Gaussian blur pyramid, and sequentially optimize with each pyramid level, with a default of three octaves. For the two first

levels Gaussian blur pyramid, the kernel sizes, and standard-deviations are respectively $h_1 = (10, 10)$ and $\sigma_1 = 5$ and $h_2 = (5, 5)$ and $\sigma_2 = 2.5$. At the finest level, we do not apply any Gaussian blur. For the three pyramid levels, we run Gauss-Newton until either convergence is reached (with the total cost difference between two consecutive iterations being strictly lower than $1e-4$) or a fixed number of iterations have passed (we use $\kappa = 20$ iterations).

4.4.3 Experimental Validation

We divide the experimental validation into two parts. First, we analyze the convergence basin of our energy function through perturbation analysis. This is to understand how sensitive our formulation is to the initial solution, and fundamentally, whether the NRSfM problem can be cast as an energy-based minimization with a strong local minimum near the true solution. Second, we compare performance to state-of-the-art NRSfM methods, using six datasets, all with ground-truth.

4.4.3.1 Methods Compared and Datasets

We compare with the following competitive NRSfM methods [14, 60, 81], denoted, respectively, with **Va09**, **Ch14**, **Pa16**. We compare to these methods because they reconstruct dense surfaces. To see the contribution of some constraints of Eq. (4.1), we compare with four versions of our method, **NoS**, where shading is not used, **NoB**, where the boundary constraint is not used in stages 2 and 4, **NoI**, where the quasi-isometry constraint is not used in stages 2 and 4, and **NoSm**, where the smoothing constraint is not used in stages 2 and 4.

We evaluated on six real-world datasets which mostly respect the Lambertian assumption: *floral paper* and *paper fortune teller* from [27], *creased paper*, *pillow cover* and *hand bag* from [28] and *Kinect paper* from [80]. Each dataset consists of a disc-topology surface in 5 different deformed states, with one state per image. We show them in Figs. 4.6 and 4.7. The five first datasets have the following conditions: (i) the object has a poorly-textured surface, (ii) several images show the surface creased, (iii) a highly-accurate depth-map associated with each image, (iv) the illumination vector is in 3D camera coordinates. These five datasets have been acquired with the structured light system [19]. As the *Kinect paper* has no accompanying illumination parameters and no camera response function, these are computed prior to the reconstruction. More details are given in [25]. Each dataset has a set of point correspondences between the first and all other images. As all datasets, except the *Kinect paper* dataset, are poorly-textured, the correspondences are sparse. We note that manual correspondences are commonly used to evaluate NRSfM methods and this is why the correspondences of our datasets were computed manually. These correspondences are distinctive points such as the texture discontinuities along the printed numbers of the *paper fortune teller* dataset, visible in Fig. 4.4a. The datasets *floral*

paper, *paper fortune teller*, *creased paper*, *pillow cover*. and *hand bag* have, respectively, 20, 24, 20, 69, and 155 correspondences and their image size is 1288×964 pixels. The *Kinect paper* dataset presents images with 640×480 pixels and 1503 correspondences computed by [29]. The datasets *floral paper*, *paper fortune teller*, *creased paper*, *pillow cover*, and *hand bag* are publicly available.²

4.4.3.2 Implementation Details and Evaluation Metrics

We constructed, for all experiments, the reference meshes by laying a triangulated 100×100 vertex regular grid on the reference image which was then cropped to Ω . We also discretized the boundary points of the texture-map to $N_{\mathcal{B}} = 1000$ uniformly spaced points. For the compared methods, there is no way to automatically optimize their hyperparameters. We then tried our best to do this by hand, to obtain the best reconstruction accuracy on all datasets. For our method, all experiments were ran using the same hyperparameters, which were manually set. In Appendix Sect. 4.5, Tables 4.1 and 4.2 give the weights of the different constraints and the hyperparameters for our method and the compared methods.

To measure reconstruction accuracy, we compared 3D distances and normals with respect to ground-truth using, respectively, the Mean Shape Error (MSE) and the Mean Normal Error (MNE). To investigate the contribution of the shading constraint, this was done at two locations: (i) densely across the ground-truth surface, and (ii) densely at creased regions, which are any points on the ground-truth surfaces that are within 5 mm of a surface crease. Both grids were constructed by sampling uniformly the respective locations. Because reconstruction is up to scale, we computed for each method the best-fitting scale factor that aligns the predicted point correspondences with their true locations in the ℓ_2 sense, then measured accuracy with the scale-corrected reconstruction.

4.4.3.3 Convergence Basin Analysis

We performed this with 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, as we do not expect them in the initial solution), and randomly perturbed the vertex positions using smooth deformation functions. For each perturbation, we optimized Eq. (4.1), by performing stages 3 and 4 in Sect. 4.4.2. The perturbation was implemented using a $4 \times 4 \times 4$ B-spline enclosing the reconstructed surfaces and randomly perturbing the spline control points at 7 different noise levels, with 30 random perturbations per noise level. Figure 4.4d reports results as box-plots for the *paper fortune teller* dataset. The x -axis gives the average perturbation in % for each noise level from the initial solution. The y -axis gives the dense surface MSE for each random sample. Similar results are obtained for the *floral paper* and *creased paper*

²The datasets are available at http://igt.ip.uca.fr/~ab/code_and_datasets/index.php.

datasets and are reported in [25]. 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. This supports our claim that the NRSfMS problem can be cast as an energy-based minimization (via Eq. (4.1)). 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.

4.4.3.4 Quantitative and Qualitative Results

We show in Figs. 4.6 and 4.7, the six datasets and their reconstructions from our method and the best performing previous method (the one with lowest MSE 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 Fig. 4.5, we give the reconstruction accuracy statistics across all test datasets and all compared methods. The *Kinect paper* dataset 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. For the *Kinect paper* dataset, 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. However, our method still obtains competitive results on this dataset. We observe that the use of shading improves globally the shape of the reconstructions and that the boundary contour constraint allows using shading better. The reduced performance with **NoSm** confirms that the smoothing constraint acts as a regularizer. An observation of the 3D surfaces reconstructed without the smoothing constraint shows that the creases cannot be formed and the surfaces are very smooth. Figure 4.5 does not show the results for **NoI** because, for every dataset, the surfaces collapse to the origin during stage 2. This is consistent with the fact that isometry constraint makes the problem well-posed as it has been mentioned in Sect. 4.3.1.1. In Appendix Sect. 4.5, Table 4.3 gives the processing times for our method and the three compared methods with respect to each dataset.

4.4.3.5 Limitations and Failure Modes

We discuss here the main limitations and the failure modes of our solution to NRSfMS. Our method is limited by the assumptions made in Sect. 4.4.1. These are that we have isometric deformations, piecewise-constant albedo, fixed and known illumination vector, and known camera responses. One important limitation of our solution is the parameter tuning since the parameters of our method are set manually and may vary with the datasets. This is because we observe that we did not find default parameters for all datasets yielding to the best reconstruction accuracy. It would be interesting to investigate whether there exist fixed tuning parameters which work well on all datasets, using, e.g., grid search. Our approach regarding

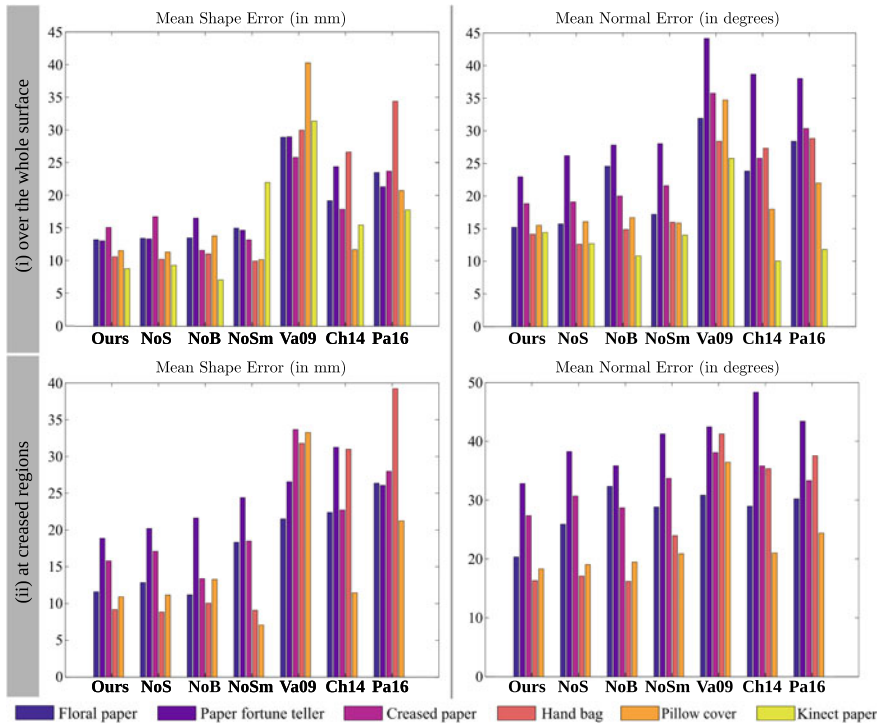


Fig. 4.5 Reconstruction accuracy statistics across all test datasets and all compared methods. Also, the *Kinect paper* dataset does not present any crease

the estimation of the albedo segments and their values present a drawback. Our method uses a single image, the reference image, to estimate the albedo segments and it cannot split or merge them during the following steps because of our modeling given in Sect. 4.4.1.2. Particularly, this may be problematic when some deformations occurring in the reference image lead the albedo initialization to merge two albedo segments with significantly different values. The constraint on the fixed number of albedo segments can be relaxed and mechanisms to automatically adjust them during the refinement step can be studied, such as the cost term of [33] which encourages piecewise-constant albedo segments by penalizing gradients on the albedo value through a ℓ_0 norm. Another limitation is that we perform our experiments with batches of 5 images and we have not performed a theoretical analysis to establish the minimum number of images to solve NRSfMS. Some failure modes are caused by the joint use of shading and motion visual cues. The first one is that, as in SfS, our method may then suffer from localized convex/concave ambiguities, which tends to worsen flawed initial solutions. The second failure mode is the under segmentation of the albedo-map, which may lead to incorrect surface orientation. The third one is the presence of some false positive creases, as Figs. 4.6 and 4.7, show. This failure mode is linked to the first one, but is more general and can integrate other sources

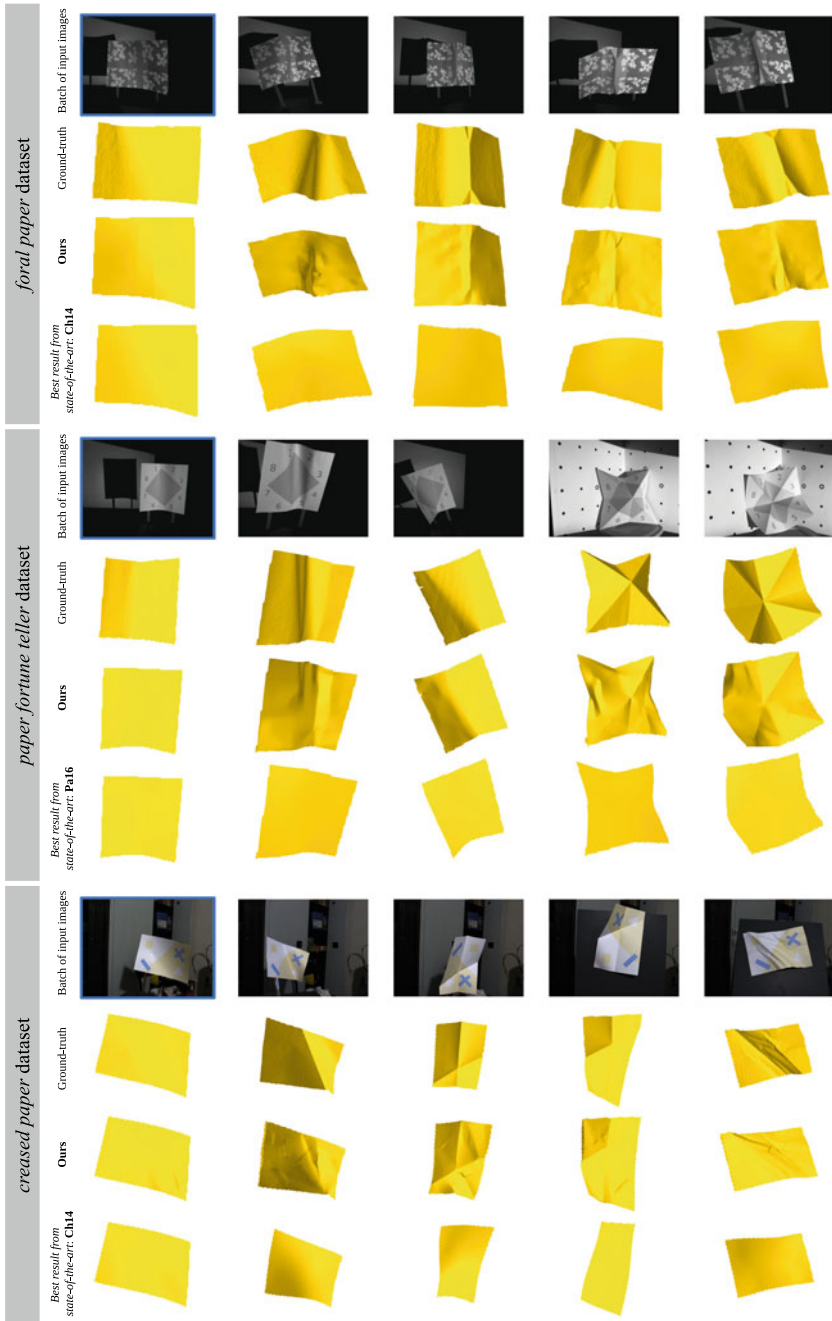


Fig. 4.6 Renderings for the *floral paper*, the *paper fortune teller* and *creased paper* datasets with ground-truth. Here we show the images from each dataset, and sample reconstructions from one of the images using our method and the best performing NRSfM method. We frame the reference image in blue

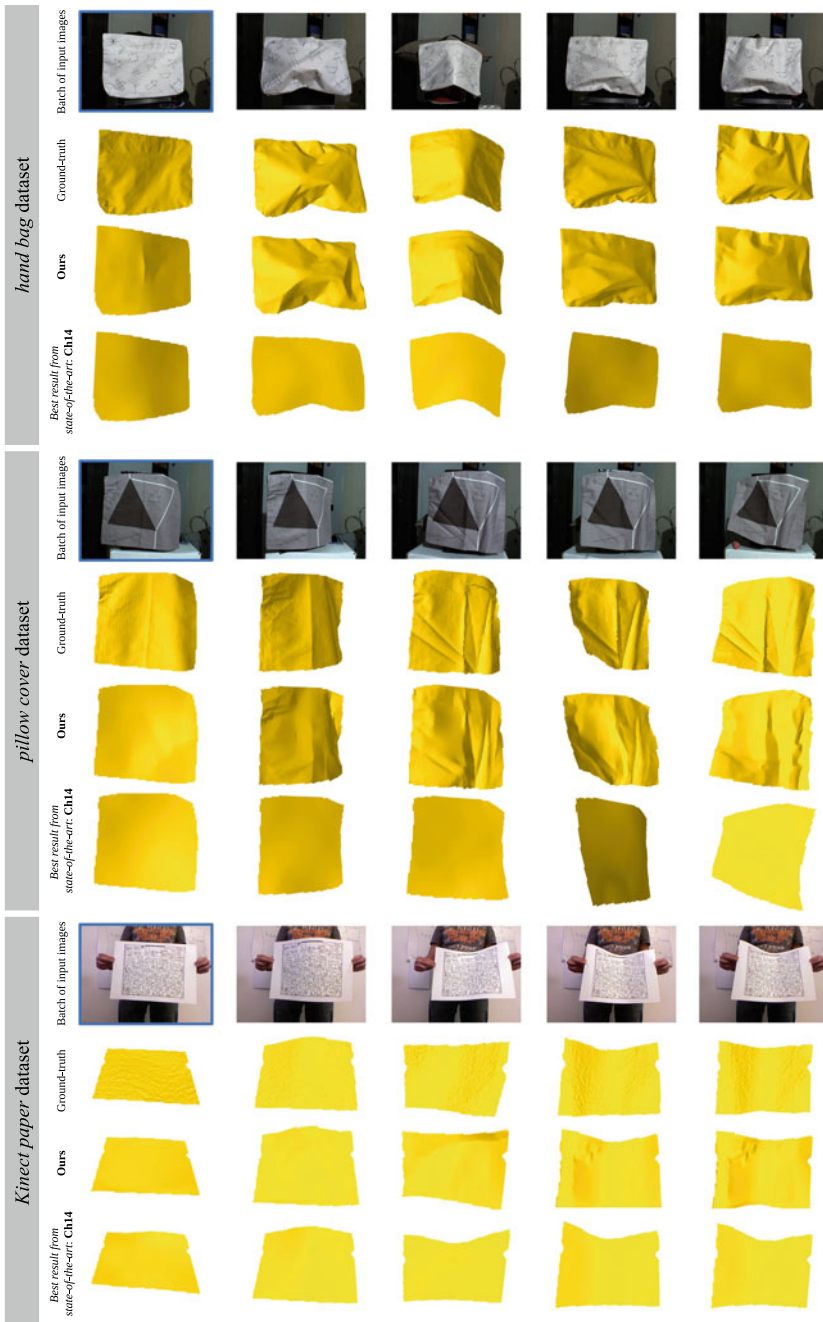


Fig. 4.7 Renderings for the *hand bag*, the *pillow cover* and *Kinect paper* datasets with ground-truth. Here we show the images from each dataset, and sample reconstructions from one of the images using our method and the best performing NRSfM method. We frame the reference image in blue

of errors such as misregistration or the robust estimator applied in the shading constraint. The last failure mode, which is not caused by the use of shading, is when the initial solutions given by stage 1 are not reliable. 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 SFT method. For unorganized 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 [17].

4.4.4 Conclusion and Open Problems

We have presented the first study of the NRSfMS problem as an illustration of the combination of motion and shading visual cues to infer the 3D shape of deforming objects from a single camera. NRSfMS does not assume the 3D geometry of the surface is known prior to reconstruction, and solves the problem using only a set of images and models pertaining to camera projection, scene illumination, surface reflectance, and camera response. We have shown for the first time that it is possible to solve NRSfMS when some of the model parameters are unknown (specifically surface reflectance) and solve jointly with reconstruction. NRSfMS is a hard and important vision problem, needed for high-accuracy dense reconstruction of poorly-textured surfaces undergoing non-smooth deformation from 2D images. We have proposed an energy-based solution and a cascaded numerical optimization strategy, and have shown encouraging results on six real-world datasets, for which all competitive NRSfM methods fail. This marks the first time that strongly creased, deformable, poorly-textured surfaces with unknown albedos have been densely reconstructed and registered from 2D image sets without shape prior knowledge on the object.

There are many possible future directions for NRSfMS. These involve both theoretical analysis to understand the problem well-posedness, and explorations to release some of the practical limits of our approach as given in Sect. 4.4.3.5. Regarding the former, NRSfM can be solved up to ambiguities with two images and SfS can be solved with one image when the illumination and the surface reflectance are known. At first sight, two images seem to be sufficient to solve NRSfMS, however, a thorough theoretical study would be required for the version of NRSfMS presented in Sect. 4.4 and also for the other possible instantiations of NRSfMS.

Regarding the latter, we propose two main research directions. The first direction is to consider more uncontrolled settings and examine other strategies to use motion in order to improve the use of shading. Examples of these settings are when the scene illumination is not known a priori or when the relative position between the camera and the light source changes over time. This will require a careful theoretical study of well-posedness, innovative initialization and optimization strategies. Our work shows that motion can provide an accurate reconstruction of surface normals at well-textured regions, and these regions can be used to estimate photometric parameters. It may be possible to estimate other photometric parameters such as camera response,

using data at the reconstructed textured regions. A second direction considers occlusions and shadows. Some solutions have been proposed for handling occlusions in SfT [17, 52, 57] and external occlusions in NRSfM [60], and for handling shadows in SfS [22]. However, there is no attempt to reason simultaneously about occlusions and shadows in NRSfMS, which is required to achieve robust reconstruction in the wild. Reasoning first with the correspondence constraints, i.e., features motion over the surface may provide more robust initial solutions which can be then more easily refined by reasoning with shading constraints.

4.5 Appendix

Tables 4.1 and 4.2 give the hyperparameters which we used to produce the results on the six datasets used in Sect. 4.4.3. We denote the SfT method [26], used in stage 2 of our NRSfMS method, with **Ga16**.

Table 4.3 gives the average processing times for our NRSfMS method and the compared methods for each dataset. We refer to Sect. 4.4.3.2 for the implementation details and Sect. 4.4.3.1 for the dataset details.

Table 4.1 Hyperparameter values used to evaluate our NRSfMS method

		<i>floral paper</i>	<i>paper fortune teller</i>	<i>creased paper</i>	<i>pillow cover</i>	<i>hand bag</i>	<i>kinect paper</i>
Ga16	M	1e4					
	$N_{\mathcal{G}}$	1e3					
	$\lambda_{contour}$	1e-5	4e-4	4e-4	4e-4	4e-4	0.04
	λ_{iso}	4e-4	0.16	4e-3	4e-3	0.04	0.04
	λ_{smooth}	6e-15	2.4e-13	1.6e-14	1.6e-14	1.6e-14	4e-13
Ours	M	1e4					
	$N_{\mathcal{G}}$	1e3					
	k_{shade}	5e-3	5e-3	5e-3	5e-3	5e-3	5e-3
	λ_{motion}	0.088	0.154	1	10	10	10
	$\lambda_{contour}$	1.25e-4	0.011	0.01	1.67e-4	1.67e-4	1.67e-4
	λ_{iso}	3.8e-3	0.025	0.167	0.167	0.167	0.5
	λ_{smooth}	2.5e-12	9.2e-12	3.33e-11	3.33e-11	3.33e-11	8.33e-10

Table 4.2 Hyperparameter values used to evaluate all compared NRSfM methods

	<i>floral paper</i>	<i>paper fortune teller</i>	<i>creased paper</i>	<i>pillow cover</i>	<i>hand bag</i>	<i>kinect paper</i>	
Va09	depth.nC	30	30	30	28	30	30
	depth.er	6	0.06	0.2	8	0.2	6
	embedding.nC	30					
	embedding.er	0.01	1e-6	1e-6	1e-4	1e-6	0.01
	homographies.neigh	100					
Ch14	depth.nC	28	30	28	16	28	30
	depth.er	5	1	0.7	1	8	0.9
	warps.nC	28	20	28	16	28	30
	warps.er	0.01	1e-3	9e-4	1e-4	1e-3	0.01
	homographies.neigh	40	40	40	80	40	40
Pa16	schwarzianParam	2e-5					1e-3
	warps.nC	60					
	warps.er	1e-4					
	depth.nC	100					
	depth.er	1					10

Table 4.3 Processing time in minutes for our method and the three compared methods with respect to each dataset. For the compared methods, we explain the differences in time for the *Kinect paper* dataset because of the number of correspondences which is significantly larger than the other five datasets

	<i>floral paper</i>	<i>paper fortune teller</i>	<i>creased paper</i>	<i>pillow cover</i>	<i>hand bag</i>	<i>kinect paper</i>
Ours	28'39	41'40	20'4	45'23	34'47	35'58
Va09	0'6	0'5	0'6	0'6	0'58	11'11
Ch14	0'18	0'15	0'14	0'8	0'29	1'33
Pa16	9'41	14'10	9'19	7'15	12'38	30'54

References

- 3Dflow (2017) 3DF Zephyr. <https://www.3dflow.net>
- Agisoft (2014) PhotoScan version 1.2.3 build 2331. <http://www.agisoft.com>
- Ahmed AH, Farag AA (2007) Shape from shading under various imaging conditions. In: CVPR
- Akhter I, Sheikh Y, Khan S, Kanade T (2009) Nonrigid structure from motion in trajectory space. In: NIPS
- Barron JT, Malik J (2015) Shape, illumination, and reflectance from shading. IEEE Trans Pattern Anal Mach Intell 37(8):1670–1687
- Bartoli A, Gérard Y, Chadebecq F, Collins T, Pizarro D (2015) Shape-from-template. IEEE Trans Pattern Anal Mach Intell 37(10):2099–2118
- Bartoli A, Özgür E (2016) A perspective on non-isometric shape-from-template. In: ISMAR
- Beeler T, Bickel B, Beardsley P, Sumner B, Gross M (2010) High-quality single-shot capture of facial geometry. In: SIGGRAPH

9. Belhumeur PN, Kriegman DJ, Yuille AL (1997) The bas-relief ambiguity. In: CVPR
10. Bell S, Bala K, Snavely N (2014) Intrinsic images in the wild. *ACM Trans Graph (SIGGRAPH)* 33(4)
11. Bregler C, Hertzmann A, Biermann H (2000) Recovering non-rigid 3D shape from image streams. In: CVPR
12. Brostow GJ, Hernández C, Vogiatzis G, Stenger B, Cipolla R (2001) Video normals from colored lights. *IEEE Trans Pattern Anal Mach Intell* 33(10):2104–2114
13. Brunet F, Hartley R, Bartoli A (2014) Monocular template-based 3D surface reconstruction: convex inextensible and nonconvex isometric methods. *Comput Vis Image Underst* 125:138–154 August
14. Chhatkuli A, Pizarro D, Bartoli A (2014) Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In: BMVC
15. Chhatkuli A, Pizarro D, Bartoli A, Collins T (2017) A stable analytical framework for isometric shape-from-template by surface integration. *IEEE Trans Pattern Anal Mach Intell* 39(5):833–850
16. Chhatkuli A, Pizarro D, Collins T, Bartoli A (2016) Inextensible non-rigid shape-from-motion by second-order cone programming. In: CVPR
17. Collins T, Bartoli A (2015) Realtime shape-from-template: system and applications. In: ISMAR
18. Dai Y, Li H, He M (2014) A simple prior-free method for non-rigid structure-from-motion factorization. *Int J Comput Vis* 107(2):101–122
19. David 3D Scanner (2014) <http://www.david-3d.com/en/products/david4>
20. Durou J-D, Falcone M, Sagona M (2008) Numerical methods for shape-from-shading: a new survey with benchmarks. *Comput Vis Image Underst* 109(1):22–43 January
21. Ecker A, Jepson AD (2010) Polynomial shape from shading. In: CVPR
22. Falcone M, Sagona M, Seghini A (2003) A scheme for the shape-from-shading model with “black shadows”. *Numerical mathematics and advanced applications*, pp 503–512
23. Fayad J, Agapito L, Del Bue A (2010) Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In: ECCV
24. Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans Inf Theory* 21(1):32–40
25. Gallardo M (2018) Contributions to monocular deformable 3D reconstruction: curvilinear objects and multiple visual cues. Theses, Université Clermont Auvergne. <https://tel.archives-ouvertes.fr/tel-01930477>
26. Gallardo M, Collins T, Bartoli A (2016) Can we jointly register and reconstruct creased surfaces by shape-from-template accurately? In: ECCV
27. Gallardo M, Collins T, Bartoli A (2016) Using shading and a 3D template to reconstruct complex surface deformations. In: BMVC
28. Gallardo M, Collins T, Bartoli A (2017) Dense non-rigid structure-from-motion and shading with unknown albedos. In: ICCV
29. Garg R, Roussos A, Agapito L (2013) Dense variational reconstruction of non-rigid surfaces from monocular video. In: CVPR
30. Gotardo PFU, Martinez AM (2011) Kernel non-rigid structure from motion. In: ICCV
31. Gotardo PFU, Simon T, Sheikh Y, Matthews I (2015) Photogeometric scene flow for high-detail dynamic 3D reconstruction. In: ICCV
32. Gumerov N, Zandifar A, Duraiswami R, Davis LS (2004) Structure of applicable surfaces from single views. In: ECCV
33. Haefner B, Quéau Y, Möllenhoff T, Cremers D (2018) Fight ill-posedness with ill-posedness: single-shot variational depth super-resolution from shading. In: CVPR
34. Haefner B, Ye Z, Gao M, Wu T, Quéau Y, Cremers D (2019) Variational uncalibrated photometric stereo under general lighting. In: ICCV
35. Haouchine N, Dequidt J, Berger MO, Cotin S (2014) Single view augmentation of 3D elastic objects. In: ISMAR
36. Hartley RI, Zisserman A (2003) *Multiple view geometry in computer vision*, 2nd edn. Cambridge University Press, Cambridge

37. Horn BKP (1970) Shape from shading: a method for obtaining the shape of a smooth shape of a smooth opaque object from one view. Technical report, Cambridge, MA, USA
38. Horn BKP (1989) Shape from shading, pp 123–171
39. Ikeuchi K, Horn BKP (1981) Numerical shape from shading and occluding boundaries. *Artif Intell* 17(1):141–184
40. Innmann M, Zollhöfer M, Nießner M, Theobalt C, Stamminger M (2016) VolumeDeform: real-time volumetric non-rigid reconstruction. In: ECCV
41. Jin H, Cremers D, Wang D, Prados E, Yezzi A, Soatto S (2008) 3-D reconstruction of shaded objects from multiple images under unknown illumination. *Int J Comput Vis* 76(3):245–256
42. Kim K, Torii A, Okutomi M (2016) Multi-view inverse rendering under arbitrary illumination and albedo. In: ECCV
43. Kimmel R, Bruckstein AM (1994) Global shape-from-shading. In: ICPR
44. Koo B, Özgür E, Le Roy B, Buc E, Bartoli A (2017) Deformable registration of a preoperative 3D liver volume to a laparoscopy image using contour and shading cues. In: MICCAI
45. Langguth F, Sunkavalli K, Hadap S, Goesele M (2016) Shading-aware Multi-view Stereo. In: ECCV
46. Lee KM, Kuo C-C (1997) Shape from shading with a generalized reflectance map model. *Comput Vis Image Underst* 67(2):143–160
47. Lee KM, Kuo CCJ (1993) Shape from shading with a linear triangular element surface model. *IEEE Trans Pattern Anal Mach Intell* 15(8):815–822
48. Liu-Yin Q, Yu R, Agapito L, Fitzgibbon A, Russell C (2016) Better together: joint reasoning for non-rigid 3D reconstruction with specularities and shading. In: BMVC
49. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
50. Magnenat S, Ngo DT, Zund F, Ryffel M, Noris G, Rothlin G, Marra A, Nitti M, Fua P, Gross M, Sumner R (2015) Live texturing of augmented reality characters from colored drawings. *IEEE Trans Vis Comput Graph* 21(11):1201–1210
51. Malti A, Bartoli A (2014) Combining conformal deformation and cook-torrance shading for 3D reconstruction in laparoscopy. *IEEE Trans Biol Eng* 61(6):1684–1692
52. Malti A, Bartoli A, Collins T (2011) A pixel-based approach to template-based monocular 3D reconstruction of deformable surfaces. In: Proceedings of the IEEE international workshop on dynamic shape capture and analysis at ICCV
53. Malti A, Bartoli A, Hartley RI (2015) A linear least-squares solution to elastic shape-from-template. In: CVPR
54. Moreno-Noguer F, Salzmann M, Lepetit V, Fua P (2009) Capturing 3D stretchable surfaces from single images in closed form. In: CVPR
55. Newcombe RA, Fox D, Seitz SM (2015) DynamicFusion: reconstruction and tracking of non-rigid scenes in real-time. In: CVPR
56. Ngo TD, Östlund J, Fua P (2016) Template-based monocular 3D shape recovery using laplacian meshes. *IEEE Trans Pattern Anal Mach Intell* 38(1):172–187
57. Ngo TD, Park S, Jorstad AA, Crivellaro A, Yoo C, Fua P (2015) Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture. In: ICCV
58. Okatani T, Deguchi K (1996) Shape reconstruction from an endoscope image by SfS technique for a point light source at the projection Center. *Comput Vis Image Underst* 66(2):119–131 July
59. Özgür E, Bartoli A (2017) Particle-SfT: a provably-convergent, fast shape-from-template algorithm. *Int J Comput Vis* 123(2):184–205 June
60. Parashar S, Pizarro D, Bartoli A (2016) Isometric non-rigid shape-from-motion in linear time. In: CVPR
61. Pentland AP (1984) Local shading analysis. *IEEE Trans Pattern Anal Mach Intell* 6(2):170–187
62. Pentland AP (1988) Shape information from shading: a theory about human perception. In: ICCV
63. Perriollat M, Hartley R, Bartoli A (2011) Monocular template-based reconstruction of inextensible surfaces. *Int J Comput Vis* 95(2):124–137 November

64. Prados E, Faugeras O (2005) Shape from shading: a well-posed problem? In: CVPR
65. Pumarola A, Agudo A, Porzi L, Sanfeliu A, Lepetit V, Moreno-Noguer F (2018) Geometry-aware network for non-rigid shape prediction from a single view. In: CVPR
66. Quéau Y, Mérou J, Castan F, Cremers D, Durou J (2017) A variational approach to shape-from-shading under natural illumination. In: EMMCVPR
67. Richardson E, Sela M, Kimmel R (2016) 3D face reconstruction by learning from synthetic data. In: 3DV
68. Richter SR, Roth S (2015) Discriminative shape from shading in uncalibrated illumination. In: CVPR
69. Rouy E, Tourin A (1992) A viscosity solutions approach to shape-from-shading. *SIAM J Numer Anal* 29(3):867–884 June
70. Salzmann M, Fua P (2009) Reconstructing sharply folding surfaces: a convex formulation. In: CVPR
71. Salzmann M, Fua P (2011) Linear local models for monocular reconstruction of deformable surfaces. *IEEE Trans Pattern Anal Mach Intell* 33(5):931–944
72. Salzmann M, Pilet J, Ilic S, Fua P (2007) Surface deformation models for nonrigid 3D shape recovery. *IEEE Trans Pattern Anal Mach Intell* 29(8):1481–1487
73. Sorkine O, Alexa M (2007) As-rigid-as-possible surface modeling. In: Symposium on geometry processing, pp 109–116
74. Sundaram N, Brox T, Keutzer K (2010) Dense point trajectories by GPU-accelerate large displacement optical flow. In: ECCV
75. Tankus A, Sochen N, Yeshurun Y (2005) Shape-from-shading under perspective projection. *Int J Comput Vis* 63(1):21–43 June
76. Taylor J, Jepson AD, Kutulakos K (2010) Non-rigid structure from locally-rigid motion. In: CVPR
77. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M (2016) Face2Face: real-time face capture and reenactment of RGB videos. In: CVPR
78. Tsai P-S, Shah M (1994) Shape from shading using linear approximation. *Image Vis Comput* 12(8):487–498
79. Valgaerts L, Wu C, Bruhn A, Seidel H-P, Theobalt C (2012) Lightweight binocular facial performance capture under uncontrolled lighting. In: SIGGRAPH
80. Varol A, Salzmann M, Fua P, Urtasun R (2012) A constrained latent variable model. In: CVPR
81. Varol A, Salzmann M, Tola E, Fua P (2009) Template-free monocular reconstruction of deformable surfaces. In: ICCV
82. Varol A, Shaji A, Salzmann M, Fua P (2012) Monocular 3D reconstruction of locally textured surfaces. *IEEE Trans Pattern Anal Mach Intell* 34(6)
83. Vicente S, Agapito L (2012) Soft inextensibility constraints for template-free non-rigid reconstruction. In: ECCV
84. Vicente S, Agapito L (2013) Balloon shapes: reconstructing and deforming objects with volume from images. In: 3DV
85. Wang X, Salzmann M, Wang F, Zhao J (2016) Template-free 3D reconstruction of poorly-textured nonrigid surfaces. In: ECCV
86. Wu C, Narasimhan SG, Jaramaz B (2010) A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *Int J Comput Vis* 86(2):211–228
87. Wu C, Varanasi K, Liu Y, Seidel HP, Theobalt C (2011) Shading-based dynamic shape refinement from multi-view video under general illumination. In: ICCV
88. Xiong Y, Chakrabarti A, Basri R, Gortler SJ, Jacobs DW, Zickler T (2015) From shading to local shape. *IEEE Trans Pattern Anal Mach Intell* 37(1)
89. Yu R, Russell C, Campbell NDF, Agapito L (2015) Direct, dense, and deformable: template-based non-rigid 3D reconstruction from RGB video. In: ICCV
90. Zhang Z (1997) Parameter estimation techniques: a tutorial with application to conic fitting. *Image Vis Comput* 15(1):59–76
91. Zhou Z, Wu Z, Tan P (2013) Multi-view photometric stereo with spatially varying isotropic materials. In: CVPR