Online Dense Non-Rigid 3D Shape and Camera Motion Recovery

Antonio Agudo[1]
aagudo@unizar.es
J. M. M. Montiel[1]
josemari@unizar.es
Lourdes Agapito[2]
l.agapito@cs.ucl.ac.uk
Begoña Calvo[1,3]
bcalvo@unizar.es

[1] Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Zaragoza, Spain.

[2] Department of Computer Science University College London, London, United Kingdom.

[3] Centro de Investigación en Red en Bioingeniería, Biomateriales y Nanomedicina, Zaragoza, Spain.

Abstract

This paper describes a sequential solution to dense non-rigid structure from motion that recovers the camera motion and 3D shape of non-rigid objects by processing a monocular image sequence as the data arrives. We propose to model the time-varying shape with a probabilistic linear subspace of mode shapes obtained from continuum mechanics. To efficiently encode the deformations of dense 3D shapes that contain a large number of mesh vertexes, we propose to compute the deformation modes on a down-sampled rest shape using finite element modal analysis at a low computational cost. This sparse shape basis is then grown back to dense exploiting the shape functions within a finite element. With this probabilistic low-rank constraint, we estimate camera pose and non-rigid shape in each frame using expectation maximization over a sliding window of frames. Since the time-varying weights are marginalized out, our approach only estimates a small number of parameters per frame, and hence can potentially run in real time. We evaluate our algorithm on both synthetic and real sequences with 3D ground truth data for different objects ranging from inextensible to extensible deformations and from sparse to dense shapes. We show the advantages of our approach with respect to competing sequential methods.

1 Introduction

Recovering the 3D shape of rigid scenes along with the camera motion from monocular image sequences, or rigid Structure *f*rom Motion (S*f*M), is one of the most active areas in computer vision. In the last decade, S*f*M methods have made significant progress to simultaneously recover camera motion and shape in real-time for a sparse set of salient points [16, 20] and even to perform dense reconstructions [21] from video sequences acquired with a hand-held camera. On the other hand, Non-Rigid Structure *f*rom Motion (NRS*f*M) methods can simultaneously estimate camera motion and 3D reconstruction of deformable objects from monocular video. Most approaches model deformations assuming a low-rank shape

[5, 8, 9, 10, 34] or trajectory basis [4, 15]. Typically these NRS*f*M methods can only reconstruct a sparse set of salient points. Moreover, they process all the frames in the sequence simultaneously in batch after acquisition has taken place, preventing them from online real-time performance. However, recently NRS*f*M methods have been extended both to *dense shape reconstruction* [12] and to *sequential processing* using either sparse [1, 23] or dense tracks [3]. While [3] exhibits an advantageous accuracy vs. computational cost trade-off, the computation time to compute the shape basis may become prohibitive for some dense cases.

In this paper, we present a sequential solution to solve monocular NRS*f*M. Our approach recovers camera pose and the non-rigid 3D shape of deforming surfaces including large deformations and displacements and it is valid for sparse and dense data. We use a linear subspace of mode shapes with a Gaussian prior on each mode shape to encode the non-rigid 3D shape. Our online system begins estimating a rest shape from the first few frames where continuum mechanics is applied to compute a shape basis. To efficiently encode the deformations of dense 3D models, the dense rest shape is down-sampled to a sparse mesh where modal analysis is applied at a low computational cost. The sparse mode shapes are then grown back to dense. With this low-rank shape basis, the system only optimizes the camera pose and a measurement noise per frame by using Expectation Maximization (EM) over a sliding temporal window of frames. Since the basis weights in the subspace are marginalized out, we only estimate a small number of parameters per frame, and hence our system can potentially run in real time. We show our approach is also adequate for highly extensible surfaces without any 3D training data prior.

2 Related Work

NRS*f*M is an inherently ill-posed problem unless additional a priori knowledge about the shape or the camera motion is considered. In [8] was proposed a low-rank shape constraint as an extension of rigid factorization algorithm [32] to estimate non-rigid shape and camera motion from monocular images. They modeled a time-varying shape as a linear combination of an unknown shape basis under orthography. Although this prior has proved to be a powerful constraint, it is insufficient to solve the inherent ambiguities in NRS*f*M. Most approaches have required to use additional priors such as temporal smoothness [5, 10, 34, 35], smooth-time trajectories [4, 15], spatial smoothness [12, 34] and inextensibility constraints [35]. Piecewise approaches have obtained accurate results estimating local reconstructions to enforce a single smooth surface [26, 30].

Hierarchical Bayesian priors have not been extensively used in NRS*f*M. These priors allow to marginalize out hidden data that does not have to be explicitly computed, thus simplifying the optimization problem and avoiding overfitting. Hierarchical Bayesian priors were used in NRS*f*M to model deformation weights in a low dimensional subspace based on Principal Component Analysis (PCA) [33] estimating the remaining model parameters by EM. These priors were also used in template-based methods, where a Gaussian Process Latent Variable Model was employed to learn a prior over the deformations of local surfaces patches [27]. The unknown shape is encoded as a linear combination of deformation modes learned on-the-fly for a relatively small deformation [33], or in advance from a relatively large set of training data [27]. However, the deformations of real-world shapes can need larger values of rank, and hence the reconstruction becomes underconstrained for methods that learn deformation modes on-the-fly [33]. This ambiguity can be reduced using a pre-defined shape basis. In [1, 2], hierarchical priors were used to model nodal forces in a

physics-based deformation model combining an Extended Kalman Filter (EKF) with the Finite Element Method (FEM) to predict non-rigid displacements.

While sequential real-time SfM [16, 21] solutions exist for rigid scenes, online estimation of non-rigid shape from monocular images remains as a challenging problem. Recently, sequential formulations to NRSfM have emerged [1, 3, 23]. An online solution based on Bundle Adjustment (BA) over a sliding window was proposed in [23]. Though efficient, it did not achieve real-time performance. The first real-time online solution to NRSfM was proposed in [1, 2] by combining an EKF with FEM to estimate a small set of salient points which belong to a deformable object.

In [3] a sequential system was proposed that combines BA with a low-rank physics-based model. The time-varying shape is modeled as a linear combination of a shape basis computed by applying modal analysis over a shape at rest. The non-rigid modes of the basis can be classified by frequency in two families. The first one is affordable to compute even in the dense case but it can only encode bending out-of-plane deformations. For scenes with stretching in-plane deformations, a few stretching modes have to be included to have a representative basis. Unfortunately, computing these dense modes may become prohibitive –sometimes unfeasible– in terms of computational and memory requirements. To solve this limitation, in this paper we propose to increase the density of some initial sparse deformation modes obtained applying modal analysis to a down-sampled shape basis. This allows to easily compute all the frequency spectrum while keeping the computational cost low.

3 Physics-Based Deformation Model

A common way to model non-rigid 3D shapes in computer vision consists in representing them as a linear combination of shape basis [3, 5, 8, 9, 19, 23, 27, 34]. The problem complexity can be reduced using dimensionality reduction techniques such as PCA [7, 19] or modal analysis [3, 24, 28]. We do not use non-rigid 3D training data to compute a shape basis, but instead just consider that a rest shape can be estimated to apply modal analysis. Let us consider a 3D object at rest $\bar{\mathbf{S}}$ of p points, with its coordinates arranged as a $3p$-vector:

$$\bar{\mathbf{S}} = \begin{bmatrix} \bar{\mathbf{S}}_1^\top & \dots & \bar{\mathbf{S}}_j^\top & \dots & \bar{\mathbf{S}}_p^\top \end{bmatrix}^\top = \begin{bmatrix} X_1\, Y_1\, Z_1 \dots X_j\, Y_j\, Z_j \dots X_p\, Y_p\, Z_p \end{bmatrix}^\top. \tag{1}$$

In order to model the dynamic behavior of this object under external actions, we can use the discretized version of the Lagrangian dynamics equation $\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{r}$ where $\mathbf{M}$ and $\mathbf{K}$ are $3p \times 3p$ mass and stiffness matrices respectively, numerically evaluated by FEM [6]. $\mathbf{u}$ and $\mathbf{r}$ are $3p \times 1$ vectors of 3D nodal displacements and external forces respectively, and derivatives with respect to time are abbreviated by superposed dots, i.e. $\ddot{\mathbf{u}}(t) \equiv \frac{d^2\mathbf{u}(t)}{dt^2}$. Recently, [3] proposes to model the stiffness matrix employing a combination of plane-stress and Kirchoff's plate [11], and the mass matrix by means of mass conservation to compute a mode shape basis. These modes can be computed as the undamped free vibrations response of the structure $\bar{\mathbf{S}}$ solving the generalized eigenvalue problem in ω^2 [6]:

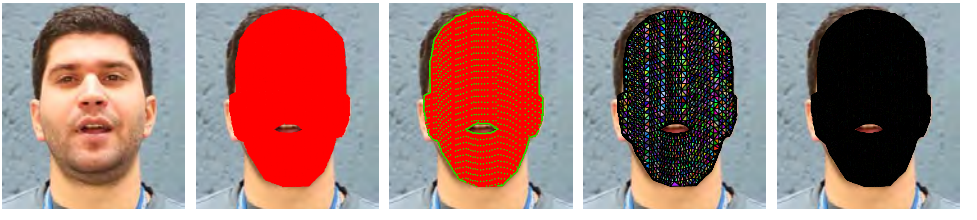
$$\mathbf{K}\psi_k = \omega_k^2 \mathbf{M}\psi_k \tag{2}$$

where the tuple $\left\{\psi_k, \omega_k^2\right\}, k = 1, \dots, 3p$ are mode shapes and vibration frequencies respectively. In this work, we use the stiffness and the lumped mass matrices, as proposed in [3], to compute a mode shape basis. A closed-form expression to approximate any 3D displacement

field can be expressed by means of a transformation matrix $\mathcal{S} \in \mathbb{R}^{3p\times r}$ that concatenates r mode shapes, and an r-vector of weights γ to obtain a low-rank representation as:

$$\mathbf{u} = \mathcal{S}\gamma = \begin{bmatrix}\psi_1 & \dots & \psi_k & \dots & \psi_r\end{bmatrix}\begin{bmatrix}\gamma_1 & \dots & \gamma_k & \dots & \gamma_r\end{bmatrix}^\top. \tag{3}$$

Although we can directly obtain mode shapes solving the eigenvalue problem Eq. (2) for any scene, this problem quickly becomes difficult to solve for dense maps, especially if stretching modes are necessary. To achieve the stretching modes in the frequency spectrum, it is necessary to know a priori the frequency of these modes or to previously compute all bending modes [3]. It is our main contribution to resolve this limitation by computing the eigenvalue problem in a sparse mesh and then the density of the deformation modes is grown –exploiting the shape functions within a finite element– to obtain mode shapes in the dense mesh. This *growth of modes* permits easily to solve dense problems and drastically to reduce the computational and memory requirements for computing both bending or stretching modes.



(a) **(b)** **(c)** **(d)** **(e)**

Figure 1: **Growth-of-modes methodology. (a):** Reference image plane to compute optical flow. **(b):** Dense 2D tracking of p points. **(c):** Subsample of dense shape into q points (green points). **(d):** Delaunay triangulation for sparse mesh. **(e):** Active search to match every point in the sparse mesh. Best viewed in color and with zooming.

3.1 Growth of Modes: Efficient Computation for Dense Cases

We propose to exploit the shape functions used to define the displacement field within a finite element [6]. Therefore, we solve the eigenvalue problem for $q << p$ scene points obtaining $\mathcal{S}^* \in \mathbb{R}^{3q\times r}$ basis shapes and then computing $\mathcal{S} \in \mathbb{R}^{3p\times r}$ for p points using the shape functions. First, we subsample the scene points to convert the p-dimensional map –dense mesh– (Fig. 1 (b)) into q-dimensional map –sparse mesh– (Fig. 1 (c-d)). Then, each point in the dense mesh must be matched with an element of the sparse mesh (Fig. 1 (e)). To find out an element in sparse mesh $\triangle\left(\bar{\mathbf{S}}_a\bar{\mathbf{S}}_b\bar{\mathbf{S}}_c\right)$ with nodal labels $\{a,b,c\}$ per point $\bar{\mathbf{S}}_j$, we suggest an active search computing several cross products over a 2-dimensional space:

$$\sum_{\substack{\tau=0\\ \tau+1\in \mathrm{mod}(3)}}^{2}\left(\chi_\kappa\left(\overrightarrow{\bar{\mathbf{S}}_\tau\bar{\mathbf{S}}_j}\times\overrightarrow{\bar{\mathbf{S}}_\tau\bar{\mathbf{S}}_{\tau+1}}\right) = \begin{cases}1 & \text{if } \left(\overrightarrow{\bar{\mathbf{S}}_\tau\bar{\mathbf{S}}_j}\times\overrightarrow{\bar{\mathbf{S}}_\tau\bar{\mathbf{S}}_{\tau+1}}\right)\in\kappa\\ 0 & \text{if } \left(\overrightarrow{\bar{\mathbf{S}}_\tau\bar{\mathbf{S}}_j}\times\overrightarrow{\bar{\mathbf{S}}_\tau\bar{\mathbf{S}}_{\tau+1}}\right)\notin\kappa\end{cases}\right) = \begin{cases}3 & \text{if } \bar{\mathbf{S}}_j\in\triangle\left(\bar{\mathbf{S}}_0\bar{\mathbf{S}}_1\bar{\mathbf{S}}_2\right)\\ \leq 2 & \text{if } \bar{\mathbf{S}}_j\notin\triangle\left(\bar{\mathbf{S}}_0\bar{\mathbf{S}}_1\bar{\mathbf{S}}_2\right)\end{cases} \tag{4}$$

where the labels $\{a,b,c\}\equiv\{0,1,2\}$ are renumbered and χ_κ represents a step function with $\kappa\equiv[0,\infty)$. $\bar{\mathbf{S}}_j$ is inside triangle element $\triangle\left(\bar{\mathbf{S}}_a\bar{\mathbf{S}}_b\bar{\mathbf{S}}_c\right)$ when all cross products are non-negative. Note that this 2-dimensional space is a dimensional reduction of the shape at rest, and hence it can be both the projection on the image plane after estimating it and the projection on the reference image plane (Fig. 1 (a)) used to compute optical flow [12, 13]

before estimating it. When the active search is completed, we need to compute its natural coordinates (ξ_j, η_j) within the element $\triangle\left(\bar{\mathbf{S}}_a \bar{\mathbf{S}}_b \bar{\mathbf{S}}_c\right)$. First, we transform from the global system to a local system $\triangle\left(\bar{\mathbf{S}}_a^{\mathcal{L}} \bar{\mathbf{S}}_b^{\mathcal{L}} \bar{\mathbf{S}}_c^{\mathcal{L}}\right)$ defined on the plane of each triangle element and then obtaining the natural coordinates as:

$$\begin{bmatrix} \xi_j \\ \eta_j \end{bmatrix} = \left[\begin{bmatrix} \bar{\mathbf{S}}_b^{\mathcal{L}} & \bar{\mathbf{S}}_c^{\mathcal{L}} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{S}}_a^{\mathcal{L}} \otimes \mathbf{1}_2^\top \end{bmatrix} \right]^{-1} \begin{bmatrix} \bar{\mathbf{S}}_j^{\mathcal{L}} - \bar{\mathbf{S}}_a^{\mathcal{L}} \end{bmatrix}, \tag{5}$$

where $\mathbf{1}_2$ is a vector of ones and $\otimes$ indicates the Kronecker's product. The 3D displacement can be obtained using the linear shape functions $\mathfrak{N}^l(\xi_j, \eta_j) \equiv \begin{bmatrix} N_1^l N_2^l N_3^l \end{bmatrix}$ (for more details over these functions, see [6]) within the element. The 3D displacement for every mode shape can be computed applying $\mathcal{S}_a^*$, $\mathcal{S}_b^*$ and $\mathcal{S}_c^*$ are $3 \times r$ displacement vectors for mode shapes basis corresponding to the triangle element $\{a,b,c\}$ which $\bar{\mathbf{S}}_j$ belongs to. Finally, $\mathcal{S}_j$ is placed in rows $3j-2$ through $3j$ in $\mathcal{S}$:

$$\mathcal{S}_j = \begin{bmatrix} \mathfrak{N}^l(\xi_j, \eta_j) \otimes \mathbf{I}_3 \end{bmatrix} \begin{bmatrix} \mathcal{S}_a^* & \mathcal{S}_b^* & \mathcal{S}_c^* \end{bmatrix}^\top. \tag{6}$$

4 Sequential NRS*f*M

We use a linear subspace shape model where the shape basis results from modal analysis to represent the non-rigidly deforming scene. Our aim is to sequentially estimate camera motion and time-varying 3D shape from uncalibrated 2D point correspondences. This section is devoted to describing the details of our sequential approach to NRS*f*M.

4.1 Probabilistic Non-Rigid Shape Model

Let us consider a 3D structure $\mathbf{S}_f$ of p points onto image frame f, the orthographic projection $\mathbf{w}_f$ into vectors can be expressed as:

$$\mathbf{w}_f = \begin{bmatrix} u_{f1} & v_{f1} & \dots & u_{fj} & v_{fj} & \dots & u_{fp} & v_{fp} \end{bmatrix}^\top = \mathbf{G}_f \mathbf{S}_f + \mathbf{T}_f + \mathbf{N}_f \tag{7}$$

where $\mathbf{G}_f = \mathbf{I}_p \otimes \mathbf{R}_f$ with $\mathbf{I}_p$ a $p \times p$ identity matrix, $\mathbf{R}_f = \Pi \mathbf{Q}_f$ are the first two rows of a full rotation matrix $\mathbf{Q}_f$ and Π is the 2×3 orthographic camera matrix. Due to orthographic projection, the depth coordinate of the translation vector cannot be resolved. Considering this ambiguity, we model 2D translations $\mathbf{t}_f$ defined as $\mathbf{T}_f = \mathbf{1}_p \otimes \mathbf{t}_f$ with $\mathbf{1}_p$ a vector of ones and $\mathbf{t}_f = \mathbf{R}_f \mathbf{d}_f$ where $\mathbf{d}_f$ is a 3×1 translation vector. Finally, $\mathbf{N}_f$ is a $2p$-dimensional zero-mean Gaussian noise process vector to model the noise in image tracks. The noise vector for a generic point j is $\mathbf{n}_{fj} \sim \mathcal{N}\left(0; \sigma^2 \mathbf{I}\right)$ with variance of the measurements σ in each dimension. We model the non-rigid 3D shape at each frame $\mathbf{S}_f$ as a linear combination of a mean shape $\bar{\mathbf{S}}$ and r deformation modes $\mathcal{S}$ with the corresponding weight vector γ_f. The non-rigid structure at frame f considering Eq. (3) can be written defining a concatenation matrix of the shape at rest and the mode shapes as $\tilde{\mathcal{S}} = \begin{bmatrix} \bar{\mathbf{S}} & \mathcal{S} \end{bmatrix}$ and $\tilde{\gamma}_f = \begin{bmatrix} 1 & \gamma_f^\top \end{bmatrix}^\top$ as:

$$\mathbf{S}_f = \bar{\mathbf{S}} + \mathbf{u}_f = \bar{\mathbf{S}} + \mathcal{S}\gamma_f = \tilde{\mathcal{S}}\tilde{\gamma}_f. \tag{8}$$

We propose to replace the previous linear subspace modal-analysis based model with a probabilistic model using a Gaussian prior on each shape in the subspace inspired by probabilistic PCA [25, 31, 33]. The weight coefficients γ_f are modeled with a Gaussian prior

distribution with zero-mean $\gamma_f \sim \mathcal{N}(\mathbf{0};\mathbf{I}_r)$. These deformation weights γ_f become latent variables that can be marginalized out and are never explicitly computed. Employing this Gaussian prior over γ_f, the weights for each shape are similar to each other being the non-rigid shape distribution $\mathbf{S}_f \sim \mathcal{N}\left(\bar{\mathbf{S}};\mathcal{S}\mathcal{S}^\top\right)$, producing smooth deformations with respect to rest shape. Note that using this prior, we do not need to add additional temporal smoothness priors to penalize strong variations, avoiding tuning regularization weights [3, 12, 23, 26].

By assuming Gaussian noise over the shape and the observations, the distribution to be estimated over the projected points $\mathbf{w}_f$ is also Gaussian and is equivalent to solving the NRS*f*M problem. Our second contribution is to propose an online EM-based algorithm to solve maximum likelihood estimation as the data arrives in this latent variable problem:

$$\mathbf{w}_f \sim \mathcal{N}\left(\mathbf{G}_f\bar{\mathbf{S}}+\mathbf{T}_f;\mathbf{G}_f\mathcal{S}\mathcal{S}^\top\mathbf{G}_f^\top+\sigma^2\mathbf{I}\right). \tag{9}$$

4.2 Online Expectation Maximization Algorithm

We propose an online version of the EM algorithm –similar to EM for factor analysis [14]– over a sliding window on the last $\mathcal{W}$ frames as proposed in [3, 23] to perform BA. We denote $\Theta_f \equiv \{\mathbf{R}_f,\mathbf{t}_f,\sigma^2\}$ the set of model parameters to estimate, γ_f as hidden data and $\{\mathbf{w}_f,\gamma_f\}$ as complete data. Given the observable data $\mathbf{w}_{f-\mathcal{W}+1:f}$ over the sliding temporal window of frames with indexes $_{f-\mathcal{W}+1:f}$, we estimate the model parameters over all frames in the current window denoted as $\hat{\mathcal{W}}$. The joint probability of $\mathbf{w}$ over sliding window, assuming samples independent and identically distributed, may be computed considering the Gaussian distribution per frame Eq. (9) as $p\left(\mathbf{w}_{\hat{\mathcal{W}}}|\mathbf{G}_{\hat{\mathcal{W}}},\mathbf{T}_{\hat{\mathcal{W}}},\sigma^2\right) = \prod_{i=f-\mathcal{W}+1}^{f} p\left(\mathbf{w}_i|\mathbf{G}_i,\mathbf{T}_i,\sigma^2\right)$.

The EM algorithm estimates iteratively until convergence the likelihood alternating between two steps: *E-step* and *M-step*. In the *E-step*, we compute the posterior distribution over latent variables given the measurements and the current model parameters on the sliding window using Woodbury's matrix identity [36] as:

$$p\left(\gamma_{\hat{\mathcal{W}}}|\mathbf{w}_{\hat{\mathcal{W}}},\Theta_{\hat{\mathcal{W}}}\right) \sim \prod_{i=f-\mathcal{W}+1}^{f} \mathcal{N}\left(\beta_i\left(\mathbf{w}_i-\mathbf{G}_i\bar{\mathbf{S}}-\mathbf{T}_i\right);\mathbf{I}_r-\beta_i\mathbf{G}_i\mathcal{S}\right), \tag{10}$$

$$\beta_i = \mathcal{S}^\top\mathbf{G}_i^\top\sigma^{-2}\left(\mathbf{I}-\sigma^{-2}\mathbf{G}_i\mathcal{S}\left(\mathbf{I}_r+\sigma^{-2}\mathcal{S}^\top\mathbf{G}_i^\top\mathbf{G}_i\mathcal{S}\right)^{-1}\mathcal{S}^\top\mathbf{G}_i^\top\right). \tag{11}$$

We just need to compute on-the-fly by updating the model parameters when a new observation is available the expectations $\mu_f \equiv \mathbb{E}\left[\gamma_f\right] = \beta_f\left(\mathbf{w}_f-\mathbf{G}_f\bar{\mathbf{S}}-\mathbf{T}_f\right)$ and $\phi_f \equiv \mathbb{E}\left[\gamma_f\gamma_f^\top\right] = \mathbf{I}_r-\beta_f\mathbf{G}_f\mathcal{S}+\mu_f\mu_f^\top$ for each frame f on the sliding window.

In the *M-step*, the expected value of log-likelihood function is optimized by replacing the latent variables by their expected values to update the model parameters. We update motion variables maximizing likelihood with respect to parameters Θ holding hidden distribution fixed:

$$\begin{aligned}\mathcal{A}(\Theta_i) &= \arg\min_{\Theta_i}\ \mathbb{E}\left[-\sum_{i=f-\mathcal{W}+1}^{f}\log p\left(\mathbf{w}_i|\Theta_i\right)\right] = \\ \arg\min_{\mathbf{G}_i,\mathbf{T}_i,\sigma^2}\ &\frac{1}{2\sigma^2}\sum_{i=f-\mathcal{W}+1}^{f}\mathbb{E}\left[\|\mathbf{w}_i-\mathbf{G}_i\left(\bar{\mathbf{S}}+\mathcal{S}\gamma_i\right)-\mathbf{T}_i\|_2^2\right]+p\mathcal{W}\log\left(2\pi\sigma^2\right).\end{aligned} \tag{12}$$

This function can not be optimized in closed-form to compute a global optimum and partial M steps are necessary. The Θ vector is individually updated in closed-form, except for

camera rotation. Similar to batch-mode EM [33], the noise variance and the translation vector can be updated in sequential-mode processing considering $\tilde{\mu}_i = \mathbb{E}\left[\tilde{\gamma}_i\right]$ and $\tilde{\phi}_i = \mathbb{E}\left[\tilde{\gamma}_i\tilde{\gamma}_i^\top\right]$:

$$\sigma^2 = \frac{1}{2p\mathcal{W}} \sum_{i=f-\mathcal{W}+1}^{f} \left(\|\mathbf{w}_i - \mathbf{T}_i\|^2 - 2\left(\mathbf{w}_i - \mathbf{T}_i\right)^\top \mathbf{G}_i \tilde{\mathcal{S}} \tilde{\mu}_i + \mathrm{tr}\left(\tilde{\mathcal{S}}^\top \mathbf{G}_i^\top \mathbf{G}_i \tilde{\mathcal{S}} \tilde{\phi}_i\right)\right), \tag{13}$$

$$\mathbf{t}_i = \frac{1}{p} \sum_{j=1}^{p} \left(\mathbf{w}_{ij} - \mathbf{R}_i \tilde{\mathcal{S}}_j \tilde{\mu}_i\right). \tag{14}$$

However, the camera rotation $\mathbf{Q}_i$ is subject to orthonormality constraints, and a closed-form update is not possible. Rotation matrices lie exactly on a smooth manifold based on the orthogonal group $SO(3)$, where it is possible to generalize a Riemannian-Newton algorithm [11, 29]. We use Riemannian manifold optimization to update the rotation matrices. First, we rewrite the expected negative log-likelihood function Eq. (12) dropping the dependence on σ^2 as:

$$\mathcal{A}(\mathbf{Q}_i) = \underset{\mathbf{Q}_i \in SO(3)}{\arg\min} \sum_{i=f-\mathcal{W}+1}^{f} \sum_{j=1}^{p} \mathbb{E}\left[\|\mathbf{w}_{ij} - \Pi \mathbf{Q}_i \tilde{\mathcal{S}}_j \tilde{\gamma}_i - \mathbf{t}_i\|_{\mathcal{F}}^2\right], \tag{15}$$

where $\mathbf{Q}_i \in SO(3)$ and its tangent $\Delta_{Q_i} \in T_Q(SO(3))$ can be expressed as $\Delta_{Q_i} = \mathbf{Q}_i [\delta]_\times$ with $[\delta]_\times$ the skew-symmetric matrix. On $SO(3)$, the geodesic at $\mathbf{Q}_i$ in the tangent direction can be expressed by means of the Rodrigues' rotation formula $\mathbf{Q}\left(\hat{\delta}, \alpha\right) = \mathbf{Q}\Delta_Q = \mathbf{Q}\left(\mathbf{I}_3 + \left[\hat{\delta}\right]_\times \sin(\alpha) + \left[\hat{\delta}\right]_\times^2 (1 - \cos(\alpha))\right)$, where $[\delta]_\times \in \mathfrak{so}(3)$ is the Lie algebra of $SO(3)$ group and $[\delta]_\times = \alpha \left[\hat{\delta}\right]_\times$. This explicit formula for geodesics is necessary to compute the gradient $d\mathcal{A}(\Delta_{Q_i})$ and the Hessian $Hess\,\mathcal{A}(\Delta_{Q_i}, \Delta_{Q_i})$ of the cost function Eq. (15) along the geodesics on the manifold. Finally, the optimal update rotation vector along the geodesic is computed as $\Delta_{Q_i} = \mathbf{Q}_i [\delta]_\times$ where δ is estimated as $\delta = -\mathbf{H}^{-1}\mathbf{g}$. The Hessian matrix $\mathbf{H}$ and the gradient vector $\mathbf{g}$ can be obtained applying $d\mathcal{A}(\Delta_{Q_i})$ and $Hess\,\mathcal{A}(\Delta_{Q_i}, \Delta_{Q_i})$ over a standard algorithm for optimization on manifolds [17, 29].

4.3 Initialization

Our method assumes that the rest shape can be estimated similarly to [3, 23, 26]. We use a rigid factorization algorithm [18] on a few initial frames (nr mostly rigid frames) to recover it. Our rest shape is a tuple $(\mathcal{P}, \mathcal{E})$ where $\mathcal{P}$ is a finite set of p nodes and $\mathcal{E}$ of m triangular elements –over q points for dense cases– obtained by means of a Delaunay triangulation [6] on the image plane or on the reference image for dense cases [12, 13]. To initialize the model parameters for a new incoming frame, the camera pose is initialized as $\mathbf{R}_i = \mathbf{R}_{i-1}$ and $\mathbf{t}_i = \mathbf{t}_{i-1}$, while the latent variables as $\mathbb{E}\left[\gamma_i\right] = \mathbb{E}\left[\gamma_{i-1}\right]$.

5 Experimental Results

We present experimental results on both synthetic and real sequences, providing both qualitative and quantitative evaluation where we compare our approach with respect to state-of-the-art methods[1]. We use the error metric defined as $e_{3\mathcal{D}} = \frac{1}{f}\sum_{i=1}^{f} \frac{\|\mathbf{S}_i - \mathbf{S}_i^{GT}\|_{\mathcal{F}}}{\|\mathbf{S}_i^{GT}\|_{\mathcal{F}}}$ [3] where $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm, $\mathbf{S}_i$ is the 3D reconstruction and $\mathbf{S}_i^{GT}$ is the ground truth.

[1] Videos of the experimental results can be found on website http://webdiis.unizar.es/~aagudo

5.1 Synthetic Data

We propose a synthetic dense sequence of 109 frames where we simulate an elastic ribbon deformation with $p = 2{,}263$ points observed by an orthographic camera. The elastic ribbon is modeled with a non-linear Yeoh hyperelastic material [37] valid for large deformations. This sequence was generated with the simulation tool Abaqus, where material properties, boundary conditions and nodal forces are necessary to compute the deformation. In contrast, our shape basis is computed using linear elasticity only valid for small deformations and the material properties, boundary conditions and external forces are unknown.

We use our EM-FEM algorithm with $\mathcal{W} = 3$ for this challenging 43% stretching deformation computing the mode shapes with a sparse mesh of $q = 78$ points and applying a growth of modes to $p = 2{,}263$ points. We obtain an error $e_{3\mathcal{D}}$ of 2.95% when we use 5 stretching modes, 1.47% with 20 and 0.79% with 40. Applying the MP algorithm [22], the error $e_{3\mathcal{D}}$ is 17.72% with 15 shape basis. Fig. 2 shows our 3D reconstruction with $r = 40$ for a few selected frames including where the stretching is maximum, as well as a qualitatively comparison with respect to ground truth. While this problem would incur a high computational cost to compute stretching mode shapes [3], in our algorithm it is negligible since we reduce the dimension of the eigenvalue problem from $2{,}263\times3$ to 78×3.

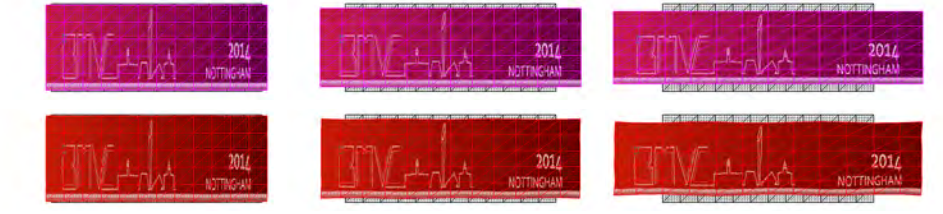


Figure 2: **Dense stretching ribbon sequence.** Reconstruction of the dense 2,263 point ribbon at frames #20, #60 and maximum deformation at #109. Shape at rest is displayed with a black mesh. **Top:** Ground truth deformation with a magenta mesh. **Bottom:** Renderings and thin lines for our reconstruction. Sparse mesh in thick lines. Best viewed in color.

5.2 Real Data

In this section, we evaluate our method on a challenging dense dataset of motion capture (MoCap) [13] with $p = 9{,}622$, corresponding to a flag waving in the wind. We present a comparison with respect to competing sequential NRS*f*M methods, considering the methods: SBA [23] and BA-FEM [3]. We exactly use for each method the same $\mathbf{w}_{1:nr}$ to compute the shape at rest by means of a rigid factorization algorithm [18].

To validate the scalability of our method, we also propose a $p = 594$ sparse version of this dense flag sequence which is the result of the subsampling process. We apply the growth of modes to this sparse sequence. We show quantitative 3D reconstruction for both the sparse and the dense flag sequence in Table 1 and qualitative results in Fig. 3. Both BA-FEM [3] and our EM-FEM outperform the SBA [23] method in terms of accuracy and efficiency for both cases. Although SBA [23] estimates the mode shapes on-the-fly –with smaller initialization computation time only for shape at rest– it is not able to overcome the mode shapes for this sequence. Note that the three methods use exactly the same initialization, although both BA-FEM [3] and our EM-FEM exploit the shape at rest to compute a mode shape basis. The error

$e_{3\mathcal{D}}$ obtained with BA-FEM [3] is smaller in the dense case than the sparse case due to the accuracy of dense mode shapes estimation –without growth of modes–. However, applying our EM-FEM, the error is smaller for sparse case due to the approximation of growth of modes with little reduction in accuracy. Note that the initialization computation time for the sparse problem is the same using both approaches since we do not use a growth of modes. Applying this method to the dense case, the computation time increment is negligible – 0.03 sec– with respect to the sparse case, and it is dominated for the rigid factorization step with 25.67sec, similar to SBA [23]. Moreover, our approach is more efficient in the sequential estimation, with a better scalability in the number of modes. We can conclude that our method outperforms the sequential state-of-the-art in terms of accuracy and efficiency. Although our method is implemented in unoptimized Matlab code without parallelization over a commodity computer Intel core i7@2.67 GHz, the results show a low computational cost per frame and it could potentially run in real time.

Algorithm	**Sparse Flag**		**Dense Flag** [13]	
	$e_{3\mathcal{D}}(\%)$	*in / op* (sec)	$e_{3\mathcal{D}}(\%)$	*in / op* (sec)
SBA [23]	7.10(114*)	0.58/82.32	13.48(114*)	25.67/895
BA-FEM [3]	3.72(10)	19.50/1.96	3.50(10)‡	300/75‡
	3.49(40)	19.50/24.83	3.29(25)‡	300/416‡
EM-FEM	3.28(10)	19.50/1.53	3.41(10)	44.62/62
	2.81(40)	19.50/2.28	3.08(25)	44.62/68

Table 1: **Quantitative comparison for MoCap sequence.** We show error $e_{3\mathcal{D}}$ for sequential methods SBA [23], BA-FEM [3] and for our method EM-FEM. In both cases, we show in brackets the number of shapes in the basis. We also show computation time for initialization *in* and optimization process per frame *op*. ‡: value reproduced from [3]. $*$: SBA [23] reports the rank r, we detail $3r$ in brackets because it is equivalent to the number of weights in both BA-FEM [3] and our EM-FEM. In all cases, $\mathcal{W} = 5$.

Figure 3: **Flag MoCap sequence:** 3D Reconstruction for a few frames with red dots. **Top:** Sparse 594 points flag. Ground truth is overlaid with a black mesh and black circles. **Bottom:** Dense 9,622 points flag. Best viewed in color and with zooming.

Finally, we evaluate our approach using dense 28,332 tracks provided by [13] to show a qualitative evaluation on the face sequence with respect to [12], where a subject is performing natural expressions and moving his head. In Fig. 4 we show few frames with our 3D reconstruction using $r = 30$ modes and $\mathcal{W} = 3$. We compute mode shapes with $q = 1{,}442$ points and applying a growth of modes to $p = 28{,}332$ points.

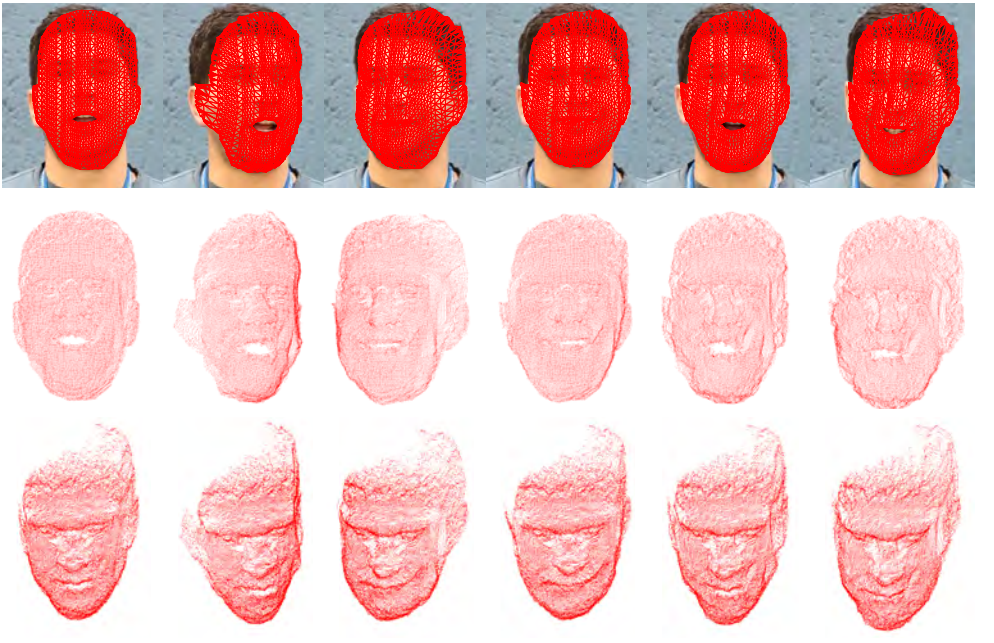


Figure 4: **Face sequence: dense 3D point cloud estimation. Top:** Selected frames #17, #43, #55, #75, #100 and #121 with 3D reprojected mesh. **Middle:** Original viewpoint. **Bottom:** General view.

6 Conclusions and Future Work

This paper presents an online system that takes NRS*f*M closer to real-time operation. We use a physics-based linear subspace model with hierarchical priors to encode a time-varying shape without a learning step. To compute the dense shape basis, we have proposed a growth of modes that uses a small set of points to obtain a sparse shape basis that is later extended to dense at quite affordable cost. Camera motion and non-rigid shape are estimated on-the-fly within a low-cost EM framework as the data arrives. Experimental results on challenging sequences show that our approach can sequentially estimate from isometric to extensible deformations. Our future work is to incorporate feature tracking and outlier detection into a single process.

Acknowledgments

This work was partly funded by the MINECO projects DIP2012-32168 and DPI2011-27939-C02-01; by the ERC Starting Grant agreement 204871-HUMANIS; and by a scholarship FPU12/04886. The authors wish to thank M. A. Ariza for the stretching sequence.

References

[1] A. Agudo, B. Calvo, and J. M. M. Montiel. Finite element based sequential bayesian non-rigid structure from motion. In *CVPR*, 2012.

[2] A. Agudo, B. Calvo, and J. M. M. Montiel. 3D reconstruction of non-rigid surfaces in real-time using wedge elements. In *Workshop on NORDIA*, 2012.

[3] 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.

[4] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *TPAMI*, 33(7):1442–1456, 2011.

[5] 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.

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

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. ACM SIGGRAPH*, pages 187–194, 1999.

[8] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

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

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

[11] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(20):303–353, 1998.

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

[13] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *IJCV*, 104(3):286–314, 2013.

[14] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. *Technical Report CRG-TR-96-1, University of Toronto*, 1996.

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

[16] G. Klein and D. W. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007.

[17] Y. Ma, J. Kosecka, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *IJCV*, 44(3):219–249, 1999.

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

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

[20] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Generic and real-time structure from motion using local bundle adjustment. *Image Vision Computing*, 27(8):1178–1193, 2009.

[21] R. Newcome, S. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *ICCV*, 2011.

[22] 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.

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

[24] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *TPAMI*, 13 (7):730–742, 1991.

[25] S. Roweis. EM algorithms for PCA and SPCA. In *NIPS*, 1997.

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

[27] M. Salzmann, R. Urtasun, and P. Fua. Local deformation models for monocular 3D shape recovery. In *CVPR*, 2008.

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

[29] A. Shaji and S. Chandran. Riemannian manifold optimisation for non-rigid structure from motion. In *Workshop on NORDIA*, 2008.

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

[31] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

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

[33] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3D shape from 2D motion. In *NIPS*, 2003.

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

[35] S. Vicente and L. Agapito. Soft inextensibility constraints for template-free non-rigid reconstruction. In *ECCV*, 2012.

[36] M. A. Woodbury. Inverting modified matrices. *Statistical Research Group, Memorandum Rept. 42*, 1950.

[37] O. H. Yeoh. Some forms of the strain energy function for rubber. *Rubber Chemistry and technology*, 66(5):754–771, 1993.