Mode-Shape Interpretation: Re-Thinking Modal Space for Recovering Deformable Shapes

Antonio Agudo[1] J. M. M. Montiel[2] Begoña Calvo[2] Francesc Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
[2]Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain

Abstract

This paper describes an on-line approach for estimating non-rigid shape and camera pose from monocular video sequences. We assume an initial estimate of the shape at rest to be given and represented by a triangulated mesh, which is encoded by a matrix of the distances between every pair of vertexes. By applying spectral analysis on this matrix, we are then able to compute a low-dimensional shape basis, that in contrast to standard approaches, has a very direct physical interpretation and requires a much smaller number of modes to span a large variety of deformations, either for inextensible or extensible configurations. Based on this low-rank model, we then sequentially retrieve both camera motion and non-rigid shape in each image, optimizing the model parameters with bundle adjustment over a sliding window of image frames. Since the number of these parameters is small, specially when considering physical priors, our approach may potentially achieve real-time performance. Experimental results on real videos for different scenarios demonstrate remarkable robustness to artifacts such as missing and noisy observations.

1. Introduction

The simultaneous 3D reconstruction of rigid structures and the camera motion using a set of uncalibrated images has been extensively studied over the last few decades. The rigidity prior has proven to be a powerful constraint to solve the problem, allowing practical and robust solutions [21]. However, rigid reconstruction techniques fail when applied directly to time-deforming objects such as a piece of cloth or a beating heart. In order to overcome this limitation, Non-Rigid Structure-*f*rom-Motion (NRS*f*M) techniques have been proposed [5, 8, 15, 18, 38], allowing one to recover the 3D shape of non-rigid objects over time. This task is a fundamental problem in computer vision, with potentially many real-world applications in sports, the movie industry or medical imaging. Since many different 3D shapes can have similar image measurements, NRS*f*M is an inherently ill-posed problem and additional a priori knowledge about the camera motion and the deformation of the object have been considered [9, 16, 38]. Only very recently, this problem has been also tackled in a sequential manner [2, 3, 7, 28]. In this case, the estimation of time-varying objects can only be done by considering the observations until current frame. However, this scenario is paramount for bringing such algorithms to real situations (e.g., in augmented reality or operating rooms) that require real-time solutions. For these cases, solving the problem assuming neither deformable 3D training data nor a deformation model is particularly challenging.

In this paper, we present a new shape basis interpretation to code non-rigid shapes. We only need a rest shape estimation of the deformable object, which we obtain from the first few frames of the monocular video. Using this shape, we compute a matrix encoding the distances between every pair of points of the shape, which then allows obtaining a reduced shape basis by means of spectral decomposition. The shape basis has direct physical interpretation and is subsequently used to span the deformation of the object in a low-rank space in which the time-varying coefficients have to be estimated. We propose incorporating this low-rank model into an on-line Bundle Adjustment (BA) framework to simultaneously retrieve the camera pose and the time-varying shape. Our approach may potentially run in real time, since the number of parameters to optimize per frame is relatively small. In addition, our method estimates the shape basis at low computational cost compared to state-of-the-art algorithms, since the eigenvalue problem we need to solve is fairly simple. We show our approach to be adequate to encode both inextensible and extensible deformations without knowing any 3D training data in advance. The complexity of our on-line method is linear with the number of points, so it can handle a wide variety of scenarios, going from sparse to semi-dense or dense objects. Further, our method is robust to corrupted observations such as missing data and noise.

2. Related Work

Modeling a deformable 3D object using low-rank models has become a very popular approach in computer vision [8, 11, 15, 18, 29, 33, 36, 38]. Low-rank shape models were firstly proposed to code the time-varying shape by means of a linear subspace of a set of deformation modes. These models, together with orthonormality constraints on the camera motion, have proven successful in the 3D reconstruction of many real-world non-rigid objects. Both unknown shape basis and coefficients were estimated by factorization-based algorithms [11, 15], or adding additional priors such as temporal and spatial smoothness [9, 25, 38] by means of optimization techniques. On the other hand, [8] proposed applying the low-rank constraint to the temporal evolution of each 3D point instead of applying it over the spatial configuration of the shape basis. To this end, each 3D point position was independently coded at every instant by means of a linear combination of trajectory basis based on the Discrete Cosine Transform (DCT). Later, this compact DCT representation was used in [20] to approximate the time evolving shape basis coefficients, by implicitly imposing temporal smoothness on each 3D point trajectory.

Many approaches estimate the shape basis on the fly, but the problem quickly becomes under-constrained when large deformations need to be represented [11, 15, 29, 38]. To reduce the number of parameters to estimate one can use either a pre-defined shape or trajectory basis. For pre-defined shape bases, the problem complexity can be reduced using dimensionality reduction techniques such as Principal Component Analysis (PCA) [10, 27, 33] over a set of non-rigid 3D training data. In a similar way, a 3D shape basis can be built using trained 2D shapes by means of an active appearance model [41, 42] for 3D reconstruction of faces or using latent states in a directed acyclic graph [35]. However, the accuracy of these methods depends on the appropriateness of the training data, which is hard to obtain in practice. Modal Analysis (MA) was proposed to obtain a mode shape basis based on a physical model of a known object [30, 34], or on the shape at rest estimated by means of an initial exploration of the object [2, 4].

On the other hand, invariant transformations for isometric deformations have been proposed applying MultiDimensional Scaling (MDS) on geodesic distance matrices over a template [13, 17, 22, 37] in 3D shape recognition. These approaches rely on obtaining new configurations where point-wise euclidean distances are approximately equal to the original point-wise geodesic distances for both 2D and 3D cases. These methods were extended to represent quasi isometric deformations for 3D face recognition [12].

In this work, we exploit the available information from an initial exploration of the deformable object acquired by an orthographic monocular camera, to compute a pre-defined shape basis and model its deformation over time. Our approach uses this exploration to obtain a 3D shape at rest that is used to compute a dissimilarity measure based on a representation of the structure. Then, we apply an algorithm similar to MDS to obtain a reduced shape basis, which can be interpreted and used to encode both inextensible and extensible deformations. Although we also use a rest shape estimation, unlike MA [2] our method does not require a deformation model, nor 3D training data like PCA-based methods [27, 33]. In fact, our model can be seen as a simplification of the standard MA, that reduces the computational cost while still being valid for a large variety of objects, in particular when the object's material properties are quasi-homogeneous.

3. Proposed Deformation Model

Euclidean distance constraints have typically been used to recover isometric transformations of a deforming object over time [32]. Although these constraints are very restrictive, they have proven to be a powerful prior to solve the inherent ambiguities of both template-based [27, 31, 39] and template-free [14, 39] methods. However, these constraints cannot be applied when the object surface undergoes stretching and/or shearing deformations. In this work, we propose using the *distance information* on the 3D rest shape to compute a shape basis that is valid to code both inextensible and extensible deformations, without any other prior information.

3.1. Modeling Distance Matrices

Let us consider a 3D shape at rest of a dynamic object made of p 3D points, represented by the matrix $\bar{\mathbf{S}} = [\bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, \ldots, \bar{\mathbf{s}}_j, \ldots, \bar{\mathbf{s}}_p]$, where the columns represent the 3D coordinates for each point $\bar{\mathbf{s}}_j \in \mathbb{R}^3$. We describe the object using a triangular mesh, where each vertex corresponds to a 3D scene point and the list of vertexes is written as $\mathcal{S} := \{\bar{\mathbf{s}}_j \in \mathbb{R}^3\}_{j=1}^{p}$, and the index set of $\mathcal{S}$ as $N_p := \{1, \ldots, p\}$. An edge represents a line segment connecting two different vertexes of $\mathcal{S}$, and can be expressed by a tuple of indexes (j, h), $j, h \in N_p, j \neq h$. The list of edges $\mathcal{E} \subset N_p \times N_p$ is denoted by $\mathcal{E} := \{(j,h)_e\}_{e=1}^{n}$ where n is the number of edges. Finally, we obtain a triangular mesh by means of a Delaunay's tessellation, where we represent the list of triangles $\mathcal{T} \subset N_p \times N_p \times N_p$ as $\mathcal{T} := \{(j,h,l)_t\}_{t=1}^{m}$ with $j, h, l \in N_p, j \neq h \neq l \neq j$, and m being the number of triangles in the mesh. Each triangle contains three edges, and we eliminate all the repeated edges from all triangles to obtain $\mathcal{E}$ from $\mathcal{T}$. A path between the points $\bar{\mathbf{s}}_j$ and $\bar{\mathbf{s}}_h$ is a sequence $\Theta(j,h) = \{\bar{\mathbf{s}}_j\}_{j=1}^{p}$, following the piecewise non-directed edges denoted by the set $\mathcal{E}$ that connect the points on the set $\mathcal{S}$.

We next exploit the distance information to compute a symmetric $p \times p$ distance matrix $\mathbf{D}$ from the shape at rest.

First of all, we present different alternatives to compute this matrix.

A distance matrix $\mathbf{D}$ could be modeled employing geodesic distances –considering the shortest path between all pairs of points– on $\bar{\mathbf{S}}$, applying the fast marching approach for curved domains [23] or length estimation based on graph search [24]. However, for simplicity, in this work we will approximate the distance matrix $\mathbf{D}$ by using Euclidean or Manhattan distances, respectively.

Euclidean Distance Matrix: We first define the Euclidean distance matrix $\mathbf{D}_E$ that contains the Euclidean distances between pairs of points on $\bar{\mathbf{S}}$ as:

$$\mathbf{D}_E = \left(\mathbf{b}\mathbf{1}_p^\top + \mathbf{1}_p\mathbf{b}^\top - 2\bar{\mathbf{S}}^\top\bar{\mathbf{S}}\right)^{\frac{1}{2}} \odot \left(\mathbf{1}_p\mathbf{1}_p^\top - \mathbf{I}_p\right), \quad (1)$$

with $\mathbf{b} = \sum(\bar{\mathbf{S}} \odot \bar{\mathbf{S}})$ a $p \times 1$ vector. $\mathbf{1}_p$ and $\mathbf{I}_p$ represent a $p \times 1$ vector of ones and a $p \times p$ identity matrix, respectively. $\odot$ indicates the Hadamard product, i.e., entrywise product, and $\frac{1}{2}$ the element-wise square root. Note that the second product is to fix a null diagonal, avoiding numerical errors. This matrix is a good approximation for quasi-planar shapes and it is the same as the geodesic distance matrix for perfectly planar shapes (see Fig. 1).

Manhattan Distance Matrix: We also define the generalized Manhattan distance matrix $\mathbf{D}_M$ for 3D irregular domains, that contains the Euclidean distances between pairs of points following the path of minimal cost from j to h by means of the Dijkstra's shortest path:

$$\mathbf{D}_M = \bigmathbf{D}_{n=1}^{p(p-1)/2} d_{m_n}(j,h), \quad d_m(j,h) = \min_{\Theta} \sum_{j=1}^{p-1} d_e(j, j+1)$$

where $\bigmathbf{D}$ is the distance assembly operator, i.e., this matrix is assembled from distances between points $d_m(j,h)$ just considering $p(p-1)/2$ terms since the matrix is symmetric with null diagonal. This matrix is a good approximation when it is considered a small neighborhood, such as into dense shapes (see Fig. 1).

3.2. Shape Basis Computation

We now describe how the deformable shape basis is estimated. Following MDS [1, 17], we apply a double centering and normalization transformation to the distance matrix $\mathbf{D}$ by means of the centering matrix $\mathbf{C} = \mathbf{I}_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top$. Then we perform a spectral decomposition of $\bar{\mathbf{D}} \equiv -\frac{1}{2}\mathbf{C}\mathbf{D}\mathbf{C}$ by solving the standard eigenvalue problem:

$$\bar{\mathbf{D}}\boldsymbol{\psi}_j = \omega_j^2\boldsymbol{\psi}_j, \quad (2)$$

where $(\boldsymbol{\psi}_j, \omega_j^2)$, $j \in N_p$ represent the $p \times 1$ eigenvectors and eigenvalues of $\bar{\mathbf{D}}$, respectively. We compute the normalized to one length eigenvectors $\|\boldsymbol{\psi}_j\|_2 = 1$ to satisfy the orthonormality conditions: $\boldsymbol{\psi}_j^\top\bar{\mathbf{D}}\boldsymbol{\psi}_h = \omega_j^2\boldsymbol{\psi}_j^\top\boldsymbol{\psi}_h$ and $\boldsymbol{\psi}_j^\top\boldsymbol{\psi}_h = \delta_{jh}$ with δ_{jh} the Kronecker's delta. Note that in the literature, MDS is typically applied to a distance matrix to obtain new configurations where point-wise distances remain approximately constant [17, 37], i.e., it is applied to isometric deformations. In contrast, this will not be the case of our approach since thanks to the physical interpretation we propose, both inextensible and extensible deformations can be coded.

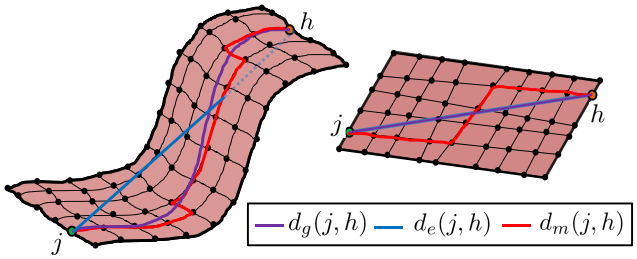

Figure 1. **Distance definition.** We represent Euclidean $d_e(j,h)$ and Manhattan $d_m(j,h)$ distances on non-planar and planar shapes between the points j and h. Geodesic distance $d_g(j,h)$ is well approximated by Euclidean distance in planar or nearly planar objects. For non-planar objects it is better approximated by the Manhattan distance. Best viewed in color.

3.3. Shape Basis with Physical Interpretation

Modeling the non-rigid deformation of an object through a linear combination of mode shapes is a standard practice in computer vision [8, 11, 15, 18, 29, 33, 38]. While most techniques use a full shape basis ($3p$-order vectors), we propose using the eigenmodes computed in previous section, that is a reduced shape basis (p-order vectors). We can represent some non-rigid 3D displacement $\mathbf{U}$ over a rest shape, by a linear combination of r of these modes, each with different physical interpretations, as:

$$\mathbf{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Upsilon}, \quad (3)$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{3\times3}$ is a transformation matrix and $\boldsymbol{\Upsilon} \in \mathbb{R}^{r\times p}$ contains the r reduced mode shapes associated to a p-points structure $\bar{\mathbf{S}}$:

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\psi}_1^\top \\ \boldsymbol{\psi}_2^\top \\ \vdots \\ \boldsymbol{\psi}_r^\top \end{bmatrix} = \begin{bmatrix} \psi_{11} & \psi_{12} & \dots & \psi_{1p} \\ \psi_{21} & \psi_{22} & \dots & \psi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{r1} & \psi_{r2} & \dots & \psi_{rp} \end{bmatrix}. \quad (4)$$

$\boldsymbol{\Lambda} \in \mathbb{R}^{3\times r}$ is a deformation transformation matrix that contains the time-varying coefficients to interpret each reduced mode shape as:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Gamma}_\alpha^\top \\ \boldsymbol{\Gamma}_\beta^\top \\ \boldsymbol{\Gamma}_\tau^\top \end{bmatrix} = \begin{bmatrix} \gamma_{\alpha1} & \gamma_{\alpha2} & \dots & \gamma_{\alpha r} \\ \gamma_{\beta1} & \gamma_{\beta2} & \dots & \gamma_{\beta r} \\ \gamma_{\tau1} & \gamma_{\tau2} & \dots & \gamma_{\tau r} \end{bmatrix}, \quad (5)$$

where we get three different mode shapes per eigenvector in the reduced basis solving Eq. (2), since we model deformations in a 3D space. Note that every $\boldsymbol{\Gamma}$ component is a

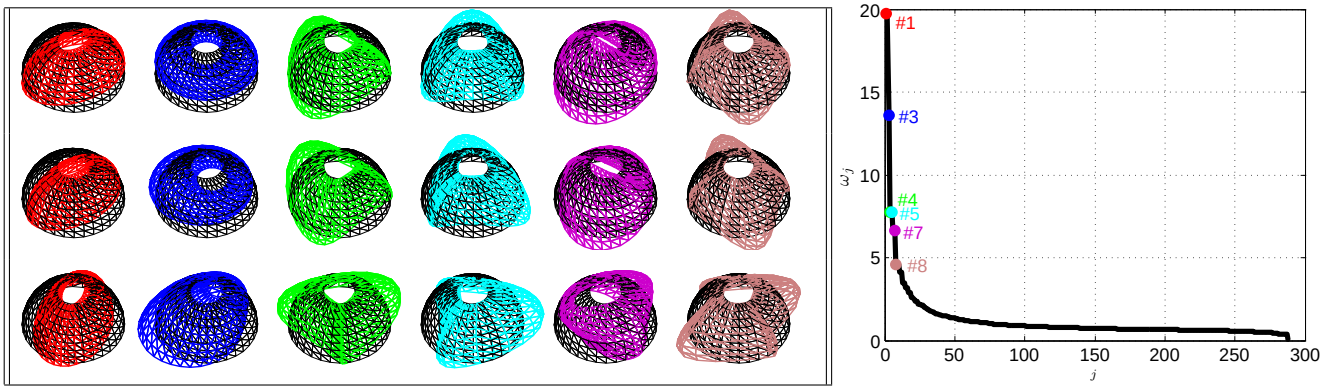

Figure 2. **Mode-shape interpretation. Left:** Three interpretations of the contribution of the some modes of the proposed reduced shape basis $\boldsymbol{\Upsilon}$ over a rest shape (black mesh). Each column corresponds to a different eigenvector, and each row, to its specific interpretation: $\boldsymbol{\Gamma}_\alpha$, $\boldsymbol{\Gamma}_\beta$ and $\boldsymbol{\Gamma}_\tau$. That is, every deformation is generated by setting one specific element of $\boldsymbol{\Lambda}$ to an arbitrary positive weight, and all the rest to zero. **Right:** Eigenvalue frequency spectrum of $\bar{\mathbf{D}}$, computed using Euclidean distances: we show eigenvalues ω_j in decreasing magnitude. Note how the energy drops until the last one, that is zero up to numerical precision. Best viewed in color.

$r \times 1$ vector that corresponds to a different interpretation of the reduced basis.

Since the principal directions in which our data varies are not normally aligned with the global axis system, we have to transform the computed eigenvectors before adding them to the shape at rest. First, we obtain a 3×3 covariance matrix $\boldsymbol{\Xi}$ as:

$$\boldsymbol{\Xi} = \Big(\bar{\mathbf{S}} - \big(\bar{\mathbf{s}}_* \otimes \mathbf{1}_p^\top\big)\Big)\Big(\bar{\mathbf{S}} - \big(\bar{\mathbf{s}}_* \otimes \mathbf{1}_p^\top\big)\Big)^\top, \qquad (6)$$

where $\bar{\mathbf{s}}_*$ is the mean values vector of all the data points in the rest shape and $\otimes$ denotes the Kronecker product. After that, we compute a transformation matrix $\boldsymbol{\Phi}$ by stacking the three eigenvectors of $\boldsymbol{\Xi}$ together as columns, to transform the eigenvectors to the global system. To obtain $\boldsymbol{\Phi}$, we use the properties of the orthogonal matrices, i.e., $\boldsymbol{\Phi} \equiv \boldsymbol{\Phi}^{-\top}$.

To analyze the computed eigenvectors of $\bar{\mathbf{D}}$, we sort them in a frequency spectrum from higher to lower frequency (see Fig. 2(right)). We observe that most deformation energy is in the eigenvectors with higher frequency, and these dominate the global deformation. Therefore the largest eigenvalues of $\bar{\mathbf{D}}$ contribute the most to the variance, and justifies the fact of using a low dimensional surface representation with only the first eigenvectors. Therefore, in practice it is not necessary to solve the full eigenvalue problem in Eq. (2), and only first r eigenvectors have to be computed, leading to a reduced computational cost. In Fig. 2(left) we plot the interpretations the some mode shapes, for a shape at rest corresponding to a hemisphere with hole.

In Table 1 we provide a qualitative comparison against other techniques that make use of shape bases. Similar to MA techniques [2, 4, 30], we only need to estimate the resting shape instead of using non-rigid 3D training data –with deformation– like PCA-based [27, 33] methods. However, our method does not need a deformation model to compute

Qua. / *Met.*	Training	Model	Accurate	Complexity
PCA	$\mathcal{X}$		✓	$(3p)^2r$
MA		$\mathcal{X}$	✓	$(3p)^2r$
Ours			✓	p^2r

Table 1. **Shape-basis techniques comparison.** We provide a comparison with other methods to obtain a shape-basis family. We consider learning methods such as PCA, physics-based ones such as MA and our interpreted model. While PCA-based methods can become very accurate if appropriate training data is available, in practice this is not often the case. Physics-based methods do not need training data, but a deformation model is mandatory to define the behavior. In contrast, our interpreted model needs neither training data nor a deformation model. In addition, since our shape basis is computed by solving a p-order eigenvalue problem, instead of a $3p$-order, our technique is more efficient. We represent the computational cost by a function $f(p, r)$ with p and r the number of points and modes, respectively. We show strong (✓) and weak ($\mathcal{X}$) qualities.

stiffness and mass matrices where material properties (such as the Poisson's ratio) are known, reducing the amount of prior knowledge. Regarding complexity, our method solves a p-order eigenvalue problem instead of a $3p$-order such as PCA or MA, and hence the memory requirements are much smaller. This is an important advantage for real applications with limited computational resources [40]. As a consequence, our method reduces the complexity from $f(p, r) = 9p^2r$ to $f(p, r) = p^2r$ [19] to solve the eigenvalue problem in Eq. (2).

We observe that our model is equivalent to the 3D-implicit low-rank shape model proposed in [28], where the shape basis is also represented by p-dimensional vectors. However, we just use a distance matrix to compute the predefined reduced shape basis. Hence, our problem is bilinear instead of trilinear –including the camera motion– reducing

thus the number of parameters to estimate. This means that our method is able to make the most of the available information, since both formulations use exactly the same initialization. In any case, since our reduced shape basis is predefined from the shape at rest, we can also define an orthogonal shape basis as $\mathbf{B}_\iota = \mathbf{\Phi}\mathbf{\Lambda}_\iota^* \mathbf{\Upsilon} \in \mathbb{R}^{3\times p}$ with $1 \leq \iota \leq 3r$, where the $\mathbf{\Lambda}_\iota^*$ matrix only contains a component different to 0, and hence obtaining $3r$ components of the linear subspace, if all combinations are considered. Finally, the 3D displacement could be computed as $\mathbf{U} = \Sigma_{\iota=1}^{3r} \phi_\iota \mathbf{B}_\iota$ with ϕ_ι the weight coefficients of the linear subspace of rank $3r$.

3.4. Physical Constraints

An interesting advantage of the proposed model is that we can associate the entries of the deformation transformation matrix $\mathbf{\Lambda}$ with physical behaviors. To do this, we can easily use prior knowledge about the deformation of an object to pre-define some of the entries in $\mathbf{\Lambda}$. For example, when we process sequences of non-rigid objects that cannot have stretching deformations –like a flag waving in the wind– the entries in $\mathbf{\Gamma}_\alpha$ and $\mathbf{\Gamma}_\beta$ can be directly set to zero, because the surface can not undergo in-plane deformations. On the other hand, if the object has no bending deformations –like an elastic hair ribbon with planar forces– the entries in $\mathbf{\Gamma}_\tau$ should be set to zero. For the general case, all entries in $\mathbf{\Lambda}$ should be considered. We observe that while the high-order bending modes can approximate better shape deformation, high-order stretching modes are very restrictive and they can model unrealistic shape deformations.

4. NRS*f*M with the Proposed Basis

Assuming the 3D shape $\mathbf{S}^f$ with p points is observed by a scaled orthographic camera, the projection $\mathbf{W}^f$ onto image frame f can be written as:

$$\mathbf{W}^f = \begin{bmatrix} u_1^f & u_2^f & \dots & u_p^f \\ v_1^f & v_2^f & \dots & v_p^f \end{bmatrix} = \mathbf{R}^f \mathbf{S}^f + \mathbf{T}^f, \quad (7)$$

where $\mathbf{R}^f$ is the truncated 2×3 rotation matrix (i.e., $\mathbf{R}^f \mathbf{R}^{f\top} = \mathbf{I}_2$) and $\mathbf{T}^f$ stacks p copies of a 2×1 translation vector $\mathbf{t}^f$. Our aim is to sequentially recover the camera motion $(\mathbf{R}^f, \mathbf{t}^f)$ and the 3D reconstruction of a deformable object $\mathbf{S}^f$ in every frame f from 2D point tracks $\mathbf{W}^f$ in a monocular image sequence. As our measurement matrix can have lost tracks due to occlusions or outliers, we define the binary vector $\mathbf{h}^f \in \{0,1\}^{p\times 1}$ that indicates absence or presence of entries in $\mathbf{W}^f$, respectively. We propose using the previously proposed deformation model to represent the non-rigidity of the object over time.

4.1. Interpreted Deformation Model

We approximate the deformable shape using a combination of mode shapes. Estimating the 3D coordinates of the deforming object at each frame f boils down to estimating the deformation matrix $\mathbf{\Lambda}^f$ in Eq. (3). Therefore we have to estimate $3r$ parameters for each frame to define the current configuration of the object. We can nevertheless reduce this number of parameters considering physical constraints as discussed in § 3.4.

In order to estimate $\mathbf{\Lambda}^f$, we will rewrite the orthographic projection in Eq. (7) at frame f as:

$$\mathbf{W}^f = \mathbf{R}^f \left(\bar{\mathbf{S}} + \mathbf{\Phi}\mathbf{\Lambda}^f \mathbf{\Upsilon} \right) + \mathbf{T}^f. \quad (8)$$

In the following section we detail the optimization process.

5. On-line Non-linear Optimization

In this section, we present our on-line approach to simultaneously retrieve the camera motion and the 3D reconstruction of deformable objects. We use a few initial frames –assuming a dominant rigid motion– to initialize and estimate the rest shape and motion by using a rigid factorization algorithm [26], a standard practice in sequential NRS*f*M [2, 28]. In our case, we also compute a distance matrix $\mathbf{D}$ to obtain a shape basis. Note that both $\mathbf{\Phi}$ and $\mathbf{\Upsilon}$ matrices are computed using the first frames as described in § 3. Therefore, our problem is reduced to the on-line estimation of the deformation matrix $\mathbf{\Lambda}^i$ and the camera motion $(\mathbf{R}^i, \mathbf{t}^i)$ per image. This means the estimation of just a few parameters per frame, which leads to a low computational cost system that may potentially run in real time. For obtaining an on-line estimation as the data arrives, we perform bundle adjustment on a temporal window of the last $\mathcal{W}$ frames, similar to [2, 6, 28]. Concretely, the model parameters are estimated by minimizing the following energy function $\mathcal{A}\left(\mathbf{R}^i, \mathbf{t}^i, \mathbf{\Lambda}^i\right)$ of all observed points over all frames in the current temporal window $\mathcal{W}$:

$$\mathcal{A} = \sum_{i=f-\mathcal{W}+1}^{f} \| \left(\mathbf{1}_2 \otimes \mathbf{h}^{i\top} \right) \odot \left(\mathbf{W}^i - \mathbf{R}^i \left(\bar{\mathbf{S}} + \mathbf{\Phi}\mathbf{\Lambda}^i \mathbf{\Upsilon} \right) - \mathbf{T}^i \right) \|_{\mathcal{F}}^2$$
$$+ \lambda_q \sum_{i=f-\mathcal{W}+2}^{f} \|\Delta \mathbf{q}^i\|_{\mathcal{F}}^2 + \lambda_t \sum_{i=f-\mathcal{W}+2}^{f} \|\Delta \mathbf{t}^i\|_{\mathcal{F}}^2 + \lambda_\gamma \sum_{i=f-\mathcal{W}+2}^{f} \|\Delta \mathbf{\Lambda}^i\|_{\mathcal{F}}^2$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm. The operator Δ is the variation $\Delta \mathbf{X}^i \equiv \mathbf{X}^i - \mathbf{X}^{i-1}$ over the variable $\mathbf{X}$. The rotation matrices $\mathbf{R}^i(\mathbf{q}^i)$ are internally parameterized using quaternions, which guarantees orthonormality. Our energy also includes a data term to penalize deviations of the image measurements and temporal smoothness priors to penalize strong variations in the model parameters whose influence is regulated by λ_q, λ_t and λ_γ. These weights were determined empirically and unchanged for all experiments. It is worth noting that our approach does not use additional inextensibility constraints, and hence extensible deformations can be recovered. We minimize the energy $\mathcal{A}\left(\mathbf{R}^i, \mathbf{t}^i, \mathbf{\Lambda}^i\right)$

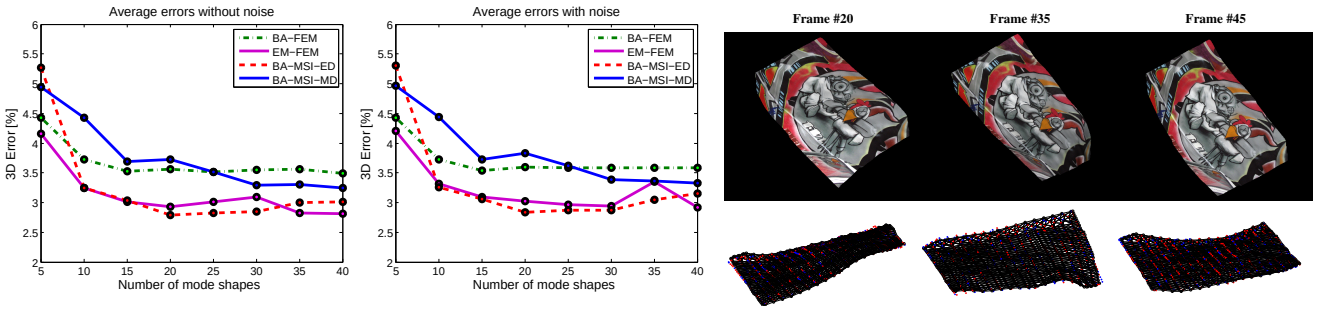

Figure 3. **Quantitative evaluation and comparison for Flag MoCap sequence. Left:** Error $e_{3\mathcal{D}}$ for sequential methods BA-FEM [2] and EM-FEM [4]; and for our methods BA-MSI-ED and BA-MSI-MD with varying number of mode shapes. We provide the performance for both noise-free and noisy observations. **Right:** Input frames and 3D reconstruction with red (observed) and blue dots (missing points). Black circles correspond to the 3D ground truth. Best viewed in color.

using sparse Levenberg-Marquardt. In order to initialize the model parameters for a new incoming frame, we simply set $\mathbf{\Lambda}^i \equiv \mathbf{\Lambda}^{i-1}$, $\mathbf{R}^i \equiv \mathbf{R}^{i-1}$ and $\mathbf{t}^i \equiv \mathbf{t}^{i-1}$.

6. Experimental Evaluation

We now present our evaluation on real monocular videos, providing both qualitative and quantitative results where we compare our method to state-of-the-art techniques based on low-rank models. For this comparison, we report the RMS error across all non-rigid frames n_f, which is defined as $e_{3\mathcal{D}} = \frac{1}{n_f}\sum_{i=1}^{n_f} \frac{\|\hat{\mathbf{S}}^i - \mathbf{S}^i_{GT}\|_{\mathcal{F}}}{\|\mathbf{S}^i_{GT}\|_{\mathcal{F}}}$ where $\hat{\mathbf{S}}^i$ is the 3D reconstruction and $\mathbf{S}^i_{GT}$ is the 3D ground truth. Our algorithms are denoted as BA-MSI-ED and BA-MSI-MD, using Euclidean and Manhattan distance matrices, respectively.

First, we evaluate our approach on a 594-point sequence of a waving *Flag*, provided by [4]. Since the deformation contains little stretching (it can also be successfully modeled using inextensibility, as shown in [39]), we can easily apply the physical constraints discussed in § 3.4 and set to zero the first two rows of the matrix $\mathbf{\Lambda}$. We compare the proposed approach with competing sequential methods based on low-rank models, BA-FEM [2] and EM-FEM [4]. The parameters of both methods were set in accordance to their original papers. For all cases, we exactly use the same initial exploration. We also present experiments with noisy measurements, adding a zero-mean Gaussian noise to every point in the object, with standard deviation $\sigma = 0.01\text{max}_j\{|d_e(j,\kappa)|\}$, where the κ-index corresponds to the centroid of all the points.

Figure 3(left) shows the consistent reduction of the error as more mode shapes are considered. We observe that BA-MSI-ED yields better results than BA-MSI-MD for this sequence, since the rest shape is quasi-planar and the points are sparsely distributed. This situation favors the modal shapes computed by Euclidean distances. Our BA-MSI-ED algorithm consistently outperforms BA-FEM [2], and performs comparably to EM-FEM [4] for both noise-free and noisy observations, with the additional advantage of not requiring a deformation model. In any event, our methods outperform the sequential algorithm SBA [28] with an error $e_{3\mathcal{D}}[\%]$ of 7.10(38); and batch algorithms MP [29], PTA [8], CSF2 [20] and EM-PND [25] with an error of 16.02(2), 14.11(2), 8.80(2) and 8.65, respectively. For low-rank methods, we show the basis rank (in brackets) that yielded the lowest error. Finally, we also represent our 3D reconstruction for a few frames in Fig. 3(right), where we randomly set a level of 40% missing points. For this case, our method is quite robust, with an error $e_{3\mathcal{D}} = 2.92\%$ when 20 mode shapes are used. In fact, our 3D reconstruction does not significantly degrade until a breaking point around 80% of missing data in the measurement matrix.

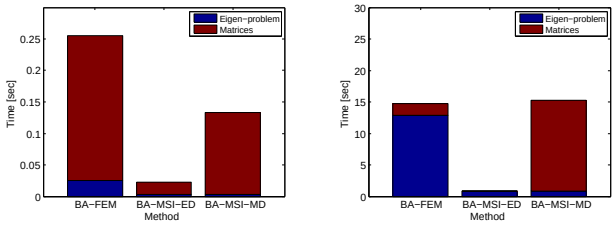

Figure 4. **Run-time comparison.** We show run-time to compute the shape basis for BA-FEM [2], and for our methods BA-MSI-ED and BA-MSI-MD. Note that since EM-FEM [4] also uses MA, the results are very close to BA-FEM [2]. For every case, we represent the computational cost to compute the matrices (in red) and to solve the eigen-value problem (in blue). **Left:** Talking face sequence of 56 points. **Right:** Flag sequence of 594 points.

Regarding computational cost, we analyze the run-time using non-optimized Matlab code to compute the shape basis, showing the matrices-computation complexity (stiffness/mass for BA-FEM [2], and distance matrices for our methods) and the solution of the eigen-value problem. Figure 4(right) summarizes these results. It can be seen that our methods have significantly lower computational cost than BA-FEM [2] to solve the eigen-value problem. Yet, while the time for computing the Euclidean matrix is almost negligible, the computation of the Manhattan matrix can become more expensive when the number of points increases

(we use an unoptimized Dijkstra's algorithm to obtain the distance matrix). In any event, the reduction of the computational cost using our BA-MSI-ED with respect to existing approaches is remarkable. It is worth pointing that an optimized implementation to solve the eigen-value problem would produce similar boosts in efficiency for every algorithm.

We also process a 100-frame real video showing a *Paper bending* and rotating, and provide a qualitative evaluation. We employ the semi-dense 828-point tracks from [2]. Since extensible deformations are not possible for this material, we impose the physical constraints on the matrix $\mathbf{\Lambda}$. Further, we consider a shape basis with $r = 30$ mode shapes, and use the metric of the BA-MSI-ED method. Figure 5 shows the reprojection of the deforming mesh into the image plane accurately describing the real 2D motion. We also display the recovered 3D reconstruction, retexturing the paper surface with a logo. Recall that this augmentation is performed on-line, upon the arrival of new frames.

We next test the *Talking face* sequence taken from a video of a man simultaneously talking and moving his head. We use 249 frames and 56 features tracks of the face. In this case, we use our BA-MSI-MD method –similar results can be obtained by BA-MSI-ED– with physical constraints that prevent pure-bending deformations. In Fig. 6 we show the reprojection of the deforming 3D mesh into the image plane and our 3D reconstruction for several views using $r = 30$ mode shapes. We also show the run-time for this sequence in Fig. 4(left).

Finally, we process a challenging *Laparoscopic* sequence of a beating heart captured during bypass surgery [18]. This shows the generality of our approach to recover the 3D reconstruction of extensible objects. In this case, since obtaining a priori knowledge over the type of deformation may be difficult, we optimize with BA-MSI-MD method the $3r(r = 10)$ parameters of the linear subspace without applying any physical constraint. Figure 7 shows some images and our 3D reconstruction for this semi-dense sequence of 3024 points.

7. Conclusion

In this paper we have shown how to exploit the distance information of a shape at rest to compute a shape basis that can model both inextensible and extensible deformations. We first compute a reduced shape basis at low computational cost by means of a spectral decomposition of the data, that we will interpret later. The obtained shape basis is then used to encode the time-varying shape, without a training step and in combination with simple regularization priors, in order to sequentially retrieve camera pose and deformable shape within a low-cost BA-based algorithm. Our claims have been experimentally validated on challenging real videos, showing accurate results obtained on-the-fly.

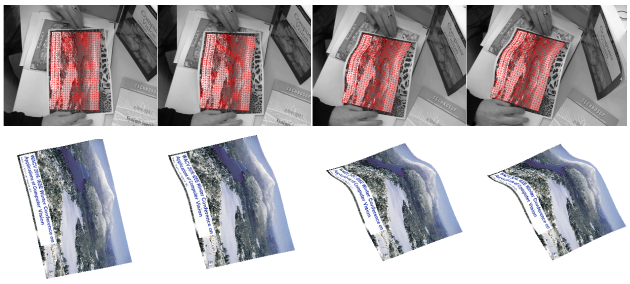

Figure 5. **Paper bending sequence. Top:** Images of a deforming piece of paper with reconstructed mesh. Notice how the 3D mesh is correctly projected and bent into the image. We also show our automatic retexturing of the paper sequence that is sequentially computed. **Bottom:** General view of the textured 3D reconstruction seen from a different viewpoint.

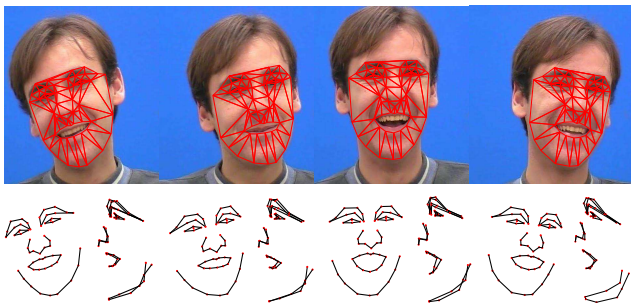

Figure 6. **Talking face sequence. Top:** Images of a face with reconstructed mesh. **Bottom:** Original viewpoint and side views of our 3D reconstruction.

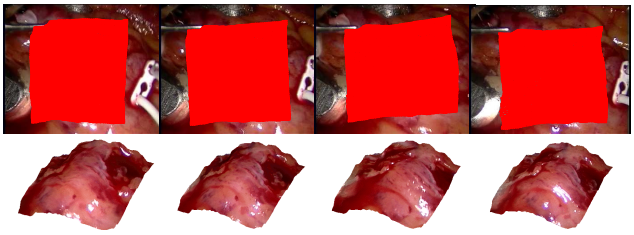

Figure 7. **Laparoscopic sequence. Top:** Images of a beating heart with reconstructed mesh. **Bottom:** Textured rendering of the recovered 3D reconstruction from a different viewpoint.

Regarding real-time capability, we consider that our method is as a suitable groundwork for augmented-reality applications in real time, and have shown an experiment along these lines. Further exploring this is part of our future work. An interesting avenue for future research is to try other dissimilarity measures such as Mahalanobis or chi-squared distances as well as the Jensen-Shannon divergence.

Acknowledgments

This work has been partially supported by the Spanish MCI under projects RobInstruct TIN2014-58178-R, SVMap DIP2012-32168 and Keratocono DPI2014-54981-R; by the ERA-net CHISTERA projects VISEN PCIN-2013-047 and I-DRESS PCIN-2015-147; and by a scholarship FPU12/04886 of the Spanish MECD.

References

[1] H. Abdi, A. J. O'Toole, D. Valentin, and B. Edelman. DISTATIS: The analysis of multiple distance matrices. In *CVPR*, 2005.

[2] 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.

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

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

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

[6] A. Agudo and F. Moreno-Noguer. Simultaneous pose and non-rigid shape with particle dynamics. In *CVPR*, 2015.

[7] A. Agudo, F. Moreno-Noguer, B. Calvo, and J. M. M. Montiel. Sequential non-rigid structure from motion using physical priors. *TPAMI, to appear*, 2016.

[8] 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.

[9] 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.

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

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

[12] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant 3D face recognition. In *AVBPA*, 2003.

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

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

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

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

[17] A. Elad and R. Kimmel. On bending invariant signatures for surfaces. *TPAMI*, 25(10):1285–1295, 2003.

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

[19] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Univ Pr, 1996.

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

[21] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[22] V. Jain and H. Zhang. A spectral approach to shape-based retrieval of articulated 3D models. *CAD*, 39(5):398–407, 2007.

[23] R. Kimmel and J. A. Sethian. Computing geodesics paths on manifolds. *Proc. Natl. Academy Sciences.*, 95(15):8431–8435, 1998.

[24] N. Kiryati and G. Szekely. Estimating shortest paths and minimal distances on digitized three-dimensional surfaces. *PR*, 26(11):1623–1637, 1993.

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

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

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

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

[29] 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.

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

[31] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *IJCV*, 95(2):124–137, 2011.

[32] M. Salzmann and P. Fua. *Deformable surface 3D reconstruction from monocular images*. Synthesis Lectures on Computer Vision, 2010.

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

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

[35] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2D and 3D pose estimation from a single image. In *CVPR*, 2013.

[36] E. Simo-Serra, A. Ramisa, G. Alenyá, C. Torras, and F. Moreno-Noguer. Single image 3D human pose estimation from noisy observations. In *CVPR*, 2012.

[37] D. Smeets, J. Hermans, D. Vandermeulen, and P. Suetens. Isometric deformation invariant 3D shape recognition. *PR*, 45(7):2817–2831, 2012.

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

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

[40] A. Wendel, M. Maurer, G. Graber, T. Pock, and H. Bischof. Dense reconstruction on-the-fly. In *CVPR*, 2012.

[41] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D active appearance models. In *CVPR*, 2004.

[42] J. Zhu, S. C. H. Hoi, and M. R. Lyu. Real-time non-rigid shape recovery via active appearance models for augmented reality. In *ECCV*, 2006.