Template-free 3D Reconstruction of Poorly-textured Nonrigid Surfaces

Xuan Wang[1], Mathieu Salzmann[2], Fei Wang[1], Jizhong Zhao[1]

[1]Xi'an Jiaotong University, China
[2]CVLab, EPFL, Switzerland
xwang.cv@gmail.com, mathieu.salzmann@epfl.ch, {wfx, zjz}@mail.xjtu.edu.cn

Abstract. Two main classes of approaches have been studied to perform monocular nonrigid 3D reconstruction: Template-based methods and Non-rigid Structure from Motion techniques. While the first ones have been applied to reconstruct poorly-textured surfaces, they assume the availability of a 3D shape model prior to reconstruction. By contrast, the second ones do not require such a shape template, but, instead, rely on points being tracked throughout a video sequence, and are thus ill-suited to handle poorly-textured surfaces. In this paper, we introduce a template-free approach to reconstructing a poorly-textured, deformable surface. To this end, we leverage surface isometry and formulate 3D reconstruction as the joint problem of non-rigid image registration and depth estimation. Our experiments demonstrate that our approach yields much more accurate 3D reconstructions than state-of-the-art techniques.

Keywords: Non-rigid 3D reconstruction, poorly-textured surfaces, template-free shape estimation

1 Introduction

This paper tackles the problem of estimating the 3D shape of a poorly-textured, nonrigid surface in all the frames of a monocular video sequence. Reconstructing the 3D shape of a deformable surface from monocular images is a challenging task, which has attracted a lot of attention over the years. The resulting algorithms can be roughly classified into two categories: Template-based methods and Non-rigid Structure from Motion (NRSfM) techniques.

Template-based methods [1–11] exploit the availability of a reference image in which the 3D shape is known and attempt to estimate the surface deformations in a new input image. To this end, they typically try to minimize an image-based cost function, which encodes how well the deformed surface reprojects in the input image. Since this information alone leaves reconstruction ambiguities, existing approaches have developed various shape priors. In particular, great progress has been made in template-based reconstruction by exploiting surface isometry [1–9], with methods yielding accurate reconstructions in an efficient manner [6, 12, 3, 13] and, in some rare cases, even tackling the case of poorly-textured surfaces [1, 2]. The main drawback of this approach, however, is its

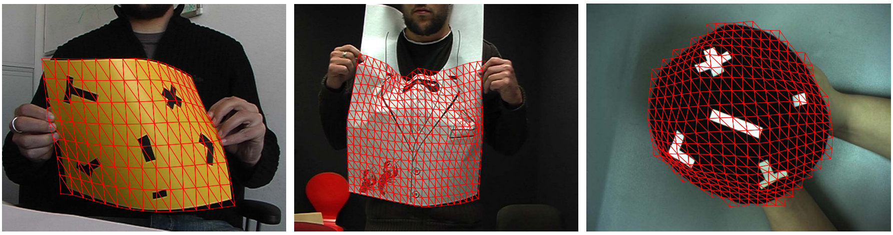


Fig. 1. Our approach: Template-free reconstruction of poorly-textured surfaces.

requirement for a 3D reference surface shape. Whether to model garments, mechanical structures, or human organs, one can in general not realistically expect having access to such prior knowledge.

By contrast, NRSfM techniques [14–20] do not require knowing the shape of the object a priori. Instead, they make use of a video sequence as input, and estimate the shape of the surface in all the frames of this sequence. To overcome the ambiguities of the resulting problem, existing methods also make use of additional priors, the most popular of which is a low-rank shape basis [15, 18]. As in the template-based case, impressive results have recently been achieved by NRSfM techniques, producing dense reconstructions [21] and working on natural video sequences [14]. However, to the best of our knowledge, all existing NRSfM approaches rely on tracking feature points throughout the input video sequence. As a consequence, they have only been applied to relatively well-textured objects and are ill-suited to handle the challenging case of poorly-texture surfaces.

In this paper, we aim to achieve the best of both worlds: We introduce a template-free reconstruction method able to tackle the poorly-textured surface scenario. To this end, we leverage the progress achieved in template-based reconstruction by exploiting surface isometry and image-based cost functions well-suited to poorly-textured objects, but perform 3D reconstruction in the NRSfM setting by not requiring prior information about the shape of the surface. To the best of our knowledge, our approach constitutes the first attempt at tackling reconstruction of poorly-textured surfaces without a shape template, thus taking a significant step towards making deformable surface reconstruction applicable to realistic scenarios.

Specifically, we model a deformable surface with a triangular mesh, and formulate 3D reconstruction as the problem of estimating an affine transformation for each mesh facet. This translates to optimizing a cost function combining an image term and a shape prior. As image term, we make use of the brightness constancy assumption, but also leverage image edges, which, while few in the poorly-textured case, provide stronger cues than image intensities. Inspired by template-based methods, our shape prior encodes surface isometry, to encourage the length of the edges of the mesh to remain constant as the surface deforms. Note, however, that, in contrast with template-based methods, we do not know these lengths a priori, and therefore need to optimize them as well. In addition to isometry, we further incorporate priors encouraging spatial and temporal smoothness of the deformations. To optimize our cost function, we make use of a fusion moves strategy, which has proven more robust to local minima than gradient-based methods.

We demonstrate the benefits of our approach on several synthetic and real sequences. Our experiments evidence, both quantitatively and qualitatively, that our method yields much more accurate reconstructions than state-of-the-art NRSfM techniques in the poorly-textured scenario.

2 Related Work

In this section, we briefly review the literature on monocular 3D reconstruction of deformable surfaces, i.e., template-based methods [1–11] and NRSfM [14–19]. In both cases, we focus the discussion on the methods that are the most relevant to our work.

Template-based nonrigid reconstruction has a long history in computer vision that can be traced back to the early physics-based models [22–24] and active appearance models [25, 26] or morphable models [27–29]. Given a template shape of the object corresponding to a reference image, the underlying task consists of deforming the template such that it reprojects at the correct location in the input image. Recently, great progress has been made in this line of research, especially by exploiting surface isometry to disambiguate the problem [1–9]. Most of the resulting methods, however, tackle the case of relatively well-textured surfaces [3, 4, 6–9]. Nevertheless, some approaches have proposed to focus on the poorly-textured scenario [1, 2, 5]. In particular, [5] exploits training data to learn local deformation models, which then act as a prior during reconstruction. By contrast, [1] and [2] avoid having to learn such a prior and rely on surface isometry. Similarly to our approach, these techniques rely on some more or less sophisticated variants of the brightness constancy assumptions, and [1, 5] exploit image edges as additional cues. However, these methods all require a known 3D template of the surface, which, in many practical situations, is difficult, if not impossible, to obtain.

Non-rigid Structure from Motion was initially introduced as an extension of the factorization method of [30] to the nonrigid scenario [15]. Given the 2D tracks of feature points throughout a video sequence, NRSfM aims at recovering the 3D locations of these points in each video frame, as well as the camera rotation and translation. As for template-based reconstruction, great progress has been achieved in NRSfM, notably by going beyond the traditional shape basis representation [16, 18]. Recently, impressive results were obtained by [21], which performs dense reconstruction by replacing the usual feature tracking step with a nonrigid registration procedure [31]. The notion of isometry has also been exploited in the context of NRSfM, with the additional difficulty, compared to the template-based case, that the true local surface distances are unknown and must therefore also be estimated. In particular, [17] introduced an approach based on a triangle soup surface representation, where triangles of neighboring feature points are assumed to move rigidly. In [20], isometry and infinitesimal planarity are employed, and reconstruction is achieved by integrating the normals obtained from the decomposition of local 2D homographies. In [19], isometry is encoded by encouraging the distance between neighboring feature points to remain constant over the entire sequence. Similarly to our approach, this method relies on a fusion moves optimization strategy, which makes it more robust to local minima than a gradient-based approach. All the above-mentioned NRSfM techniques, however, have been designed to handle relatively well-textured surfaces.

In this paper, we aim to achieve the best of both worlds: We introduce an approach that does not require a template of the surface of interest, but can nonetheless reconstruct poorly-textured surfaces. To the best of our knowledge, our work represents the first attempt at tackling this challenging template-free monocular reconstruction of poorly-textured deformable surfaces.

3 Our Approach

Let us now introduce our template-free approach to 3D reconstruction of poorly-textured, deformable surfaces. Given a monocular sequence of F frames with known intrinsic camera parameter matrix $\mathbf{K}$ depicting a deforming surface, our goal is to estimate the 3D shape of the surface in each frame of the video. Here, we represent the surface as a triangular mesh. We assume to be given a rough region of interest (ROI) containing the surface in the first frame of the sequence. Note that a similar assumption is made by most NRSfM methods, where only feature points belonging to the surface of interest are taken into account. We then cover this ROI with a regular 2D triangulation, which defines the 2D locations of the mesh vertices in the first frame. Note that this still makes no assumption about the initial 3D shape of the surface; if the surface is not flat in the first frame, the 3D mesh will simply not be regular.[1]

We formulate 3D reconstruction as the problem of jointly estimating the 2D displacement $\mathbf{u}_i^f$ of each mesh vertex i in each frame $2 \leq f \leq F$ with respect to the first frame and the depth d_i^f of the vertices in all the frames.[2] This is expressed as an optimization problem containing an image-based energy and shape priors. Below, we discuss these different terms in details.

3.1 Image-based Energy

Since we aim at tackling the poorly-textured scenario, we cannot expect to be able to reliably track feature points across the frames. Instead, we therefore exploit two sources of image information. The first one is based on the brightness constancy assumption, and the second one relies on image edges. Both terms will

[1] We acknowledge that this initialization might not be ideal when the first frame depicts very complex deformations, since a single mesh facet might then cover a large portion of the surface. However, as evidenced by our experiments, it remains effective in practice.

[2] While we do not explicitly model the camera motion, it can be accounted for by the mesh, which our parametrization allows to move freely in 3D.

require accessing image information at arbitrary points on the mesh, not just at the mesh vertices. To compute the 2D locations of such points we rely on the assumption that the mesh facets are sufficiently small such that they remain flat as the surface deforms and are not strongly affected by the perspective effect. Equivalently, this means that the barycentric coordinates of a 2D point with respect to the facet it belongs to remain constant as the mesh deforms. More formally, let $\mathbf{b}_n = [b_{n,1}, b_{n,2}, b_{n,3}]^T$ be the barycentric coordinates of a point n on the mesh, which can be obtained from the first frame of the sequence. The 2D location $\mathbf{x}_n^f$ of this point in frame f can be expressed as

$$\mathbf{x}_n^f = \mathbf{x}_n^1 + b_{n,1}\mathbf{u}_{i(n,1)}^f + b_{n,2}\mathbf{u}_{i(n,2)}^f + b_{n,3}\mathbf{u}_{i(n,3)}^f \ , \tag{1}$$

where $\mathbf{x}_n^1$ is the known location of the point in the first frame, and where the notation $i(n,j)$ indicates that the actual index of the mesh vertex to take into account depends on the point n and on which barycentric coordinate is considered. The two terms in our image-based energy are then defined as follows.

Brightness constancy. Our first image-based term relies on the intuition that the intensity under a mesh point should remain constant as the mesh deforms. Given a set of N_s points densely sampled on the mesh surface with known barycentric coordinates, our brightness constancy term can be written as

$$E_b(\mathbf{U}) = \frac{1}{F-1}\sum_{f=2}^{F}\frac{1}{N_s}\sum_{n=1}^{N_s} g(\mathbf{x}_n^1)[I^f(\mathbf{x}_n^f) - I^1(\mathbf{x}_n^1)]^2 \ , \tag{2}$$

where $\mathbf{U} = \{\mathbf{u}_i^f\}$ is the set of all unknown displacements, and where the dependency on the $\mathbf{u}_i^f$s is not explicitly written for ease of notation, but arises via the $\mathbf{x}_n^f$s, which are computed from Eq. 1. $I^1(\cdot)$ and $I^f(\cdot)$ denote the first and f^{th} images respectively, and $g(\cdot)$ is a gradient-based weighting function. This function is expressed as

$$g(\mathbf{x}) = \exp(\|\nabla I^1(x)\|_2^2) - 1 \ , \tag{3}$$

where ∇ denotes the image gradient after Gaussian smoothing. This weighting function encodes the intuition that pixels with small gradient magnitude are less reliable, and thus makes this term more robust to illumination changes.

Image edges. To further account for the image edges, which, while sparse in the poorly-textured case provide more reliable information, we follow the idea employed in [1]. More precisely, we rely on the distance transform D^f of the edge image obtained from I^f using Canny's algorithm. D^f encodes the distance of each point in frame f to the closest edge point. Given N_e edge points sampled on the edges of the first frame, we formulate our edge-based energy as

$$E_e(\mathbf{U}) = \frac{1}{F-1}\sum_{f=2}^{F}\frac{1}{N_e}\sum_{n=1}^{N_e} D^f(\mathbf{x}_n^f) \ . \tag{4}$$

Altogether, our image-based energy can thus be expressed as

$$E_I(\mathbf{U}) = E_b(\mathbf{U}) + w_e E_e(\mathbf{U}) \;, \tag{5}$$

where w_e is a weight that sets the relative influence of these two terms. This weight was set as 6 in all our experiments.

3.2 Shape Priors

Relying on image information only is known to leave many ambiguities in non-rigid shape reconstruction. Over the years, many priors have been studied in the literature. Here, we make use of three such priors: Isometry, spatial smoothness and temporal smoothness. These three terms are discussed below.

Isometry. In the context of template-based reconstruction, isometry has proven to provide a very reliable prior. This prior encodes the fact that the distance between two neighboring 3D points should not vary, or vary minimally, as the surface deforms. Here, we encourage this for every edge in our mesh. Since this prior is defined in 3D, we need to compute the 3D locations $\mathbf{v}_i^f$ of mesh vertex i in frame f using our parametrization. To this end, let $\tilde{\mathbf{u}}_i^f = [(\mathbf{x}_i^1 + \mathbf{u}_i^f)^T, 1]^T$ be the 2D location of vertex i in frame f in homogenous coordinates, where, with a slight abuse of notation, $\mathbf{x}_i^1$ denotes the 2D location of vertex i in the first frame. The 3D location of vertex i in frame f can then be obtained as

$$\mathbf{v}_i^f = d_i^f \frac{\mathbf{K}^{-1}\tilde{\mathbf{u}}_i^f}{\|\mathbf{K}^{-1}\tilde{\mathbf{u}}_i^f\|_2} \;. \tag{6}$$

Given the set of mesh edges $\mathcal{E}$, this lets us write our isometry prior as

$$E_d(\mathbf{U}, \mathbf{D}, \mathbf{L}) = \frac{1}{F}\sum_{f=1}^{F} \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left(\|\mathbf{v}_i^f - \mathbf{v}_j^f\|_2 - l_{i,j} \right)^2 \;, \tag{7}$$

where $\mathbf{D} = \{d_i^f\}$ is the set of all unknown depths, and, since we do not have a shape template, the true lengths $\mathbf{L} = \{l_{i,j}\}$ are unknown, and thus act as auxiliary variables to be determined by our algorithm. Note that, since the 2D locations of the vertices in the first frame are known, the corresponding displacements $\mathbf{u}_i^1$ are set to 0. The depth d_i^1 of these vertices, however, still needs to be determined by our approach.

Spatial Smoothness. In addition to isometry, we also rely on the intuition that the shape of the surface remains relatively smooth as it deforms, and thus the parameters of neighboring vertices should remain close to each other. In practice, we make use of both first- and second-order smoothness terms. Since our initial mesh is defined as a regular grid, we can define these terms along its vertical and horizontal edges. Let $\mathcal{E}'$ be the set of such edges and $\mathcal{T}$ the set of triplets of aligned vertices. We express spatial smoothness as

$$E_s(\mathbf{U},\mathbf{D}) = \frac{1}{F}\sum_{f=1}^{F}\left(\frac{1}{|\mathcal{E}'|}\sum_{(i,j)\in\mathcal{E}'}\frac{\|\mathbf{p}_i^f - \mathbf{p}_j^f\|_2^2}{(m_{i,j}^1)^2} + \frac{1}{|\mathcal{T}|}\sum_{(i,j,k)\in\mathcal{T}}\frac{\|\mathbf{p}_i^f - 2\mathbf{p}_j^f + \mathbf{p}_k^f\|_2^2}{(m_{i,j}^1)^2}\right) , \tag{8}$$

where $\mathbf{p}_i^f = [(\mathbf{u}_i^f)^T, d_i^f]^T$ denotes the vector of parameters for vertex i in frame f and $m_{i,j}^1$ is the distance between the mesh vertices i and j in the first frame.

Temporal Smoothness. Finally, since we use a video sequence as input, we model the natural intuition that sudden changes in our parameters are unlikely to occur between neighboring frames. This can be expressed by the energy

$$E_t(\mathbf{U},\mathbf{D}) = \frac{1}{F-1}\sum_{f=2}^{F}\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{p}_i^{f-1} - \mathbf{p}_i^f\|_2^2 , \tag{9}$$

where N is the total number of vertices in the mesh.

Altogether, our shape priors can be grouped in an energy of the form

$$E_S(\mathbf{U},\mathbf{D},\mathbf{L}) = w_d E_d(\mathbf{U},\mathbf{D},\mathbf{L}) + w_s E_s(\mathbf{U},\mathbf{D}) + w_t E_t(\mathbf{U},\mathbf{D}) , \tag{10}$$

where w_d, w_s and w_t are weights to set the relative influence of each term. These weights were set to 1.2, 0.05 and 0.05, respectively, in all our experiments.

3.3 Optimization Method

Based on the different energy terms derived above, we formulate 3D reconstruction as the solution to the optimization problem

$$\underset{\mathbf{U},\mathbf{D},\mathbf{L}}{\text{minimize}} \quad E_I(\mathbf{U}) + E_S(\mathbf{U},\mathbf{D},\mathbf{L}) . \tag{11}$$

Since this is a non-convex problem, we make use of a fusion moves strategy to optimize it, which has proven more effective than gradient-based optimization in practice [32, 33]. Note that, since we do not have any template, our formulation still suffers from one global scale ambiguity. Indeed, a larger surface observed further away from the camera will yield exactly the same images. To overcome this issue, we fix the depth of one vertex (the bottom-left corner of the mesh) to an arbitrary value (in practice, the focal length). Below, we explain the fusion moves procedure and our approach to generating good proposals.

Fusion moves. Fusion moves [32, 33] is an optimization technique for graphical models, which can handle continuous variables by iteratively solving a discrete problem. In the context of deformable surfaces, this approach was employed by [19] to address the well-textured template-based and template-free scenarios. Cast to our problem, the fusion moves algorithm works in the following manner: Given the current solution at iteration t, $(\mathbf{U},\mathbf{D},\mathbf{L})^t$, and a proposal, $(\mathbf{U},\mathbf{D},\mathbf{L})^p$, the fusion moves algorithm combines the solution and proposal into a new solution $(\mathbf{U}, \mathbf{D}, \mathbf{L})^{t+1}$ while ensuring that the energy of the new solution is at least as low as that of both the current solution and the proposal. This is achieved by translating the optimization problem to a binary problem with one boolean variable per original variable. In our case, since our energy includes terms that involve triplets of variables (i.e., the image terms and the isometry term), we rely on the order reduction method of [34] to convert it into a purely pairwise energy. We then use QPBO [35] to solve the resulting binary problem at each iteration. For more details of fusion moves, we refer the reader to [32, 33].

Proposal generation. The quality of the solution produced by fusion moves crucially depends on having a good strategy to generate new proposals. Here, we therefore introduce an approach to generating such proposals in our template-free, poorly-textured reconstruction context. Our strategy relies on the two steps discussed below.

First, given the current solution $(\mathbf{U}, \mathbf{D}, \mathbf{L})^t$, we update the image displacements. To this end, we follow a tracking approach that leverages the fact that the 2D vertex locations in the first frame are known. Specifically, we proceed frame-by-frame, starting from frame 2, and minimize our energy with respect to the 2D displacements of each frame in turn. In practice, we employ a Levenberg-Marquardt algorithm initialized using the current solution $\mathbf{U}^t$. The results of this procedure for frame f are then used in the energy minimized for frame $f+1$.

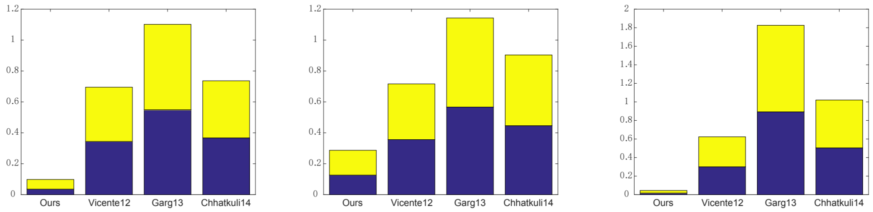
Given the updated image displacements $\mathbf{U}^p$, the depths for the proposal are generated as

$$\mathbf{D}^p = \underset{\mathbf{D}}{\operatorname{argmin}}\, E_S(\mathbf{D}|\mathbf{U}^p, \mathbf{L}^t)\ . \tag{12}$$

To solve this problem, we adopt the gradient descent strategy proposed in [36], and initialize the depths as the upper bound defined in [4]. To account for the global scale ambiguity mentioned above, we rescale the resulting depths, such that the vertex chosen to have fixed depth has the correct depth value.

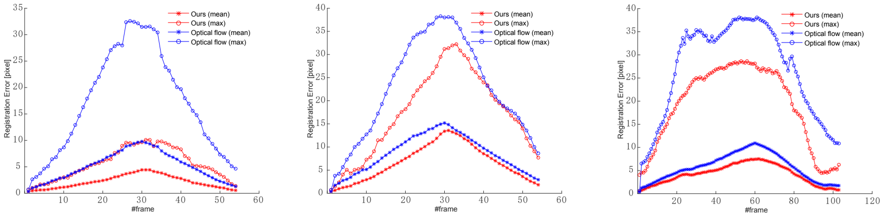
These two steps define the mesh variables, i.e., 2D displacements and depths. We then need to compute proposals for the length variables. To this end, for each edge, we alternate between using its median length in the mesh proposals and its lengths in the first frame. Note that, since the projection in the first frame is given, the corresponding lengths typically provide a good estimate.

Initialization. Before starting our fusion moves procedure, we rely on the following initialization strategy. We initialize our approach by first tracking the surface in 2D using a low-resolution mesh. We then obtain the high-resolution 2D vertex locations from the barycentric coordinates of the missing vertices w.r.t. the low-resolution mesh. Finally, we initialize the depth of all the vertices to the value of the focal length. In supplementary material, we compare this initialization with our final results.



(a) (b) (c)

Fig. 2. Reconstruction error for the synthetic data. We plot the mean and max errors of our approach and of the baselines. The different plots correspond to (a) the cardboard data, (b) the cardboard data with minimal texture and (c) the cloth data. Note that our approach outperforms the baselines significantly.



(a) (b) (c)

Fig. 3. Registration error for the synthetic data. We compare our 2D results with those obtained with the registration method of [31]. The different plots correspond to (a) the cardboard data, (b) the cardboard data with minimal texture and (c) the cloth data.

4 Experiments

To demonstrate the effectiveness of our approach, we evaluate it on six datasets containing images of poorly-textured surfaces deforming in front of a camera. In particular, we make use of three synthetic sequences to provide a quantitative evaluation, and of three real sequences that further evidence that our method generalizes to diverse real surfaces.

In all our experiments, we compare the results of our approach to the following baselines:

- **Vicente12:** This method corresponds to the template-free approach of [19], which leverages isometry in an NRSfM context.
- **Garg13:** This method was introduced in [21] and relies on a total variation regularization within a dense NRSfM framework based on the optical flow approach of [31].
- **Chhatkuli14:** This method was introduced in [20] and performs template-free 3D shape reconstruction by relying on isometry and on an infinitesimal planarity assumption.

For the baselines Vicente12 and Chhatkuli14, we used the publicly available implementations of the authors. For Garg13, since the code is not available, we re-implemented the method ourselves. To validate our implementation, we verified that we could obtain similar results as those published in [21]. Note that Vicente12 and Chhatkuli14 use correspondences obtained from image warps [20, 19]. Since we rely on video as input, we estimated these image warps using tracked feature points, which proved more reliable than just matching them without taking temporal information into account. Note that the 2D registration method employed by Garg13 already accounts for temporal smoothness.

In the remainder of this section, we first discuss our results on the synthetic data, and then move on to the real images. The video sequences of our results are provided as supplementary material.

4.1 Results on Synthetic Data

To perform our quantitative evaluation, we employed the motion capture data publicly available at [37]. This data contains two different surfaces; a piece of cardboard and a piece of cloth. It was acquired by sticking reflective markers on real surfaces in a 9 × 9 grid and deforming the surfaces in front of 6 infrared cameras. The data is provided as 3D triangular meshes. We therefore textured these meshes using poorly-textured images, and rendered the resulting surfaces using a virtual camera. This resulted in images such as those shown in the top row of Figs. 4, 5 and 6.

In our experiments, we report the 3D reconstruction error, computed as the mean point-to-point distance between the ground-truth meshes and the reconstructed ones for each frame in the sequences. To account for the global scale ambiguity that all the evaluated methods are subject to, we first re-scaled all the meshes to a fixed global scale. Note that the baselines yield dense 3D reconstruction. Therefore, we know the 3D locations of the pixels corresponding to the mesh vertices, and can thus use them to estimate this error. In addition to these 3D errors, we also compare the accuracy of our estimated 2D displacements with the registration method of [31] used in Garg13.

Figs. 2 and 3 provide the 3D reconstruction and 2D registration errors for the cardboard and cloth cases, respectively. Note that our approach yields significantly better results than the baselines. As illustrated by Figs 4 , 5 and 6, where we visualize some reconstructions, this can be attributed to the lack of reliable texture information, which, as expected, affects the feature-based methods (Vicente12 and Chhatkuli13), but maybe more surprisingly, also has a negative impact on the optical-flow-based Garg13 method. By contrast, our method is much more robust to this phenomenon. This strength of our approach over the baselines is even further evidenced by the extreme case depicted in Fig 4, where only a single black square acts as texture on the surface, Note that even in this extreme case, as shown in Figs. 2 and 4, our reconstructions remain of good quality. Note that reconstructing a well-textured version of the cardboard data gives the errors: (1.1423, 5.7545, 0.0216, 0.0591) (2D mean error, 2D max error,

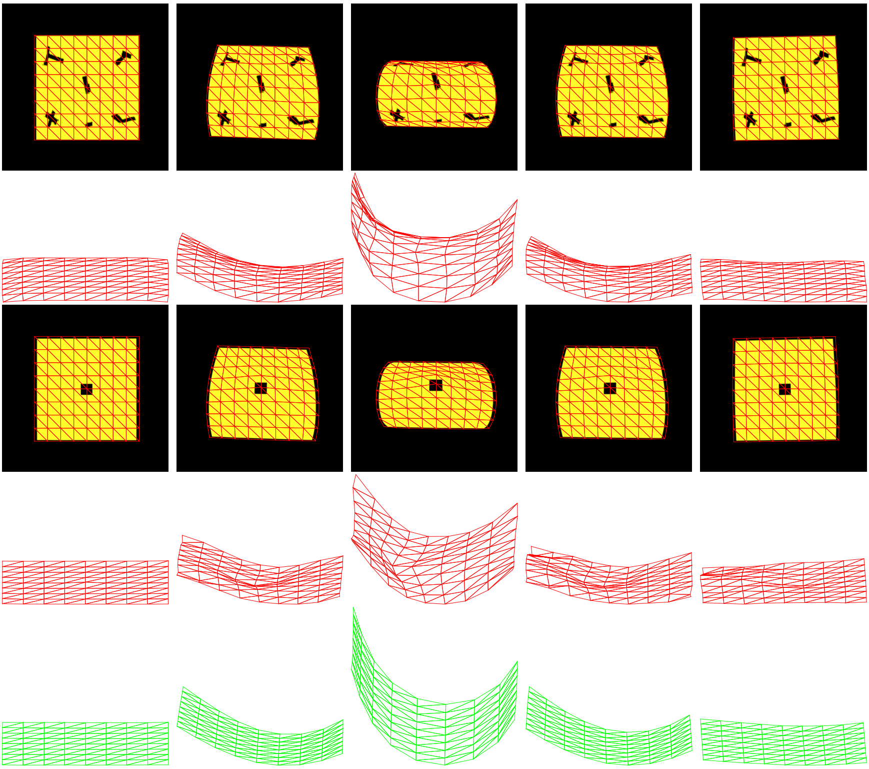


Fig. 4. Reconstructing a piece of cardboard. We show results obtained with two different textures, including one truly minimally-textured case. In each case, we show the input image with our reconstruction reprojected on it and a side view of our reconstruction. The bottom row depicts the ground-truth surfaces. Note that our method yields accurate reconstruction, even in the minimal texture case.

3D mean error, 3D max error). This shows that texture can further help our method.

Regarding efficiency, the runtimes (in sec) of the methods on the synthetic cardboard data (54 frames) are: Our approach: 381.77; Vicente12: 209.73; Garg13: 851.8 (optical flow) + 1189.9 (reconstruction); Chhatkuli14: 361.494 (image warping) + 1458.462 (reconstruction). These results should be taken with a grain of salt: 1) The reconstruction of Garg13 could be, but was not, parallelized. 2) For Chhatkuli14, reconstruction in each frame was done as in their paper by using the image warps of all the other frames, and not using the alternative mentioned by the authors consisting of reconstructing a single frame and using it as template for SfT. Nevertheless, the conclusions would remain the same: Our method is slightly slower than Vicente12 and faster than the other baselines. In supplementary material, we evaluate potential failure cases of our approach.

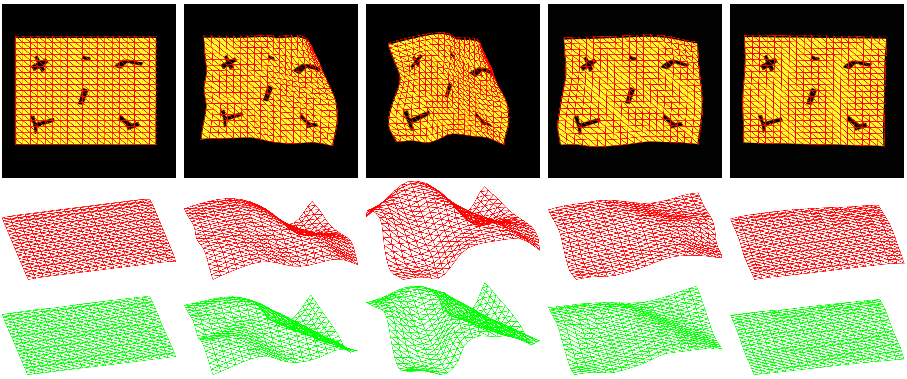


Fig. 5. Reconstructing a piece of cloth. We show the input image with our reconstruction reprojected on it and a side view of our reconstruction. The bottom row depicts the ground-truth surfaces. Note that our method yields accurate reconstruction

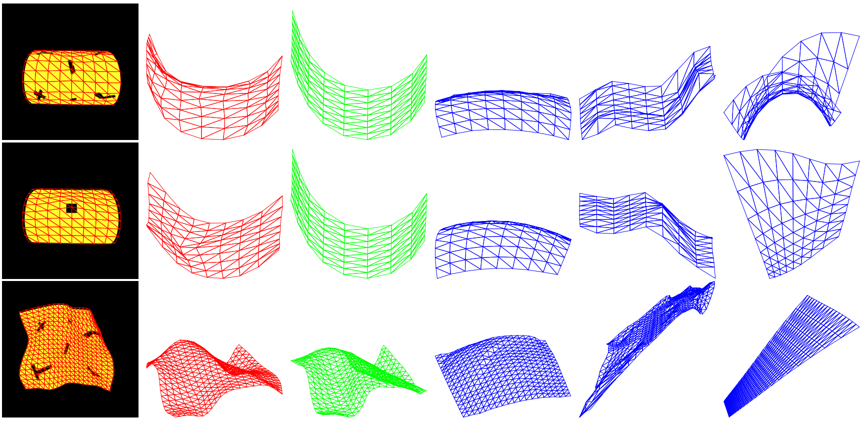


Fig. 6. Comparison to the baselines. From left to right: Input image with our reprojected reconstruction, our reconstruction, ground-truth, Vicente12, Garg13, Chhatkuli14. Note that our reconstructions are clearly more accurate than those of the baselines.

4.2 Results on Real Images

To show that our approach generalizes to real images and to diverse surfaces, we further evaluated it on three real sequences. The first two are the cardboard and cloth sequences used in [5] and publicly available at [37]. The third one depicts a deforming cap, and thus serves to show that our method applies to non-developable surfaces, and, importantly, surfaces that are not planar in the first frame.

The results of our method and the baselines on these three sequences are presented in Fig. 7, respectively. In all cases, we show the input image, the

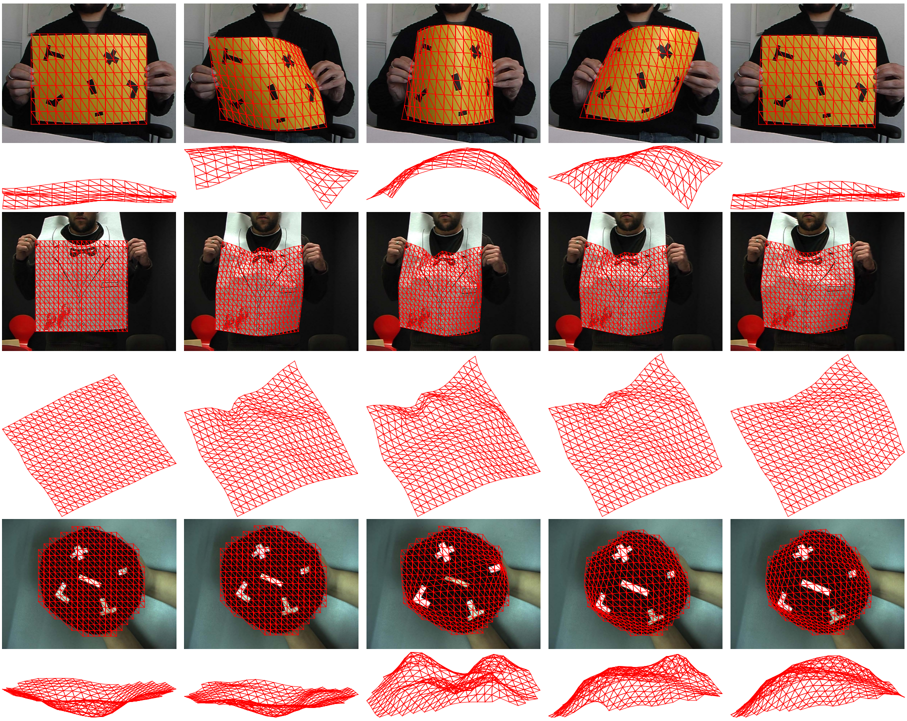


Fig. 7. Reconstructing the cardboard, napkin and cap. we show the input image with our reconstruction reprojected on it and a side view of our reconstruction. Our method yields accurate reconstruction

reconstructed surface reprojected on this image, and the reconstructed surface seen from another viewpoint. These results clearly evidence that our approach yields much more accurate results than the baselines. In particular, the baselines all suffer from the lack of texture and the illumination changes, which make the feature-matching, or 2D registration, step unreliable. By contrast, our approach that jointly performs 2D registration and 3D reconstruction essentially regularizes the matching problem with 3D constraints, and thus yields more accurate results, in terms of both 3D and 2D. From Fig. 7, note that our approach also yields accurate reconstructions when the surface is non-developable, which illustrates its generality.

5 Conclusion

We have introduced an approach to reconstructing a poorly-textured deformable surface from a monocular video sequence and without any template of the surface of interest. Our approach lies at the boundary between template-based reconstruction and NRSfM, in the sense that it leverages priors and image terms that

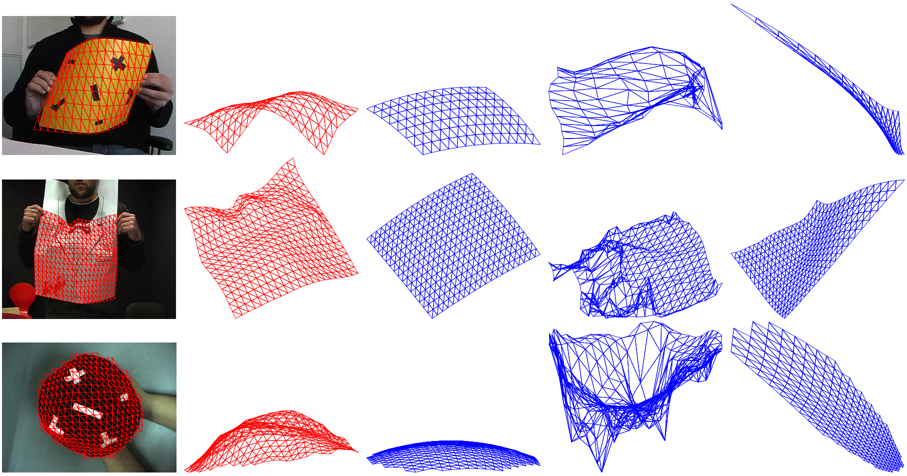


Fig. 8. Comparison to the baselines. From left to right: Input image with our reprojected reconstruction, our reconstruction, Vicente12, Garg13, Chhatkuli14. Note that our reconstructions are clearly more accurate than those of the baselines.

have proven effective in the template-based case, but, as NRSfM techniques, works in the template-free regime. In particular, we have formulated reconstruction as the problem of jointly optimizing the 2D image displacements of mesh vertices and the depth of these vertices, and have proposed a fusion moves strategy to optimize the resulting problem. Our experiments have demonstrated the effectiveness of our approach, and have shown that it yields much higher accuracy than existing template-free techniques. To the best of knowledge, this constitutes the first attempt at solving the challenging template-free and poorly-textured scenario. In the future, we intend to study solutions to address the case where the first frame depicts large, complex deformations, which our current approach remains ill-suited to handle. To this end, we will focus on automatically finding the frame with the smallest deformation, either for the entire surface, or individually for surface patches.

Acknowledgement: This work was supported in part by Natural Science Foundation of China (No.61231018, No.61273366), National Science and Technology Support Program (2015BAH31F01) and Program of Introducing Talents of Discipline to University under grant B13043. Part of this work was performed while X. Wang and M. Salzmann were respectively visiting and affiliated with NICTA, Canberra.

References

1. Salzmann, M., Urtasun, R.: Beyond feature points: Structured prediction for monocular non-rigid 3d reconstruction. In: ECCV. (2012)
2. Ngo, T.D., Park, S., Jorstad, A.A., Crivellaro, A., Yoo, C., Fua, P.: Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture. In: ICCV. (2015)
3. Chhatkuli, A., Pizarro, D., Bartoli, A.: Stable template-based isometric 3d reconstruction in all imaging conditions by linear least-squares. In: CVPR. (2014)
4. Perriollat, M., Hartley, R., Bartoli, A.: Monocular template-based reconstruction of inextensible surfaces. IJCV (2011)
5. Salzmann, M., Urtasun, R., Fua, P.: Local deformation models for monocular 3d shape recovery. In: CVPR. (2008)
6. Bartoli, A., Gerard, Y., Chadebecq, F., Collins, T.: On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In: CVPR. (2012)
7. Bartoli, A., Collins, T.: Template-based isometric deformable 3d reconstruction with sampling-based focal length self-calibration. In: CVPR. (2013)
8. Bartoli, A., Pizarro, D., Collins, T.: A robust analytical solution to isometric shape-from-template with focal length calibration. In: ICCV. (2013)
9. Salzmann, M., Fua, P.: Reconstructing sharply folding surfaces: A convex formulation. In: CVPR. (2009)
10. Salzmann, M., Urtasun, R.: Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In: CVPR. (2010)
11. Yu, R., Russell, C., Campbell, N., Agapito, L.: Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In: ICCV. (2015)
12. Malti, A., Bartoli, A., Hartley, R.: A linear least-squares solution to elastic shape-from-template. In: CVPR. (2015)
13. Ngo, T.D., Östlund, J., Fua, P.: Template-based monocular 3d shape recovery using laplacian meshes. TPAMI (2016)
14. Russell, C., Yu, R., Agapito, L.: Video-popup: Monocular 3d reconstruction of dynamic scenes. In: ECCV. (2014)
15. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: CVPR. (2000)
16. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid structure from motion in trajectory space. In: NIPS. (2008)
17. Taylor, J., Jepson, A., Kutulakos, K.: Non-rigid structure from locally-rigid motion. In: CVPR. (2010)
18. Dai, Y., Li, H., He, M.: A simple prior-free method for nonrigid structure from motion factorization. In: CVPR. (2012)
19. Vicente, S., Agapito, L.: Soft inextensibility constraints for template-free non-rigid reconstruction. In: ECCV. (2012)
20. Chhatkuliand, A., Pizarro, D., Bartoli, A.: Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In: BMVC. (2014)
21. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: CVPR. (2013)
22. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. IJCV (1988)
23. Fua, P., Leclerc, Y.G.: Object-centered surface reconstruction: Combining multi-image stereo and shading. IJCV (1995)

24. Greminger, M., Nelson, B.: Deformable object tracking using the boundary element method. In: CVPR. (2003)
25. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. In: ECCV. (1998)
26. Matthews, I., Baker, S.: Active appearance models revisited. IJCV (2004)
27. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH. (1999)
28. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2d+3d active appearance models. In: CVPR. (2004)
29. Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating faces in images and video. In: Eurographics. (2003)
30. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. IJCV (1992)
31. Garg, R., Roussos, A., Agapito, L.: A variational approach to video registration with subspace constraints. IJCV (2013)
32. Woodford, O., Torr, P., Reid, I., Fitzgibbon, A.: Global stereo reconstruction under second order smoothness priors
33. Lempitsky, V., Rother, C., Roth, S., Blake, A.: Fusion moves for markov random field optimization. TPAMI (2010)
34. Ishikawa, H.: Higher-order clique reduction in binary graph cut. In: CVPR. (2009)
35. Rother, C., Kolmogorov, V., Lempitsky, V., Szummer, M.: Optimizing binary mrfs via extended roof duality. In: CVPR. (2007)
36. Ishikawa, H.: Higher-order gradient descent by fusion-move graph cut. In: ICCV. (2009)
37. http://cvlab.epfl.ch/data/dsr