Global Model with Local Interpretation for Dynamic Shape Reconstruction

Antonio Agudo Francesc Moreno-Noguer
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), 08028, Barcelona, Spain

Abstract

The most standard approach to resolve the inherent ambiguities of the non-rigid structure from motion problem is using low-rank models that approximate deforming shapes by a linear combination of rigid basis. These models are typically global, i.e., each shape basis contributes equally to all points of the surface. While this approach has been shown effective to represent smooth deformations, its performance degrades for surfaces composed of various regions, each following a different deformation rule. Piecewise methods attempt to capture this type of behavior by locally modeling surface patches, although they subsequently require enforcing global constraints to assemble back the patches. In this paper we propose an approach that combines the best of global and local models: it locally considers low-rank models but, by construction, does not need to impose global constraints to guarantee local patch continuity. We achieve this by a simple expectation maximization strategy that besides learning global shape bases, it locally adapts their contribution to each specific surface region. Furthermore, as a side contribution, in order to split the surface into different local patches, we propose a novel physically-based mesh segmentation approach that obeys an energy criterion. The complete framework is evaluated in both synthetic and real datasets, and shows an improved performance to competing methods.

1. Introduction

Simultaneously estimating non-rigid 3D shape and camera motion from a monocular image sequence, i.e., the Non-Rigid Structure from Motion (NRSfM) problem, is severely ill-posed because many different 3D shapes can produce very similar image observations. The problem becomes even more challenging when input data is corrupted by artifacts such as noise or missing data. Over the past decade, a wide body of research has been proposed to tackle these complex situations [6, 22, 26, 39]. At the core of most these methods, lies the assumption that objects do not arbitrarily change their shapes, and that their deformations can be ruled by low-rank models. Among them, the more widely used are the low-rank shape models [10, 31], in which the 3D shape is spanned by a linear combination of rigid and global basis weighted by time-varying coefficients. This has raised a number approaches for sparse [17, 26], dense [20] and even sequential [1, 30] reconstruction.

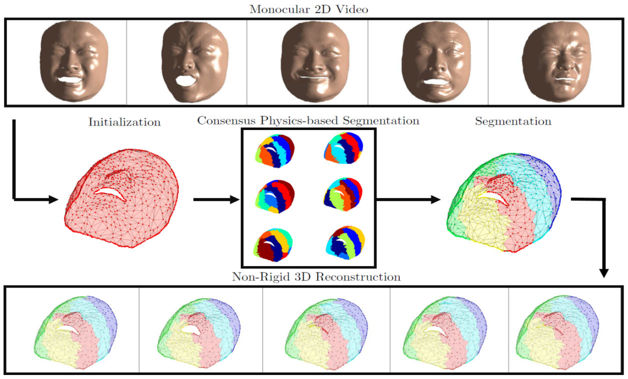
Monocular 2D Video
Initialization
Consensus Physics-based Segmentation
Segmentation
Non-Rigid 3D Reconstruction

Figure 1. **Overview of our approach for non-rigid reconstruction**. The input to our algorithm is a series of 2D point tracks in a monocular video (top). We initially compute a mean shape using rigid factorization (mid-left). We then apply a new physics-based segmentation algorithm to split this shape into regions that have a similar energy pattern (mid-left,right). These regions are fed into our novel strategy that models the shape globally but allows to locally interpret each of the segmented regions. The outcome of our algorithm is the 3D shape in all video frames (bottom) and the camera poses. Note how the results show how our approach captures correctly the local deformations produced in the mouth and eyes. The figure is best viewed in color.

While effective, global models are constrained to shapes that exhibit a homogeneous physical behavior, and are prone to fail for surfaces composed of different local deformations, like those we can find in a biological structure with different tissues, or in a human face expressing emotions, where the deformation is mostly focused on the mouth and the eyes area. There have been several attempts at tackling this by means of piecewise methods [16, 19, 33, 37] that split the shape into small overlapping patches or segments and independently model each of them. However, these approaches either require local rigidity constraints and become limited to isometric deformations [16, 37], or need to solve an optimization problem per each patch and subsequently relying on post-processing steps to assemble back all patches and enforce global consistency [19, 33].

In this paper we propose a novel low-rank global model with local interpretation that allows the fitting of models to small regions, while still retaining the global consistency of the shape without the need of post-processing operations nor requiring overlapping of neighboring regions, solving a single optimization problem. In addition, since no inextensibility constraints are imposed, our method can handle non-isometric motions. Like in global low-rank modeling, we approximate surface deformation by a linear combination of rigid shape basis. However, our main novelty is to specifically weight the contribution of each basis for each segment, that is, instead of using a single time-varying weight per shape basis, we use as many weights as patches or regions the surface is made of. By doing this, we are able to learn the more specific deformation patterns each region may undergo. Although our approach introduces additional unknowns –the per-basis weights– to the global NRSfM problem, we can learn them along with the shape basis and the camera pose parameters using a probabilistic expectation-maximization framework, similar to the one applied in [5, 39] for learning different deformation models.

An additional contribution of our work is a physically-inspired technique to perform mesh segmentation, which we use to generate the local regions that are fed to the NRSfM algorithm. This segmentation holds on a modal analysis performed to the mean shape and splits the surface into regions with similar energy patterns. An schematic of the overall approach, and how the segmentation and reconstruction algorithms are combined is depicted in Fig. 1. Quantitative results on synthetic data, as well as qualitative results on real video sequences, will show the advantages of our approach.

2. Related Work

Reconstructing a time-varying 3D surface while estimating camera pose from solely the observation of 2D point trajectories, is a severely under-constrained problem that requires additional prior information. The most standard prior consists in constraining the surface to lie on a global low dimensional shape space [24, 41]. These early approaches built upon the well-known closed-form factorization method for rigid reconstruction [38]. Later, several iterative methods were proposed to recover the shape and pose parameters [18, 26, 31, 39], which, on top of the low-rank constraints, incorporated temporal and spatial smoothness on the shape. Another way to enforce temporal smoothness is through differentials over the 3D shape matrix by directly minimizing its rank [17, 20].

While global methods have been extensively used in the literature, they may perform poorly when parts of the object obey different deformation rules. To address this problem two new families of solutions were proposed, the local trajectory-based models and the piecewise methods. [7] introduced trajectory models through a series of predefined basis of a discrete cosine transform to independently span the trajectory of each object point. Later, priors on trajectories were incorporated using 3D point differentials [40] and further combined with the global shape model in [23]. Alternatively, piecewise solutions split the surface into a number of patches and independently model and solve each of them. However, these methods usually require tracking more points than global approaches to locally enforce isometry constraints that help to disambiguate the problem [16, 37]; or on the other hand, to enforce a smooth transition from the local models to a global one, these methods rely on a large number of overlapping patches [19, 33]. This may require being able to match features between neighboring patches, which can be difficult in practice.

Contributions. We propose two main contributions. First, we present a novel solution for non-rigid reconstruction that combines the best of global and local models into a single framework. We resemble local methods in that our approach can locally model surface regions with distinct physical behavior. And, like in global models, we can do this without then having to enforce global continuity, which is inherently guaranteed by our formulation. Furthermore, this is achieved with just a small number of additional parameters compared to existing techniques, which can be learned with standard expectation-maximization. Our second contribution is a new physically-grounded shape signature that holds on a modal analysis decomposition. This signature is used to segment a reference shape into regions with similar energy patterns, which become the local patches for the reconstruction algorithm. The combination of both these contributions shows favorable results compared to state-of-the-art techniques, as we report in the results section.

3. Physically-based Mesh Segmentation

We next present our physically-based approach for mesh segmentation that will be used as input to the reconstruction technique. We first revisit concepts of modal analysis, the physical principle upon which we build a signature per each 3D point of the mesh. This descriptor is then used to generate multiple candidate segmentations, that are fed into a consensus clustering for the final segmentation.

3.1. Revisiting Modal Analysis

Modal Analysis (MA) is a standard technique in structural engineering to reduce the degrees of freedom of a deforming shape by approximating it as a linear combination of modes [11]. In MA, the N nodes of an object can be regarded as physical elements (e.g., spring, triangle or tetrahedral meshes), and assemble their local contributions into global stiffness $\mathbf{K}$ and mass $\mathbf{M}$ matrices. Both matrices

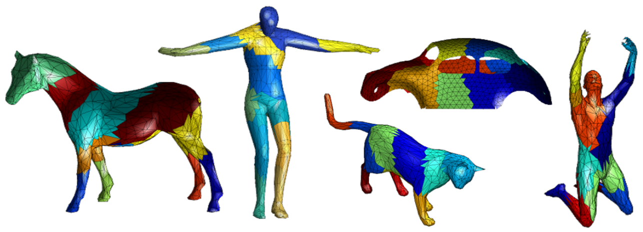

Figure 2. **Sample examples of 3D mesh segmentation in general meshes.** In the results section we show further segmentation examples specifically applied to the images we then feed to the reconstruction algorithm. The figure is best viewed in color.

are computed following [3]. We can then solve equilibrium equations and obtain the undamped free vibration response of the 3D object caused by a disturbance with respect to the shape at rest $\mathbf{s}_0$ based on the following generalized eigenvalue problem:

$$\mathbf{K}\boldsymbol{\Phi} = \mathbf{M}\boldsymbol{\Omega}\boldsymbol{\Phi} \ , \tag{1}$$

where $\boldsymbol{\Phi}{=}[\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_k,\ldots,\boldsymbol{\phi}_{3N}]$ are the $3N$ mode shapes (eigenvectors) and diag$(\boldsymbol{\Omega})$ their frequencies (eigenvalues). Each eigenmode $\boldsymbol{\phi}_k$ is a $3N \times 1$ vector representing the displacements of all N nodes with respect to $\mathbf{s}_0$.

The frequency spectrum pattern can be used to infer physical properties of the object [1, 35]. In particular, [1] used the mode frequency to classify them into bending and stretching types. Within each category, in turn, it was shown that eigenmodes with lower frequencies are mode shapes that require less energy to be excited, and govern global deformation. In contrast the most high frequency modes dominate local deformations. Recall that this principle was successfully used to code the deformations of an object exclusively by observing its mean shape, and we here use these observations to obtain region-based features under deformation and to segment a mesh based on an energy criterion. We next describe the details of the segmentation algorithm.

3.2. Energy-Preserving Segmentation Algorithm

Deformable mesh segmentation is an active research topic in 3D shape analysis. Much of the focus in this area consists in building point signatures robust to non-rigid deformations. These descriptors are then used to perform the actual segmentation in schemes like [21, 25, 32]. The nature of such signatures may be very different (global or local, geometric or topological, volumetric or superficial, intrinsic or extrinsic) [15] and its review is beyond the focus of this paper. Just to name a few, it is worth mentioning descriptors based on geodesic distances or on the Laplace-Beltrami [34] and its variants like the Heat Kernel Signature [14, 36].

Drawing inspiration in these works, given a point $\mathbf{x}$ on the surface, we define its global signature $\mathbf{G}(\mathbf{x}) \in \mathbb{R}^{3\times 2p}$ by taking the value of the first p eigenvectors on both the bending and stretching family shapes, weighted by a function of the corresponding eigenvalue. The $\mathbf{G}(\mathbf{x})$ is:

$$\Bigg[\underbrace{\frac{1}{\sqrt{\omega_{1b}}}\boldsymbol{\phi}_{1b}[\mathbf{x}],\ldots,\frac{1}{\sqrt{\omega_{pb}}}\boldsymbol{\phi}_{pb}[\mathbf{x}]}_{p\text{ bending modes}},\underbrace{\frac{1}{\sqrt{\omega_{1s}}}\boldsymbol{\phi}_{1s}[\mathbf{x}],\ldots,\frac{1}{\sqrt{\omega_{ps}}}\boldsymbol{\phi}_{ps}[\mathbf{x}]}_{p\text{ stretching modes}}\Bigg], \tag{2}$$

where $\boldsymbol{\phi}_k[\mathbf{x}]$ are the 3 components of $\boldsymbol{\phi}_k$ at the point $\mathbf{x}$, $\{\boldsymbol{\phi}_{jb}, \omega_{jb}\}$ are the pairs eigenmode/eigenvalue of the bending deformation and $\{\boldsymbol{\phi}_{js}, \omega_{js}\}$ the corresponding stretching pairs. Lower-energy mode shapes, i.e., those related with global deformations, will have a stronger contribution.

A straightforward approach to perform an energy-based segmentation, could be simply running a standard k-means clustering based on the shape embeddings defined by Eq. (2). However, the results obtained this way turned not to be stable and strongly depended on the initial seeds chosen to initialize the k-means. In order to obtain more stable segmentations, we followed the consensus segmentation approach proposed recently in [21, 32]. The main idea is to first generate an ensemble of segmentations by running the clustering algorithm multiple times, and then computing a consensus segmentation that is as close as possible to all the others (see Fig. 1). More specifically, let $\mathbf{B} \in \mathbb{R}^{B\times N}$ be a representation of the B segmentations, where $\mathbf{B}[b, j]$ is the label of the j-th point in the b-th segmentation. The consensus segmentation $\mathbf{y}$ can then be retrieved by computing the Fréchet sample mean:

$$\arg\min_{\mathbf{y}} \sum_{b=1}^{B} d^2 \left(\mathbf{B}[b, *], \mathbf{y}\right) , \tag{3}$$

where d is a semi-metric to measure the distance between segmentations, and $\mathbf{B}[b, *]$ is the b-th row in $\mathbf{B}$, i.e., the b-th partial segmentation. For further details we refer the reader to [27, 32]. Figure 2 shows segmentation results on some synthetic objects of the benchmark TOSCA [13]. Although our segmentations are not visually symmetric, they produce more accurate solutions compared to competing methods, as we show on real objects in the results section.

4. Global-to-Local Deformation Model

In this section we introduce our global shape model with local interpretations. We start reviewing classical global low-rank shape models, upon which, we will then build our proposed approach.

4.1. Classical Global Low-rank Shape Model

Representing the non-rigid deformation of an object as a linear combination of rigid, global shape bases is a well-known practice. Such a low-rank shape basis has been computed by learning techniques like principal component analysis over a set of training data [12, 29], applying modal [1, 9] or spectral [4] analysis over a rest configuration, or estimated on-line using data-driven methods [20, 26, 31, 39].

Let us consider an object represented by N 3D points and observed in T frames. Let also denote by $\mathbf{x}_n^t = [x_n^t, y_n^t, z_n^t]^\top$ the 3D coordinates of the n-th point at frame t, and by $\mathbf{s}^t = [(\mathbf{x}_1^t)^\top, \dots, (\mathbf{x}_N^t)^\top]^\top$ the $3N$-dimensional representation of the shape. We can approximate the instant shape $\mathbf{s}^t$ by a mean shape $\mathbf{s}_0$ together with a linear combination of R mode shapes $\mathbf{s}_r$, $r \in \{1, \dots, R\}$ where $\mathbf{s}_r = [(\mathbf{s}_{1,r})^\top, \cdots, (\mathbf{s}_{n,r})^\top, \cdots, (\mathbf{s}_{N,r})^\top]^\top$, and $\mathbf{s}_{n,r} = [x_{n,r}, y_{n,r}, z_{n,r}]^\top$ are the coordinates of the n-th point on the r-th shape vector. If we concatenate these modes into a matrix $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_R] \in \mathbb{R}^{3N \times R}$ we can write $\mathbf{s}^t = \mathbf{s}_0 + \mathbf{S}\boldsymbol{\psi}^t$, or equivalently:

$$\begin{bmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_n^t \\ \vdots \\ \mathbf{x}_N^t \end{bmatrix} = \begin{bmatrix} \mathbf{s}_{1,0} \\ \vdots \\ \mathbf{s}_{n,0} \\ \vdots \\ \mathbf{s}_{N,0} \end{bmatrix} + \begin{bmatrix} \mathbf{s}_{1,1} & \dots & \mathbf{s}_{1,r} & \dots & \mathbf{s}_{1,R} \\ \dots & \ddots & \dots & \dots & \dots \\ \mathbf{s}_{n,1} & \dots & \mathbf{s}_{n,r} & \dots & \mathbf{s}_{n,R} \\ \dots & \dots & \dots & \ddots & \dots \\ \mathbf{s}_{N,1} & \dots & \mathbf{s}_{N,r} & \dots & \mathbf{s}_{N,R} \end{bmatrix} \begin{bmatrix} \psi_1^t \\ \vdots \\ \psi_r^t \\ \vdots \\ \psi_R^t \end{bmatrix} \tag{4}$$

where $\boldsymbol{\psi}^t$ is an R-dimensional vector with the time-varying coefficients for the shape at time t.

4.2. Local Interpretation of Global Shape Models

We next describe our strategy to provide the global mode shapes with the ability to adapt locally. Let us assume our shape $\mathbf{s}^t$ to be partitioned into C clusters. Without loss of generality we consider the N points $\mathbf{x}_n^t$, $n \in \{1, \dots, N\}$ to be sorted in such a way that the N_1 first points belong to the first cluster, the next N_2 points belong to the second cluster, and so on, until the C-th cluster[1]. Note that $\sum_{c=1}^{C} N_c = N$, i.e., we do not assume overlapping between the points of neighboring clusters. Our goal is to let global shape basis to adapt differently to each cluster.

To allow the global shape basis $\mathbf{S}$ to adapt locally for each of the clusters, we will follow a simple strategy, where the components $\mathbf{s}_{n,r}$ of $\mathbf{S}$ will be re-arranged into a $3N \times RC$ block diagonal matrix $\tilde{\mathbf{S}}$, and the vector of coefficients $\boldsymbol{\psi}^t$ will be expanded to an RC-dimensional vector $\boldsymbol{\varphi}^t$. In particular, the global model of Eq. (4) will be rewritten as the following global model with local interpretation $\mathbf{s}^t = \mathbf{s}_0 + \tilde{\mathbf{S}}\boldsymbol{\varphi}^t$, or equivalently:

$$\begin{bmatrix} \tilde{\mathbf{x}}_1^t \\ \vdots \\ \tilde{\mathbf{x}}_c^t \\ \vdots \\ \tilde{\mathbf{x}}_C^t \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{s}}_{1,0} \\ \vdots \\ \tilde{\mathbf{s}}_{c,0} \\ \vdots \\ \tilde{\mathbf{s}}_{C,0} \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{S}}_1 & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \ddots & & & \vdots \\ \mathbf{0} & & \tilde{\mathbf{S}}_c & & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \dots & \tilde{\mathbf{S}}_C \end{bmatrix} \begin{bmatrix} \boldsymbol{\varphi}_1^t \\ \vdots \\ \boldsymbol{\varphi}_c^t \\ \vdots \\ \boldsymbol{\varphi}_C^t \end{bmatrix} \tag{5}$$

where $\tilde{\mathbf{x}}_c^t = [(\mathbf{x}_\delta^t)^\top, \dots, (\mathbf{x}_{\delta+N_c}^t)^\top]^\top$ are the N_c 3D points of the c-th cluster, and $\delta = 1 + \sum_{j=1}^{c-1} N_j$ indexes the first 3D points of the cluster (recall that we assumed all points to be sequentially ordered in $\mathbf{s}^t$). The vector $\tilde{\mathbf{s}}_{c,0}$ is formed in a similar manner from the elements of the shape at rest $\mathbf{s}_0$[2]; and $\boldsymbol{\varphi}_c^t = [\varphi_{c,1}^t, \dots, \varphi_{c,R}^t]^\top$ are the cluster specific weights. The block diagonal matrices are formed by simply re-arranging the terms of the original matrix $\mathbf{S}$. For instance $\tilde{\mathbf{S}}_c$ is a $N_c \times R$ matrix with the form:

$$\tilde{\mathbf{S}}_c = \begin{bmatrix} \mathbf{s}_{\delta,1} & \mathbf{s}_{\delta,2} & \dots & \mathbf{s}_{\delta,R} \\ \mathbf{s}_{\delta+1,1} & \mathbf{s}_{\delta+1,2} & \dots & \mathbf{s}_{\delta+1,R} \\ & & \ddots & \\ \dots & \dots & & \dots \\ \mathbf{s}_{\delta+N_c,1} & \mathbf{s}_{\delta+N_c,2} & \dots & \mathbf{s}_{\delta+N_c,R} \end{bmatrix}. \tag{6}$$

Note that with this formulation, each shape cluster can be locally adapted by means of their particular coefficients $\boldsymbol{\varphi}_c$. The global and local shape bases have exactly the same number of non-zero components (see classical global model in Eq. (4)). This is why we claim our new model is a global one, but with local interpretations. The only additional parameters correspond to the vectors of coefficients, which has grown from size R in $\boldsymbol{\psi}$, to size RC in $\boldsymbol{\varphi}$. The new matrix of shape basis $\tilde{\mathbf{S}}$ has $3NR$ non-null parameters, which exactly correspond to those of the original matrix $\mathbf{S}$. Indeed the two matrices can be related by means of an $RC \times R$ *compression matrix* $\mathbf{C}$ as follows:

$$\tilde{\mathbf{S}}\mathbf{C} = \tilde{\mathbf{S}}(\mathbf{1}_C \otimes \mathbf{I}_R) = \mathbf{S}, \tag{7}$$

where $\otimes$ denotes a Kronecker's product, $\mathbf{I}_R$ is the R-dimensional identity matrix and $\mathbf{1}_C$ is a C-dimensional vector of ones. When considering one single cluster, $\mathbf{C} \equiv \mathbf{I}_R$, and our model becomes the classical global one. By construction, the compression matrix $\mathbf{C}$ has rank R, and the inverse mapping cannot be computed. For later computations, we can alternatively represent Eq. (7) as:

$$(\mathbf{C}^\top \otimes \mathbf{I}_{3N})\text{vec}(\tilde{\mathbf{S}}) = \text{vec}(\mathbf{S}), \tag{8}$$

where vec$(\cdot)$ denotes the vectorization operator.

5. Recovering Shape and Motion with the new Deformation Model

We now describe how to introduce our global low-rank shape model with local interpretations into the formulation of the NRSfM problem, to jointly obtain camera motion and non-rigid shape by solving a single optimization problem.

5.1. Problem Formulation

Let us consider an orthographic camera observing a dynamic object which at a time instant t is represented by a

[1] Without this assumption, in the following we should define a permutation function that indexes the pixels of each cluster. But for clarity of presentation, we have declined doing so.

[2] Note that the vectors $\mathbf{s}^t$ and $\mathbf{s}_0$ are exactly the same in Eq. (4) and Eq. (5), respectively. In the latter we have just grouped the components per each of the clusters.

$3N$ vector $\mathbf{s}^t$. We can write the projection of the 3D points onto the image by:

$$\mathbf{p}^t = \mathbf{Q}^t \mathbf{s}^t + \mathbf{n}^t, \tag{9}$$

where $\mathbf{p}^t$ is a $2N$-vector with the projected points, $\mathbf{Q}^t = \mathbf{I}_N \otimes \mathbf{R}^t$ and has size $2N \times 3N$, $\mathbf{R}^t$ are the two upper rows of a full rotation matrix and $\mathbf{n}^t$ is a $2N$-dimensional vector of Gaussian noise. Note that this projection model assumes mean-centered 2D projections, and thus the translation component of the pose is not considered. Our problem consists in, given the observation of temporal point correspondences $\mathbf{p}^t$ corrupted by noise $\mathbf{n}^t$, for $t \in \{1, \dots T\}$, recovering the shape $\mathbf{s}^t$ and camera rotation $\mathbf{R}^t$ for all frames of the sequence. We can introduce the shape model of Eq. (5) into the projection Eq. (9) as:

$$\mathbf{p}^t = \mathbf{Q}^t (\mathbf{s}_0 + \tilde{\mathbf{S}} \boldsymbol{\varphi}^t) + \mathbf{n}^t. \tag{10}$$

5.2. Probabilistic Global Shape Model with Local Interpretations

In order to jointly learn shape and motion, we follow the recent works on probabilistic NRSfM [5, 39]. The overall approach consists in first defining a probabilistic distribution over the observations $\mathbf{p}^t$ on Eq. (10), and then estimating the model parameters that maximize its likelihood function using an EM-based algorithm.

To achieve this, we assume the time-varying coefficients $\boldsymbol{\varphi}^t$ become latent variables and follow a zero-mean Gaussian distribution $\boldsymbol{\varphi}^t \sim \mathcal{N}(\mathbf{0}; \mathbf{I}_{RC})$. If we also assume that the observation noise follows a Gaussian distribution $\mathbf{n}^t \sim \mathcal{N}(\mathbf{0}; \sigma^2 \mathbf{I}_{2N})$, it can be showed that the projected points $\mathbf{p}^t$ are normally distributed:

$$\mathbf{p}^t \sim \mathcal{N}\left(\mathbf{Q}^t \mathbf{s}_0; \mathbf{Q}^t \tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top {\mathbf{Q}^t}^\top + \sigma^2 \mathbf{I}_{2N}\right). \tag{11}$$

Solving the NRSfM problem is then equivalent to learning the parameters of this distribution. We next describe how we do this by performing Maximum Likelihood Estimation (MLE) on this latent variable problem using EM.

5.3. Parameter Learning with EM

The MLE distribution in Eq. (11) is learned using EM in a similar manner as done in [5, 39]. Concretely, given the 2D point trajectories $\mathbf{p} = \{\mathbf{p}^1, \dots, \mathbf{p}^T\}$, we aim at estimating the parameters $\boldsymbol{\Theta} = \{\mathbf{R}^1, \dots, \mathbf{R}^T, \tilde{\mathbf{S}}, \sigma^2\}$, taking the weight coefficients $\boldsymbol{\varphi}^t$ as latent variables. We next detail the specific equations involved in each of the $E-$ and $M-$steps over which the EM algorithm iterates.

E-Step. The first step in EM consists in estimating the posterior distribution over the variables $\boldsymbol{\varphi}^t$ given the observations $\mathbf{p}^t$ and the current model parameters $\boldsymbol{\Theta}$. Assuming i.i.d. observations, it can be shown that this distribution is:

$$p(\boldsymbol{\varphi}^t | \mathbf{p}^t, \mathbf{R}^t, \tilde{\mathbf{S}}, \sigma^2) \sim \mathcal{N}(\boldsymbol{\mu}^t_{\boldsymbol{\varphi}}; \boldsymbol{\Sigma}^t_{\boldsymbol{\varphi}}), \tag{12}$$

with:

$$\begin{aligned} \boldsymbol{\mu}^t_{\boldsymbol{\varphi}} =& \boldsymbol{\Lambda}^t (\mathbf{p}^t - \mathbf{Q}^t \mathbf{s}_0) \ ; \ \ \boldsymbol{\Sigma}^t_{\boldsymbol{\varphi}} = \mathbf{I}_{RC} - \boldsymbol{\Lambda}^t \mathbf{Q}^t \tilde{\mathbf{S}}, \\ \boldsymbol{\Lambda}^t =& \tilde{\mathbf{S}}^\top \left(\mathbf{Q}^t\right)^\top (\sigma^2 \mathbf{I}_{2N} + \mathbf{Q}^t \tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top {\mathbf{Q}^t}^\top)^{-1}. \end{aligned}$$

M-Step. In the M-step, we seek to maximize the likelihood of the observed data (or minimize its negative log-likelihood) with respect to the modal parameters $\boldsymbol{\Theta}$. For this purpose, we replace the latent variables by their expected values, and build the following negative log-likelihood function $\mathcal{L} \equiv \frac{1}{2\sigma^2} \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{p}^t - \mathbf{Q}^t(\mathbf{s}_0 + \tilde{\mathbf{S}}\boldsymbol{\varphi}^t)\|_2^2\right] + NT \log(2\pi\sigma^2)$, where $\mathbb{E}[\cdot]$ denotes the expectation operator. In order to update each model parameter, we compute its corresponding partial derivative assuming the rest of parameters are fixed, set it to zero and solve it. We next detail the update rules we obtain for every parameter.

Updating the Deformation Model $\tilde{\mathbf{S}}$. Note that the likelihood $\mathcal{L}$ is a function of the *global-to-local* matrix $\tilde{\mathbf{S}}$ we have introduced in this paper. However, as seen in Eq. (5), $\tilde{\mathbf{S}}$ is block diagonal and their only non-zero elements are the same as those of the classical *global* shape matrix $\mathbf{S}$ in Eq. (4). We therefore proceed by first estimating the matrix $\mathbf{S}$ to then build the extended version $\tilde{\mathbf{S}}$.

Since the likelihood function $\mathcal{L}$ does not explicitly depend on $\mathbf{S}$, we need to resort to Eq. (8) that maps $\tilde{\mathbf{S}}$ to $\mathbf{S}$ to then solve for $\frac{\partial \mathcal{L}}{\partial \mathbf{S}} = 0$. In this way, we can compute the mapping in closed-form (with no need to invert the matrix $\mathbf{C}$), obtaining finally the following update rule for the shape basis:

$$\begin{aligned} \text{vec}(\mathbf{S}) \leftarrow & (\mathbf{C}^\top \otimes \mathbf{I}) \left(\sum_{t=1}^{T} \left(((\boldsymbol{\Upsilon}^t_{\boldsymbol{\varphi}\boldsymbol{\varphi}})^\top \mathbf{C})^\top \otimes {\mathbf{Q}^t}^\top \mathbf{Q}^t \right) \right)^{-1} \\ & \times \text{vec}\left(\sum_{t=1}^{T} {\mathbf{Q}^t}^\top (\mathbf{p}^t - \mathbf{Q}^t \mathbf{s}_0) \boldsymbol{\mu}^t_{\boldsymbol{\varphi}} \mathbf{C} \right), \end{aligned}$$

where we use the expectation $\boldsymbol{\Upsilon}^t_{\boldsymbol{\varphi}\boldsymbol{\varphi}} = \mathbb{E}\left[\boldsymbol{\varphi}^t (\boldsymbol{\varphi}^t)^\top\right] = \boldsymbol{\Sigma}^t_{\boldsymbol{\varphi}} + \boldsymbol{\mu}^t_{\boldsymbol{\varphi}} (\boldsymbol{\mu}^t_{\boldsymbol{\varphi}})^\top$.

Updating $\mathbf{R}^t$ and σ. The update rules for the rotation matrices and 2D noise parameter can be computed in a more straightforward manner from a direct computation of the partial derivatives $\partial \mathcal{L} / \partial \mathbf{R}^t = 0$ and $\partial \mathcal{L} / \partial \sigma^2 = 0$. For the rotation matrices we just need to ensure that they lie on the smooth manifold defined by the orthogonal group $SO(3)$, which we achieve following the iterative approach proposed in [2]. Model parameters are initialized running a rigid factorization algorithm [28].

5.4. Practical Considerations

We next briefly discuss several details of our approach regarding its ability to handle outliers and the influence of the number of clusters.

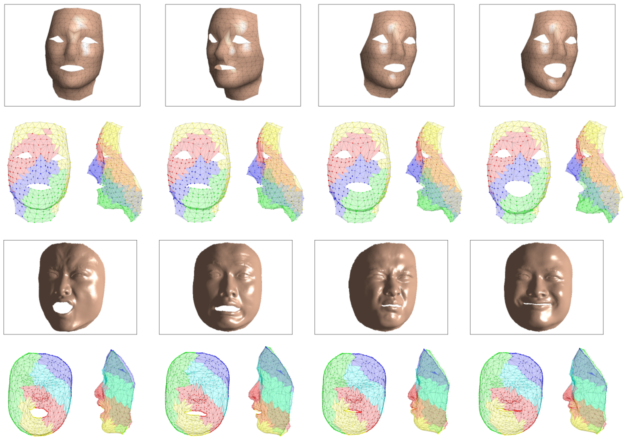

Figure 3. **Synthetic results on the *Face1* and *Expressions* sequences.** For each experiment, we show the input images at the top, and at the bottom a frontal and side views of the reconstructed shapes. For the reconstruction views, we also include colored triangles and points to represent the cluster they belong. We set the number of clusters to $C = 4$ for the *Face1* and to $C = 5$ and for *Expressions*. The rank of the shape model is set to $R = 2$ and $R = 4$, respectively. Best viewed in color.

Dealing with missing data. The EM-based optimization framework we propose, allows to naturally handle missing observations due to occlusions or outliers. To do this, we need to consider the missing observations at initialization and optimization. At initialization, we can obtain the missing entries by imposing smooth trajectories in the image plane, as done in [23]. For the optimization stage, the missing entries $\hat{\mathbf{p}}_i^t$ are updated during the M-step by considering the expected latent values and the model parameters as:

$$\hat{\mathbf{p}}_i^t \leftarrow \mathbf{Q}^t(\mathbf{s}_{i,0} + \tilde{\mathbf{S}}_{i,*}\boldsymbol{\varphi}^t), \tag{13}$$

where $\tilde{\mathbf{S}}_{i,*}$ represents the shape contribution for the i-th point. In the experimental section, we will show our approach performs robustly to large amounts missing data.

Number of clusters. The number C of clusters a shape is segmented is manually chosen. We could have used statistical measures for doing so (e.g., information criteria) but we found it not necessary. In any event, the value of C has direct implications both in the computational cost (the size of the latent variables $\boldsymbol{\varphi}^t$ is RC) and in a lesser extend, in the accuracy. In the results section we will evaluate several choices of the parameter C, which represent good compromises between computation time and accuracy.

6. Experimental Evaluation

We next present quantitative and qualitative results of our approach on a wide variety of objects and types of deformations. These results can be best viewed in the supplemental video. When a quantitative evaluation is reported, we provide the mean 3D reconstruction error as defined in [19, 20].

Data \ Met.	EM-LDS	PTA	CSF2	KSTA	EM-PND	Ours
Expressions	4.42(5)	4.77(2)	3.02(5)	3.46(4)	–	**2.56(3)**
Face1	5.08(2)	3.62(2)	2.10(4)	2.08(3)	12.12	**1.94(2)**
Face2	2.81(2)	2.67(2)	2.50(3)	2.34(3)	4.08	**1.80(2)**

Table 1. **Quantitative comparison on mocap sequences.** We compare our approach against: EM-LDS [39], PTA [7], CSF2 [23], KSTA [22], and EM-PND [26] in terms of $e_{3D}[\%]$. "–" indicates the algorithm did not manage to process the sequence. In all cases we show the results with the number of rank R in the subspace (in brackets) that gave the lowest e_{3D}.

6.1. Synthetic Data

We consider a synthetic benchmark with three sequences annotated with 3D ground truth. Two of the sequences are from [8] and show faces performing simple deformations and gestures. Each sequence consist of 100 frames and 313 points. We denote these datasets as *Face1* and *Face2*. We also use the mocap sequence from [42], which shows a 3D face performing over-exaggerated expressions. Although this data is not originally meant for evaluating NRSfM methods, we process it to generate a sequence of 384 frames and 997 2D point tracks, which we denote it as *Expressions*.

We will compare our approach against the following methods that use low-rank models on both shape and trajectory spaces. Particularly, we consider: EM-LDS [39], and EM-PND [26] for shape space; PTA [7] for trajectory space; the Column Space Fitting (CSF2) [23] and the Kernel Shape Trajectory Approach (KSTA) [22] for shape-trajectory methods[3]. The parameters of these methods were set as suggested in the original papers. For the optimization stage, we only have to set the rank of the subspace R. The number of clusters, is set to $C = 4$ for the *Face1* and *Face2* experiments and to $C = 5$ for the *Expressions* experiment.

The mean 3D reconstruction errors are summarized in Table 1. Observe that our approach consistently outperforms the rest of competing approaches. In Fig. 3 we show some qualitative results including the 2D input data and the reconstructed 3D shape, along with the regions that have been computed by the segmentation algorithm. Note that no discontinuities are observed at the boundaries of neighboring regions, indicating that our approach can naturally enforce global consistency with no need to use specific post-processing operations.

Regarding the computation time, the *Face1* and *Face2* sequences required about 0.88 sec. to be segmented, and about 14.40 sec. to compute the 3D shape. For the *Expressions* sequence the segmentation and reconstruction times were 2.13 sec. and 363.12 sec., respectively. It is worth to point that EM-PND [26], which is acknowledged to be

[3] We also considered using the block matrix approach of [17], but did not manage to make it work for none of the sequences. We presume the number of linear-matrix-inequality constraints this method uses is not sufficient for the proposed sequences.

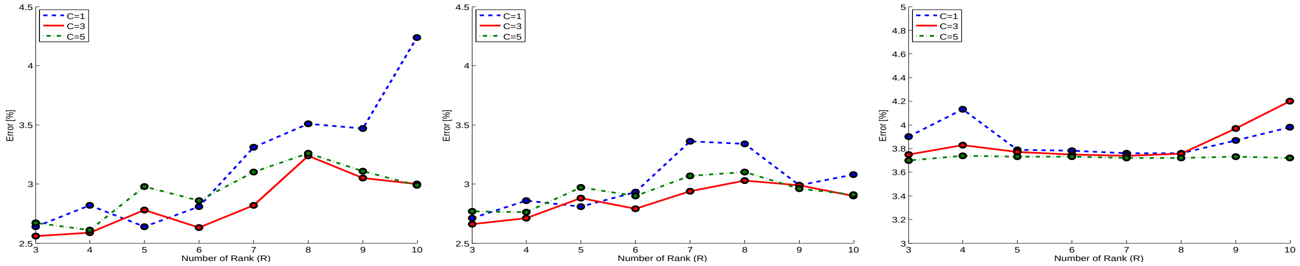

Figure 4. **Mean 3D reconstruction error** in the *Expressions* sequence, as a function of basis rank R and the number of clusters C. **Left:** Noise-less 2D data. **Middle:** The 2D observations have been corrupted with Gaussian noise of standard deviation $\sigma_{noise} = 0.01\kappa$, where κ represents the maximum distance of an image point to the centroid of all the points. **Right:** Robustness to 50% random missing data.

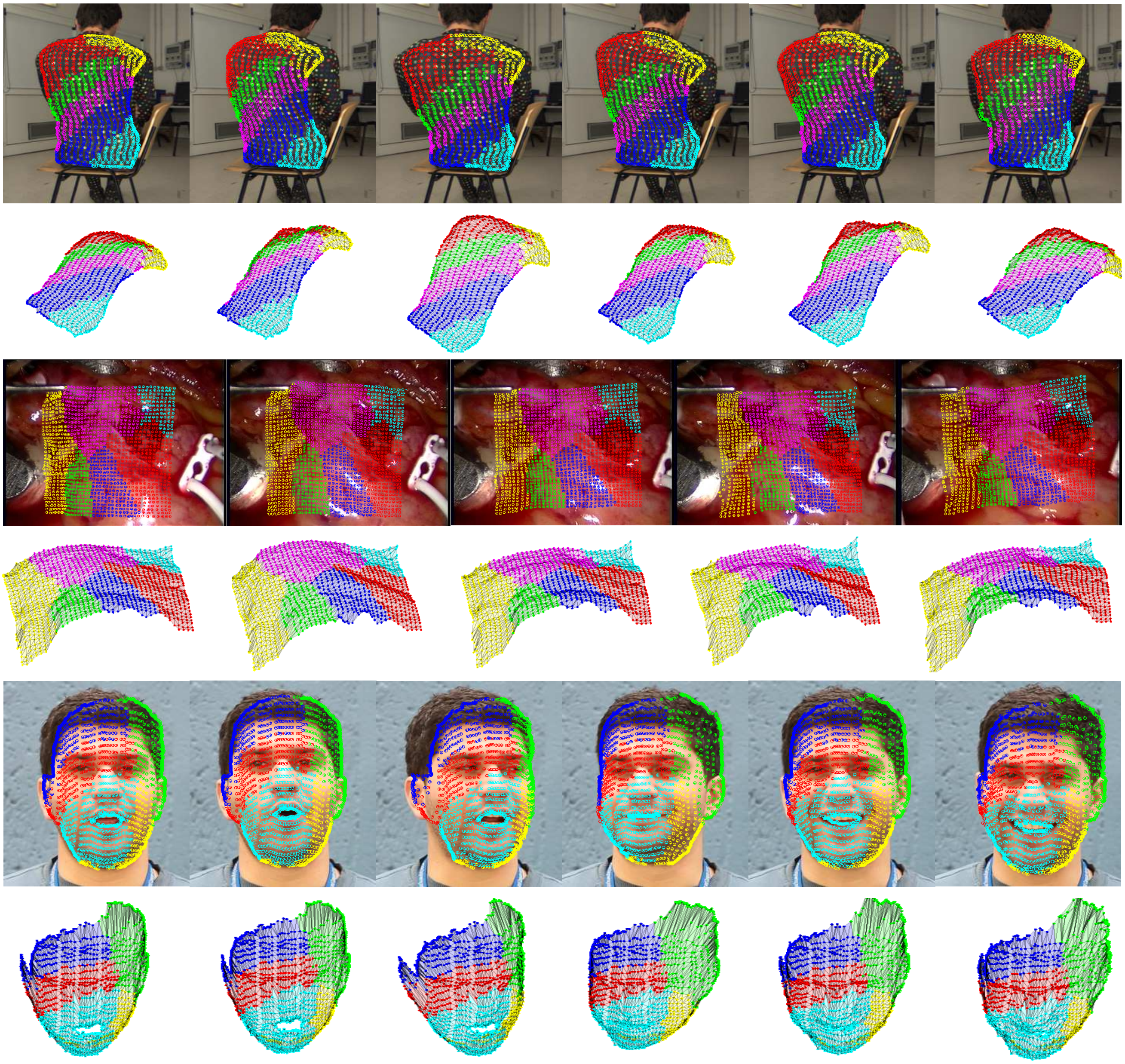

Figure 5. **Real Video Sequences.** For each sequence we show on the top the input images with the 2D tracking data (circles) and the reprojected 3D object (dots). In the row below we show the 3D reconstructed shape from a different viewpoint. Colored regions, represent the retrieved segments ($C = 6$ for *Back* and *Heart*; $C = 5$ for *Face*).

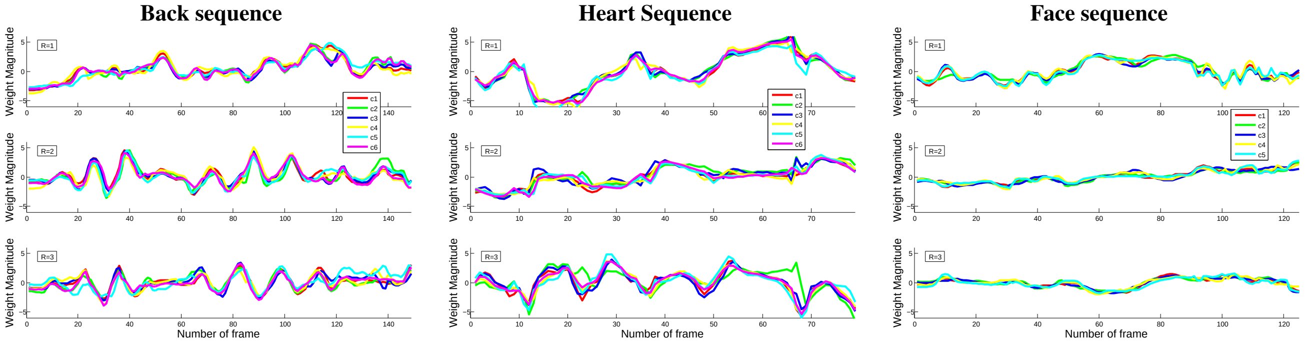

Figure 6. **Local-to-global interpretation.** We display the temporal evolution of the weight coefficients $\boldsymbol{\varphi}^t$. The number of basis is $R = 3$ for all sequences, and the number of segmented clusters is $C = 5$ or $C = 6$. Note that for the *Face* sequence the 5 interpretations are very similar, so our local interpretation model becomes a global one. For *Back* and *Heart* sequences, we observe larger deviations, and hence the local interpretation is well suited to model this type of deformations. Best viewed in color.

at the top of the state-of-the-art in low-rank shape models, took 303.87 sec. to process both *Face1* and *Face2* sequences, and hence our method also outperforms this method in terms of efficiency.

We also use the *Expressions* dataset to evaluate our approach as a function of the dimensionality R of the low-rank model, and of the number C of clusters. The results are shown in Fig. 4 for the noise-less case (left), when adding 2D noise to the input data (middle) and against missing data (right). The overall pattern is that there is little influence of these two parameters, as long as more than one cluster is chosen. This result indicates that choosing more than one cluster is advantageous for the kind of sequences we have chosen, where distinct regions obey different deformation patterns. In the *Expressions* sequence, for instance, the deformation of the mouth is very different from that of the cheek or the forehead. Finally, we also validate our physics-based segmentation compared to a Laplace-Beltrami criterion [32, 34]. The results in Table 1 for this case are: 2.90(3), 2.49(2) and 1.91(2), respectively. This means our segmentation approach provides more accurate reconstructions than competing distance-based descriptors [32, 34].

6.2. Real Video Sequences

We also qualitatively evaluate our approach on three real and semi-dense sequences (about 1,000 points). We first process the *back* sequence, with 150 frames showing the back of a person deforming sideways and flexing [33]. We use the 878 point tracks provided by [20]. Figure 5-top shows the 3D reconstruction. We also represent the physics-based segmentation. The second sequence (79 frames, 1,332 points) shows a *beating heart* acquired during bypass surgery. Figure 5-middle depicts the 3D reconstruction, where one of the main challenges is to handle the small camera motion. Finally, we process a real face sequence (125 frames, 1,442 points). Figure 5-bottom shows the 3D reconstruction and the segmentation we obtain. Interestingly, note how the clusters seem to group areas with similar physical behavior, like the two eyes, and a distinctive region around the mouth.

In Fig. 6 we show the temporal evolution of the coefficients $\boldsymbol{\varphi}^t$ in Eq. (5) for each of the clusters $C = \{1, \ldots, 6\}$, and for each of the basis $R = \{1, 2, 3\}$. Note that for the *Back* and *Heart* sequences the coefficients differ much more among clusters, than for the *Face* sequence. This responds to the fact that the latter, only shows a small amount of non-rigid deformation around the mouth and eyes areas, which could be appropriately modeled by a global model (or equivalently considering $C = 1$, one single cluster). Our approach automatically behaves this way, by making the weights of the different clusters almost identical.

7. Conclusion

In this paper we have presented a new low-rank model for representing the shape of deformable objects. We have shown that classical low-rank global models can be locally interpreted by just rearranging their terms and introducing new region-specific coefficients. With this strategy we get the best of existing global and local models: ability to specialize the model to local object regions, and no need to enforce global consistency on these local interpretations. Additionally, we have proposed a new physically-based mesh segmentation approach, that computes local regions based on an energy-deformation criteria. We illustrate the effectiveness of both these contribution in the NRSfM problem, where our solution produces more accurate 3D reconstructions than state-of-the-art approaches. Even though our results are accurate, our future work is oriented to refine the mesh clustering while learn shape and motion.

Acknowledgment: This work has been partially supported by the Spanish Ministry of Science and Innovation under project RobInstruct TIN2014-58178-R, by the ERA-net CHISTERA project I-DRESS PCIN-2015-147 and by the EU project LOGIMATIC H2020-Galileo-2015-1-687534.

References

[1] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *CVPR*, 2014.

[2] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Online dense non-rigid 3D shape and camera motion recovery. In *BMVC*, 2014.

[3] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Modal space: A physics-based model for sequential estimation of time-varying shape from monocular video. *JMIV, to appear*, 2016.

[4] A. Agudo, J. M. M. Montiel, B. Calvo, and F. Moreno-Noguer. Mode-shape interpretation: Re-thinking modal space for recovering deformable shapes. In *WACV*, 2016.

[5] A. Agudo and F. Moreno-Noguer. Learning shape, motion and elastic models in force space. In *ICCV*, 2015.

[6] A. Agudo, F. Moreno-Noguer, B. Calvo, and J. M. M. Montiel. Sequential non-rigid structure from motion using physical priors. *TPAMI*, 38(5):979–994, 2016.

[7] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Non-rigid structure from motion in trajectory space. In *NIPS*, 2008.

[8] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. *TOG*, 31(2):17:1–17:12, 2012.

[9] J. Barbic and D. James. Real-time subspace integration for st. venant-kirchhoff deformable models. *TOG*, 24(3):982–990, 2005.

[10] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *CVPR*, 2008.

[11] K. J. Bathe. *Finite element procedures in Engineering Analysis*. Prentice-Hall, 1982.

[12] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 1999.

[13] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer, 2008.

[14] M. Bronstein and I. Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *CVPR*, 2010.

[15] M. Carrière, S. Y. Oudot, and M. Ovsjanikov. Stable topological signatures for points on 3D shapes. In *SGP*, 2015.

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

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

[18] A. Del Bue, X. Llado, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *CVPR*, 2006.

[19] J. Fayad, L. Agapito, and A. Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In *ECCV*, 2010.

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

[21] A. Golovinskiy and T. Funkhouser. Randomized cuts for 3D mesh analysis. *TOG*, 27(5), 2008.

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

[23] P. F. U. Gotardo and A. M. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *CVPR*, 2011.

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

[25] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3D Mesh Segmentation and Labeling. *TOG*, 29(3), 2010.

[26] M. Lee, J. Cho, C. H. Choi, and S. Oh. Procrustean normal distribution for non-rigid structure from motion. In *CVPR*, 2013.

[27] A. Loureno, S. R. Bul, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, and M. Pelillo. Probabilistic consensus clustering using evidence accumulation. *ML*, 98(1):331–357, 2013.

[28] M. Marques and J. Costeira. Optimal shape from estimation with missing and degenerate data. In *WMVC*, 2008.

[29] F. Moreno-Noguer and J. M. Porta. Probabilistic simultaneous pose and non-rigid shape recovery. In *CVPR*, 2011.

[30] M. Paladini, A. Bartoli, and L. Agapito. Sequential non rigid structure from motion with the 3D implicit low rank shape model. In *ECCV*, 2010.

[31] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito. Factorization for non-rigid and articulated structure using metric projections. In *CVPR*, 2009.

[32] E. Rodolà, S. R. Bulò, and D. Cremers. Robust region detection via consensus segmentation of deformable shapes. *CGF*, 33(5):97–106, 2014.

[33] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *CVPR*, 2011.

[34] R. M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. In *SGP*, 2007.

[35] S. Sclaroff and A. Pentland. Modal matching for correspondence and recognition. *TPAMI*, 17(6):545–561, 1995.

[36] E. Simo-Serra, C. Torras, and F. Moreno-Noguer. DaLI: Deformation and light invariant descriptor. *IJCV*, 115(11):136–154, 2015.

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

[38] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2):137–154, 1992.

[39] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors. *TPAMI*, 30(5):878–892, 2008.

[40] J. Valmadre and S. Lucey. General trajectory prior for non-rigid reconstruction. In *CVPR*, 2012.

[41] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion. *IJCV*, 67(2):233–246, 2006.

[42] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *SIGGRAPH*, 2004.