A Bayesian Approach to Simultaneously Recover Camera Pose and Non-Rigid Shape from Monocular Images

Francesc Moreno-Noguer and Josep M. Porta

Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens Artigas 4-6, 08028, Barcelona, Spain

Abstract

In this paper we bring the tools of the Simultaneous Localization and Map Building (SLAM) problem from a rigid to a deformable domain and use them to simultaneously recover the 3D shape of non-rigid surfaces and the sequence of poses of a moving camera. Under the assumption that the surface shape may be represented as a weighted sum of deformation modes, we show that the problem of estimating the modal weights along with the camera poses, can be probabilistically formulated as a maximum a posteriori estimate and solved using an iterative least squares optimization. In addition, the probabilistic formulation we propose is very general and allows introducing different constraints without requiring any extra complexity. As a proof of concept, we show that local inextensibility constraints that prevent the surface from stretching can be easily integrated.

An extensive evaluation on synthetic and real data, demonstrates that our method has several advantages over current non-rigid shape from motion approaches. In particular, we show that our solution is robust to large amounts of noise and outliers and that it does not need to track points over the whole sequence nor to use an initialization close from the ground truth.

Key words: Deformable Surfaces, Pose Estimation, Bayesian Belief Networks, SLAM

Email address: {fmoreno,porta}@iri.upc.edu (Francesc Moreno-Noguer and Josep M. Porta).
URL: iri.upc.edu/people/{fmoreno,porta}/ (Francesc Moreno-Noguer and Josep M. Porta).

1 Introduction

Recovering the 3D shape of non-rigid objects from monocular image sequences is known to be a severely ill-conditioned problem because very different shape configurations may have a similar projection [14,35,38]. As shown in Fig. 1 the problem becomes even further underconstrained if the camera moves while the shape deforms, and both non-rigid shape and camera motion have to be simultaneously estimated. In order to resolve the inherent ambiguity between camera motion and shape deformation and turn the problem into a tractable one, prior knowledge about the object's behavior or the camera dynamics is then necessary.

Traditional approaches seek to reduce the space of possible shapes by introducing deformation models, either physically inspired ones [9,23,24,43,46] or learned from training data [6,7,8,10,17,22,28,25,30,38,39,50]. Surface deformations are then expressed as weighted combinations of modes, and estimating the shape entails at retrieving the weights of this linear combination by minimizing image based objective functions. However, since these objective functions are often complex, their convergence is only guaranteed if the shape is precisely initialized. In addition, most of these approaches either assume the pose of the camera to be known or retrieve the shape with no camera referential.

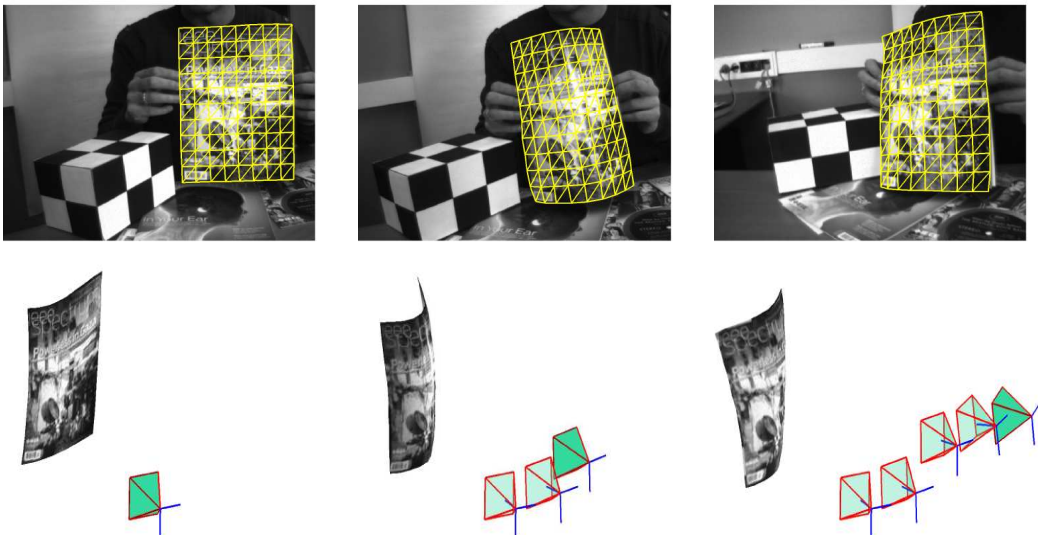

Fig. 1. Simultaneous estimation of non-rigid shape and camera pose from input images. **Top:** Three different frames of an input sequence with the reconstructed 3D mesh overlaid. **Bottom:** Re-textured side view of the retrieved surface and sample camera poses up to the current frame. Note that estimating the camera pose from only the observation of the deforming shape is very difficult even for the human eye. It would be much easier from the observation of the rigid objects, such as the calibration box, although we do not contemplate this case in the current paper.

Recent approaches in non-rigid structure-from-motion (NRSFM) have shown that deformation modes can be learned along with the shape and motion parameters [3,15,33,36,42,44,47,49]. Yet, while these techniques are especially interesting in situations where training data is hard to obtain, they typically require a number of points to be tracked throughout the whole sequence, which is difficult to satisfy in practice, especially when dealing with non-rigid objects that suffer from self occlusions. Furthermore, existing NRSFM approaches have shown to be effective only for relatively small deformations and they are quite sensitive to the presence of outliers and noisy observations.

In this paper, we propose a new formulation to the problem of simultaneously retrieving non-rigid 3D shape and camera motion that overcomes some of the limitations of previous approaches. We make two basic assumptions that are widely used in previous literature [29,35,37,38]. First, we assume that the deformation modes are available. And second, we assume that some 2D-to-3D correspondences can be established between the input images and a reference image in which the shape is already known. Yet, in contrast to NRSFM methods, we do not require tracking the points along the whole sequence, that is, each input image may have its independent set of matches. And most importantly, our method tolerates significant amounts of outliers and noise.

Our approach draws inspiration from a recent work on Simultaneous Localization and Map Building (SLAM) [13] used to estimate camera pose while mapping a rigid and static environment. We show that by appropriately parameterizing the shape and pose, the SLAM formulation can be extended to non-rigid domains. More specifically, we formulate the problem of estimating the modal weights describing non-rigid shapes and the pose parameters of the camera as a Maximum a Posteriori (MAP) estimate which can be iteratively solved using linearization and an efficient QR factorization for sparse linear systems [12]. As we will demonstrate through testing on both synthetic and real data, besides the robustness to outliers and noise, this formulation does not require a precise initialization, which is another remarkable step-forward when compared to the previous methodologies.

This work is an extended version of our earlier paper [27] where we already proposed the probabilistic framework to integrate parameters describing both the camera motion and the surface deformation. Here, we exploit the generality of this methodology and show that it allows introducing additional constraints. As a proof of concept we will show that enforcing local inextensibility naturally fits within our formulation, yielding better accuracies on the reconstructed shapes.

2 Related Work

3D surface reconstruction from monocular images has been an active research topic in Computer Vision for many years. Existing solutions may be roughly classified into those based on pre-defined or pre-learned deformation modes and those that learn the modes from input images and simultaneously retrieve shape and pose parameters.

The earliest works introduced physically-inspired deformation modes such as superquadrics [24], thin-plate splines [23] or balloons [9], used in combination with modal analysis [34] to reduce the degrees of freedom of the problem. Yet, all these approaches are only effective to capture relatively small deformations. More realistic deformations were described by complex non-linear models [4,46], although their applicability is limited to very specific materials.

This limitation has been addressed by methods that learn the deformation modes from training data, such as the Active Appearance and Shape Models [10,22] or the 3D Morphable Models [6]. These approaches represent surface deformations as linear combinations of rigid modes, and retrieving shape entails minimizing an image-based objective function. However, since this function is typically highly non-convex, it requires good pose and shape initializations to converge, which makes these methods appropriate for tracking shapes with a small inter-frame deformation, such faces [30,50]. In [17] a similar approach is used to detect human shape and pose from just a single image, although it requires manual pose initialization.

Several recent methods have been proposed to recover non-rigid shape from single images, by using deformation modes in conjunction with local rigidity constraints to reconstruct inextensible surfaces [14,28,35,37,38], and in conjunction with shading constraints to reconstruct stretchable surfaces [29]. However, none of these approaches retrieves the camera pose, and either assume that the deformation modes are aligned with the camera coordinate system or provide a solution shape for which the pose is unknown. One interesting exception is [39] which simultaneously retrieves point correspondences, pose and shape from one single image. Yet, in order to do so, it assumes hard prior constraints on the pose which may be difficult to hold in practice, and can only handle a reduced amount of points of interest.

Constraining the surface motion by linear models is also at the core of non-rigid structure-from-motion methods. Although the seminal work of Bregler *et al.* [8] used known deformation modes, current approaches [3,15,33,36,42,44,47,49] do not require to known them and, given a video sequence, they simultaneously compute modes, pose and shape. This generality, though, comes at the price of having to impose several constraints that are difficult to hold in prac-

tice, such as requiring a sufficient number of points to be tracked throughout the whole sequence. In addition, most of these methods have only been effectively used to retrieve relatively small deformations, and tend to be sensitive to noisy correspondences, missing data, and outliers. Recently, in [2,1], the strengths of both the NRSFM and the physic-based approaches based on Finite Elements are merged, yielding a system able to track the motion of the camera and estimate 3D shape of potentially extensible surfaces. Yet, these alternatives still lack an estimation of the absolute pose of the camera, and the deformations that can be modeled using FEMs are relatively mild.

To tackle these issues, we propose a SLAM-inspired solution. SLAM refers to the problem of localizing a robot or a camera in an unknown environment while simultaneously building a consistent map of this environment. Although dynamical aspects are sometimes considered [5,19,11], SLAM generally assumes a robot moving in a static and rigid environment, even in their more recent implementations [31]. In order to deal with the uncertainty, SLAM is formulated using a probabilistic MAP problem which was initially addressed as a filtering problem [20,40,41,48] or using iterative methods [16,32] closely related to bundle adjustment [45]. However, modern SLAM methods arise from the insights gained by viewing the problem as inference in a graphical model. This is the approach adopted in this paper. Using deformation modes and a formalism similar to the SLAM approach by [13], we show that the problem of simultaneously recovering pose and non-rigid shape can be formulated as a MAP estimate, where the joint probability of both the camera poses and the object deformations is maximized given a set of 3D-to-2D correspondences between each input image and a reference configuration. Furthermore, this solution is shown to have significant advantages in terms of robustness and convergence properties.

3 Simultaneous Pose and Non-Rigid Shape

The method we present in this paper combines the strengths of the mathematical framework used to solve SLAM in robotics [13], and the linear formulation of the non-rigid shape recovery problem proposed in recent works [28,38]. Given these main ingredients, we will next show that the problem of simultaneously recovering pose and non-rigid shape can be probabilistically formulated as a maximum a posteriori estimate, where both pose and shape are estimated given a set of 3D-to-2D correspondences. We then show that the solution can be iteratively approximated solving a sequence of linear least squares problems.

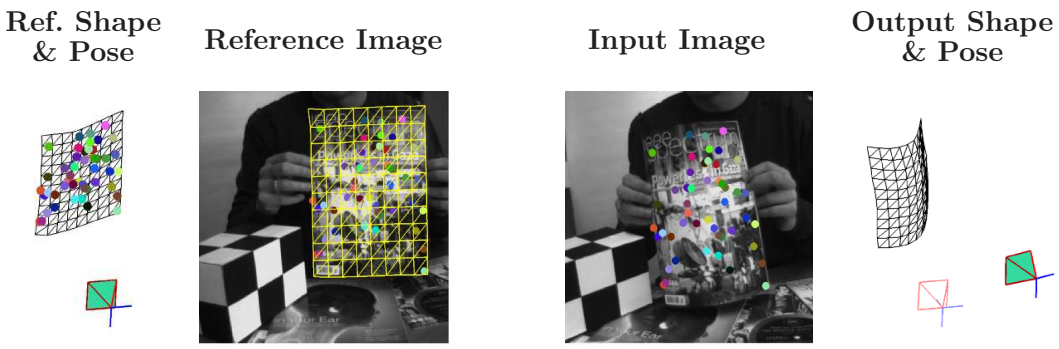

Fig. 2. **Problem Formulation.** We initially assume we are given a reference configuration in which the surface shape and camera pose are known and hence, the shape may be registered to a reference image. At running time, we seek to recover the pose of the moving camera and the shape of the deforming surface for each input image. Our approach first establishes point correspondences between each input image and the reference one. Note that these matches are in fact from 2D-to-3D, and are represented by points with the same color in the figures above. In addition, observe that these correspondences are difficult to establish because of the non-linear appearance deformations of the image, and thus, may contain gross errors and mismatches. Yet, as will be shown in the results section, our algorithm is robust to these artifacts and can handle large amounts of outliers.

3.1 Notations and Assumptions

We represent the surface as a triangulated 3D non-stretchable mesh with n_v vertices $\mathbf{v}_i$ concatenated in a $3n_v$ vector $\mathbf{x}=[\mathbf{v}_1^\top,\ldots,\mathbf{v}_{n_v}^\top]^\top$. The camera pose is represented as a 6-dimensional vector $\boldsymbol{\rho}$ parameterizing a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. Given a sequence of input images $\mathcal{I}=\{\mathbf{I}^k\}$, $1\leq k\leq n_I$, our goal is to estimate both the surface shape $\mathbf{x}^k$ and the camera pose $\boldsymbol{\rho}^k$ at each time instant k.

As shown in Figure 2, we assume that we are given a set of 3D points $\mathcal{R}^{ref}=\{\mathbf{r}_i\}$, $1\leq i\leq n_r$ on a reference configuration $\mathbf{x}^{ref}$, and that for each input image, we know $n_c^k\leq n_r$ 3D-to-2D correspondences between a subset of points of $\mathcal{R}^{ref}$ and a set of 2D points $\mathcal{U}^k=\{\mathbf{u}_i^k\}$ on $\mathbf{I}^k$.

Additionally, in order to reduce the dimensionality of the problem, we model the plausible surface deformations as a linear combination of a mean shape $\mathbf{x}_0$ and n_m deformation modes $\mathbf{Q}=[\mathbf{q}_1,\ldots,\mathbf{q}_{n_m}]$

$$\mathbf{x}^k=\mathbf{x}_0+\sum_{i=1}^{n_m}\alpha_i^k\,\mathbf{q}_i=\mathbf{x}_0+\mathbf{Q}\,\boldsymbol{\alpha}^k\;, \tag{1}$$

Geometric Parameters	
$\mathbf{x}$	Object shape
$\boldsymbol{\rho}$	Camera pose
$\mathbf{v}$	Mesh vertex
$\mathbf{A}$	Camera calibration matrix
$\mathbf{R}$	Rotation matrix
$\mathbf{t}$	Translation vector
$\mathbf{p}$	3D point on the input surface
$\mathbf{r}$	3D point on the reference surface
$\mathcal{R}^{ref}$	Set of reference points
$\mathbf{I}$	Input image
$\mathbf{u}$	2D projection of a point $\mathbf{p}$
$\mathcal{U}$	Set of 2D points on the images
$\mathbf{x}_0$	Mean shape vector
$\mathbf{Q}$	Deformation modes matrix
$\boldsymbol{\alpha}$	Modal weights vector
n_I	Number of input images
n_v	Number of mesh vertices
n_r	Number of 3D points in $\mathcal{R}^{ref}$
n_c	Number of 3D to 2D matches
n_m	Number of deformation modes
n_e	Number of edges in the mesh

Stochastic Parameters	
$\boldsymbol{\phi}$	State vector
$f(\cdot)$	Process model
$\mathbf{F}$	Jacobian of the process model
$\mathbf{w}_{\boldsymbol{\phi}}$	Process noise
$\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$	Process noise covariance matrix
$h(\cdot)$	Measurement model
$\mathbf{H}$	Jacobian of the measurement model
$\mathbf{w}_{\mathbf{u}}$	Measurement noise
$\boldsymbol{\Sigma}_{\mathbf{u}}$	Measurement noise covariance matrix
$l(\cdot)$	Edge length estimation function
σ_l	Variance in the estimated length

Table 1
Notation used in this paper. When necessary, superindices are used to refer to the parameters for a particular frame.

where $\boldsymbol{\alpha}^k = [\alpha_1^k, \ldots, \alpha_{n_m}^k]^\top$ are unknown modal weights that define the surface shape at time k. We obtain these modes by applying Principal Component Analysis over a training set of meshes which undergo similar deformations as those considered during the test.

Finally, we also assume the camera to be calibrated and denote by $\mathbf{A}$ its 3×3 matrix of intrinsic parameters. Table 1 summarizes of the notation we use.

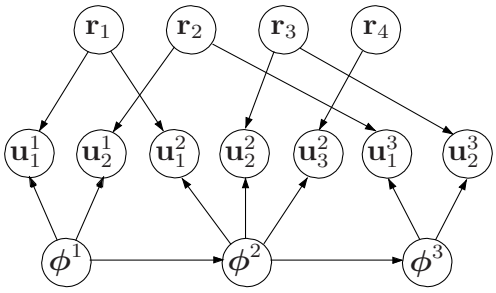

Fig. 3. Representation of the problem to solve in the form of a Bayesian belief network, for a small problem with three states $\boldsymbol{\phi}^k$, and four reference points $\mathbf{r}_i$.

3.2 Probabilistic Formulation of the Problem

Let $\boldsymbol{\phi}^k = [\boldsymbol{\rho}^{k\top}, \boldsymbol{\alpha}^{k\top}]^\top$ be the augmented $(6 + n_m)$-state vector that collects the unknown pose and shape at frame k, where the shape is represented by means of the modal weights. We can then formulate our problem as that of estimating $\Phi = \{\boldsymbol{\phi}^k\}$ given the observations $\mathcal{U} = \{\mathcal{U}^k\}$, with $1 \leq k \leq n_I$. This can be expressed in terms of the following maximum a posteriori estimate

$$\Phi^* = \arg\max_{\Phi} P(\Phi \,|\, \mathcal{U}) \propto \arg\max_{\Phi} P(\Phi, \mathcal{U}) \,, \tag{2}$$

where for the second step we have assumed a uniform distribution of $P(\mathcal{U})$.

The joint probability $P(\Phi, \mathcal{U})$ may be written as the product of a motion component describing the camera dynamics and a measurement component relating the observations with the configuration of the state vector for each frame k

$$P(\Phi, \mathcal{U}) = P(\boldsymbol{\phi}^1) \prod_{k=2}^{n_I} P(\boldsymbol{\phi}^k | \boldsymbol{\phi}^{k-1}) \prod_{i=1}^{n_c^k} P(\mathbf{u}_i^k | \boldsymbol{\phi}^k, \mathbf{r}_{i_k}) \tag{3}$$

where $P(\boldsymbol{\phi}^1)$ is a prior on the initial pose and shape, $P(\boldsymbol{\phi}^k | \boldsymbol{\phi}^{k-1})$ is the dynamic model, and $P(\mathbf{u}_i^k | \boldsymbol{\phi}^k, \mathbf{r}_{i_k})$ is the measurement model of the 3D reference point $\mathbf{r}_{i_k}$ corresponding the i-th 3D-to-2D match at time step k. The dependencies among variables of this joint probability can be represented as a belief network. For instance, Figure 3, shows a simple case with a sequence of three images, an thus three state vectors, and four reference points.

We define the dynamic model for the camera and modal weights as a stochastic process $\boldsymbol{\phi}^k = f(\boldsymbol{\phi}^{k-1}) + \mathbf{w}_{\boldsymbol{\phi}}^k$, which may be probabilistically written as

$$P(\boldsymbol{\phi}^k | \boldsymbol{\phi}^{k-1}) \propto \exp -||f(\boldsymbol{\phi}^{k-1}) - \boldsymbol{\phi}^k||^2_{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^k} \,, \tag{4}$$

where $||\cdot||^2_{\boldsymbol{\Sigma}}$ denotes the squared Mahalanobis distance, $f(\cdot)$ is the process

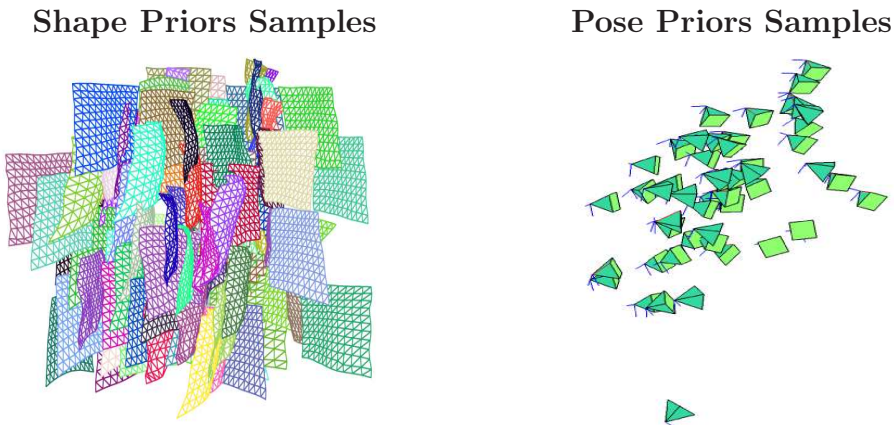

Fig. 4. Shape and pose priors samples we consider, obtained by adding Gaussian noise to a given shape and pose. Note that the priors we use are fairly ambiguous and allow representing many different configurations of shapes and poses.

model, and $\mathbf{w}_{\boldsymbol{\phi}}^k$ is a zero mean Gaussian noise with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^k$. In fact, this covariance is a block diagonal matrix, composed of a 6×6 covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\rho}}^k$ for the poses, and a $n_m \times n_m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^k$ for the modal weights. In practice, we set these covariance matrices to relatively large values in order to be able to produce a large variety of shapes and poses. This increases the generality of our approach to solve problems where input data may considerably differ from the training data we used to compute the modes. For instance, Figure 4 shows the kind of different shapes and poses that may be generated by picking random samples of a Gaussian distribution with the covariance matrices we consider in our experiments.

The measurement model is described in the form $\mathbf{u}_i^k = h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) + \mathbf{w}_{\mathbf{u}}^k$, which as above can be probabilistically expressed as

$$P(\mathbf{u}_i^k | \boldsymbol{\phi}^k, \mathbf{r}_{i_k}) \propto \exp -||h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) - \mathbf{u}_i^k||^2_{\boldsymbol{\Sigma}_{\mathbf{u}}^k} , \tag{5}$$

where $\mathbf{u}_i^k$ is the known 2D location of the 3D reference point $\mathbf{r}_{i_k}$, $h(\cdot)$ is the measurement function, and $\mathbf{w}_{\mathbf{u}}^k$ is a zero mean Gaussian noise with 2×2 covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}}^k$. The function $h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})$, corresponds to the equation that projects $\mathbf{r}_{i_k}$ onto the image, after being mapped according to the pose $\boldsymbol{\rho}^k$ and shape parameters $\boldsymbol{\alpha}^k$.

More specifically, let $\mathbf{p}_i^k$ be a point on the mesh $\mathbf{x}^k$ corresponding to the point $\mathbf{r}_{i_k}$ in the reference configuration. We can write $\mathbf{p}_i^k$ in terms of the barycentric coordinates of the face it belongs as

$$\mathbf{p}_i^k = \sum_{j=1}^{3} a_{ij} \mathbf{v}_{i_j}^k , \tag{6}$$

where the a_{ij} are the barycentric coordinates and $\mathbf{v}_{i_j}^k$ are the vertices of the face in $\mathbf{x}^k$ containing the point $\mathbf{p}_i^k$. Since we assume the mesh does not stretch, these barycentric coordinates remain constant for each point and can be easily computed from points $\mathbf{r}_{i_k}$ and the reference mesh $\mathbf{x}^{ref}$.

The measurement equation $h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})$ returns $\tilde{\mathbf{u}}_i^k$, the 2D projection of $\mathbf{p}_i^k$ onto the image given the current pose and shape parameters. If we expand $\boldsymbol{\phi}^k$ into a rotation matrix $\mathbf{R}^k$, translation vector $\mathbf{t}^k$, and modal weights $\boldsymbol{\alpha}^k$ we can write such a projection as

$$w_i \begin{bmatrix} \tilde{\mathbf{u}}_i^k \\ 1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathbf{R}^k | \mathbf{t}^k \end{bmatrix} \begin{bmatrix} \mathbf{p}_i^k \\ 1 \end{bmatrix}, \tag{7}$$

where w_i is a projective scalar. Finally, by injecting the barycentric coordinates of Equation (6) and the modal description of Equation (1), the measurement equation $\tilde{\mathbf{u}}_i^k = h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})$ can be written in terms of the pose parameters and modal weights

$$w_i \begin{bmatrix} \tilde{\mathbf{u}}_i^k \\ 1 \end{bmatrix} = \mathbf{A}\mathbf{R}^k \sum_{j=1}^{3} a_{ij} \left(\mathbf{x}_{0i_j} + \mathbf{Q}_{i_j} \boldsymbol{\alpha}^k \right) + \mathbf{A}\mathbf{t}^k \,, \tag{8}$$

where $\mathbf{x}_{0i_j}$ and $\mathbf{Q}_{i_j}$ are the subvectors of $\mathbf{x}_0$ and $\mathbf{Q}$ corresponding to the coordinates of the vertex $\mathbf{v}_{i_j}$.

3.3 Least Squares Formulation

Equation (2) expresses the problem of simultaneously retrieving pose and shape as a MAP estimate of the joint probability $P(\Phi, \mathcal{U})$. We next show that this can be iteratively solved as a simple Least Squares problem.

By taking the negative logarithm of Equation (2), and considering Equations (3), (4) and (5), the MAP problem may be reduced to the following non-linear least-squares estimation

$$\Phi^* = \arg\min_{\Phi} \sum_{k=1}^{n_I} \varepsilon_{\text{mot}}(\boldsymbol{\phi}^{k-1}, \boldsymbol{\phi}^k) + \varepsilon_{\text{meas}}(\boldsymbol{\phi}^k), \tag{9}$$

where

$$\varepsilon_{\text{mot}}(\boldsymbol{\phi}^{k-1}, \boldsymbol{\phi}^k) = \left\| f(\boldsymbol{\phi}^{k-1}) - \boldsymbol{\phi}^k \right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^k} \tag{10}$$

is the dynamic estimation component of the error, for which we define $f(\boldsymbol{\phi}^0) = \boldsymbol{\phi}^1$, and

$$\varepsilon_{\text{meas}}(\boldsymbol{\phi}^k) = \sum_{i=1}^{n_c^k} \left\| h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) - \mathbf{u}_i^k \right\|^2_{\boldsymbol{\Sigma}_{\mathbf{u}}^k} \tag{11}$$

is the measurement error component, that is, the sum of squared distances between the predicted location of the reference points in the image and their true position.

Since the measurement function $h(\cdot)$ is nonlinear and the process function $f(\cdot)$ may also be non-linear, the minimum of Equation (9) is iteratively approximated by linearizing the dynamic and measurement terms. Let us assume that $\mathbf{\Phi}_0 = [\boldsymbol{\phi}_0^{1\top}, \ldots, \boldsymbol{\phi}_0^{n_I\top}]^\top$ is a given estimation of Φ^*. We then approximate $f(\cdot)$ and $h(\cdot)$ linearizing at $\mathbf{\Phi}_0$

$$f(\boldsymbol{\phi}^{k-1}) \approx f(\boldsymbol{\phi}_0^{k-1}) + \mathbf{F}^{k-1}\boldsymbol{\delta}^{k-1}, \tag{12}$$

$$h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) \approx h(\boldsymbol{\phi}_0^k, \mathbf{r}_{i_k}) + \mathbf{H}_{i_k}^k \boldsymbol{\delta}^k, \tag{13}$$

where $\boldsymbol{\delta}^k = \boldsymbol{\phi}^k - \boldsymbol{\phi}_0^k$ is the update term, $\mathbf{F}^{k-1}$ is the $(6+n_m)\times(6+n_m)$ Jacobian of $f(\cdot)$, and $\mathbf{H}_{i_k}^k$ is the $2\times(6+n_m)$ Jacobian matrix of $h(\cdot)$, both of them evaluated at the corresponding element of $\mathbf{\Phi}_0$

$$\mathbf{F}^{k-1} = \left.\frac{\partial f(\boldsymbol{\phi}^{k-1})}{\partial \boldsymbol{\phi}^{k-1}}\right|_{\boldsymbol{\phi}_0^{k-1}} \tag{14}$$

$$\mathbf{H}_{i_k}^k = \left.\frac{\partial h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})}{\partial \boldsymbol{\phi}^k}\right|_{\boldsymbol{\phi}_0^k} \tag{15}$$

If we denote the error in the dynamic prediction by $\mathbf{c}^k = \boldsymbol{\phi}_0^k - f(\boldsymbol{\phi}_0^{k-1})$, and the error in the measurement by $\mathbf{d}_i^k = \mathbf{u}_i^k - h(\boldsymbol{\phi}_0^k, \mathbf{r}_{i_k})$, Equation (9) becomes

$$\begin{aligned}\varepsilon(\boldsymbol{\delta}^{k-1}, \boldsymbol{\delta}^k) \approx & \left\|\mathbf{F}^{k-1}\boldsymbol{\delta}^{k-1} - \mathbf{G}\boldsymbol{\delta}^k - \mathbf{c}^k\right\|^2_{\mathbf{\Sigma}_{\boldsymbol{\phi}}^k} \\ & + \sum_{i=1}^{n_c^k} \left\|\mathbf{H}_{i_k}^k \boldsymbol{\delta}^k - \mathbf{d}_i^k\right\|^2_{\mathbf{\Sigma}_{\mathbf{u}}^k},\end{aligned} \tag{16}$$

where $\mathbf{G}$ is a $(6+n_m)\times(6+n_m)$ identity matrix, introduced to simplify subsequent notation. Finally, the original least-squares problem is re-written as

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \|\mathbf{B}\,\boldsymbol{\delta} - \mathbf{b}\|^2_{\mathbf{\Sigma}}, \tag{17}$$

where $\boldsymbol{\delta} = [\boldsymbol{\delta}^{1\top}, \ldots, \boldsymbol{\delta}^{n_I\top}]^\top$, and $\mathbf{\Sigma}$ is a matrix made of all the $\mathbf{\Sigma}_{\boldsymbol{\phi}}^k$ and $\mathbf{\Sigma}_{\mathbf{u}}^k$ noise terms. The matrix $\mathbf{B}$ collects all Jacobian matrices and the vector $\mathbf{b}$ is made of all errors in dynamic and measurements predictions. Their structure is as follows:

$$\mathbf{B} = \begin{bmatrix} \mathbf{G} & & & & \\ \mathbf{F}^1 & \mathbf{G} & & & \\ & \mathbf{F}^2 & \mathbf{G} & & \\ & & \ddots & \ddots & \\ & & & \mathbf{F}^{n_I-1} & \mathbf{G} \\ \mathbf{J}^1 & & & & \\ & \mathbf{J}^2 & & & \\ & & & \ddots & \\ & & & & \mathbf{J}^{n_I} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \mathbf{c}^1 \\ \mathbf{c}^2 \\ \mathbf{c}^3 \\ \vdots \\ \mathbf{c}^{n_I} \\ \mathbf{d}^1 \\ \mathbf{d}^2 \\ \vdots \\ \mathbf{d}^{n_I} \end{bmatrix} \tag{18}$$

where we defined the matrix $\mathbf{J}^k = [\mathbf{H}_1^k, \cdots, \mathbf{H}_{n_c^k}^k]^\top$ and the vector $\mathbf{d}^k = [\mathbf{d}_1^{k\top}, \ldots, \mathbf{d}_{n_c^k}^{k\top}]^\top$.

In order to solve Equation (17) we differentiate with respect to $\boldsymbol{\delta}$ and equate to zero. Then, $\boldsymbol{\delta}^*$ may be found as the solution of

$$\hat{\mathbf{B}}^\top \hat{\mathbf{B}}\, \boldsymbol{\delta}^* = \hat{\mathbf{B}}^\top \hat{\mathbf{b}}, \tag{19}$$

with $\hat{\mathbf{B}} = \boldsymbol{\Sigma}^{-1/2}\, \mathbf{B}$ and $\hat{\mathbf{b}} = \boldsymbol{\Sigma}^{-1/2}\, \mathbf{b}$.

Note that $\mathbf{B}$ is a large but very sparse matrix. Assuming a constant number n_c of 3D-to-2D correspondences per image, $\mathbf{B}$ would be a $n_I(6+n_m+2n_c) \times n_I(6+n_m)$ matrix. Typical values in our experiments are $n_I = 50$ images, we detect about $n_c = 100$ correspondences per image, and use $n_m = 30$ deformation modes. These values would yield a 11800×1800 matrix, although with only a 2.3% of non-null entries. Taking advantage of this sparsity, $\hat{\mathbf{B}}$ and $\hat{\mathbf{b}}$ can be directly defined pre-multiplying $\mathbf{F}^k$, $\mathbf{G}$, and $\mathbf{c}_i^k$ by $(\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^k)^{-1/2}$, and $\mathbf{H}_i^k$ and $\mathbf{d}_i^k$ by $(\boldsymbol{\Sigma}_{\mathbf{u}}^k)^{-1/2}$. The resulting $\hat{\mathbf{B}}$ is also sparse and we can use a sparse QR factorization [12] on $\hat{\mathbf{B}}$ to efficiently obtain $\boldsymbol{\delta}^*$ without explicitly having to compute $\hat{\mathbf{B}}^\top \hat{\mathbf{B}}$.

The solution $\boldsymbol{\delta}^*$ is then used to update the estimation $\boldsymbol{\Phi}_0$ and the procedure described in this Section is repeated until convergence.

4 Improving the Robustness

In this section we will show how to improve the robustness of the proposed method along two directions. First, we will show that imposing inextensibility constraints on the retrieved surface, nicely fits within our framework. This is in contrast to the formulation proposed in [38,29] for which introducing these kind of constraints causes an explosion in the complexity of the system. In the second place, we propose a simple extra stage that allows to iteratively detect and remove 3D-to-2D correspondences with very large errors.

4.1 Inextensibility Constraints

Although representing the surface shape as a linear combination of modes and imposing temporal coherence will in general provide meaningful solutions, introducing additional constraints may accelerate the convergence and even provide more accurate results. One typical constrain used in previous works [28,38] consists in preventing the surface from stretching by preserving distances in local neighborhoods. Note that by doing this, any possible inaccuracy in the barycentric coordinates is minimized and, thus, there is a reduction in the overall error.

Inextensibility constraints may be readily introduced within our formulation by considering in Equation (3) an additional term that measures length errors. Let $l(\boldsymbol{\phi}^k, i) = \|\mathbf{v}_{i_1}^k - \mathbf{v}_{i_2}^k\|$ be the length of the edge defined by the neighboring vertices $\mathbf{v}_{i_1}^k$ and $\mathbf{v}_{i_2}^k$ whose coordinates are computed from the modal weights and modes using Equation (1). If we denote by l_i^{ref} the original length of the edge on the reference mesh and by n_e the number of edges of the mesh, we can expand Equation (9) with an additional multiplying term enforcing inextensibility constraints,

$$\prod_{i=1}^{n_e} P(l(\boldsymbol{\phi}^k, i)|\boldsymbol{\phi}^k, l_i^{ref}) \; . \tag{20}$$

Defining this measurement function as a stochastic process with Gaussian noise, Equation (9) can then be rewritten as

$$\begin{aligned}\Phi^* = \arg\min_{\Phi} \sum_{k=1}^{n_I} & \varepsilon_{\text{mot}}(\boldsymbol{\phi}^{k-1}, \boldsymbol{\phi}^k) \\ & + \varepsilon_{\text{meas}}(\boldsymbol{\phi}^k) + \varepsilon_{\text{inext}}(\boldsymbol{\phi}^k),\end{aligned} \tag{21}$$

where ε_{mot} and $\varepsilon_{\text{meas}}$ were defined above and

$$\varepsilon_{\text{inext}}(\boldsymbol{\phi}^k) = \sum_{i=1}^{n_e} \left\| l(\boldsymbol{\phi}^k, i) - l_i^{ref} \right\|_{\sigma_l^2}^2 \; . \tag{22}$$

The variance σ_l^2 corresponds to the uncertainty of the Gaussian noise corrupting the function that computes the length of an edge.

In order to solve this optimization problem we follow the iterative least squares procedure described in Section 3.3, i.e, we linearize the function $l(\cdot)$ and turn the inextensibility constraints into $n_e \cdot n_I$ additional rows for the matrix $\mathbf{B}$ of Equation (17). The top-left graph of Figure 5 depicts the block structure of the matrix $\mathbf{B}$ after considering all of the constraints for a simple case with $n_I = 5$ images, $n_c = 20$ matches per image, $n_m = 15$ modes and $n_e = 30$ edges per mesh.

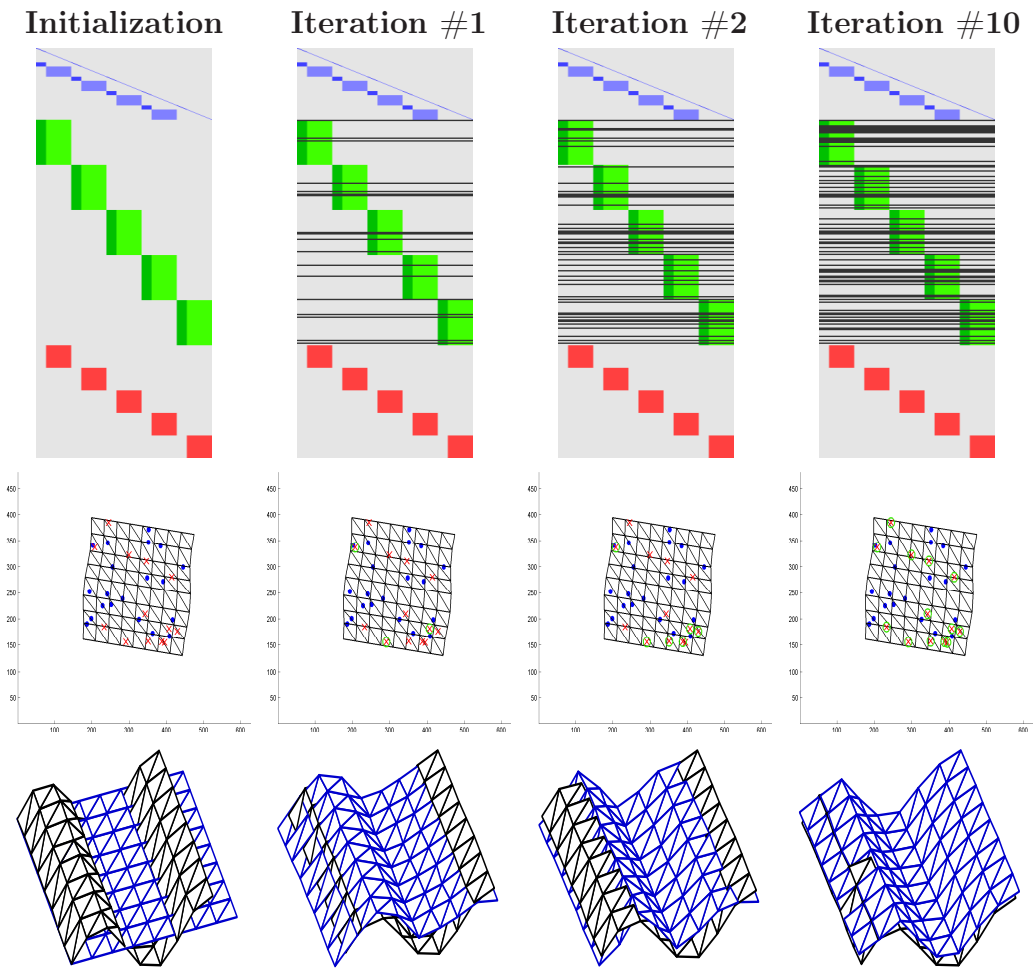

Fig. 5. Iterative fitting and outlier rejection. **Top:** Block structure of the matrix **B** for a simple case. The upper part of the matrix –blue blocks– correspond to the Jacobians of the dynamic models; the central part –green blocks– are the Jacobians of the matching constraints. In both cases, darker blocks are the constraints applied to the pose parameters and lightly colored blocks are applied on the shape parameters. The lower part of the matrix –red blocks– are the Jacobians that enforce intextensibility. Note how these Jacobians are only applied on the shape parameters. This is because length constraints only depend on the distances between vertices, and are independent of the pose. After each iteration our approach automatically detects and removes outliers from the computation. Each outlier correspondence corresponds to two rows on **B** which are black-colored on the graphs. As may be observed, we are able to handle large percentages of outliers. **Center:** Representation of the inlier –blue dots– and outlier –red crosses– correspondences for one sample frame of the input sequence. Note that we iteratively detect the outliers –green circles–. **Bottom:** Fitting the shape on one sample frame of the sequence. Our approach simultaneously resolves this fitting for several frames while computes the poses of the camera.

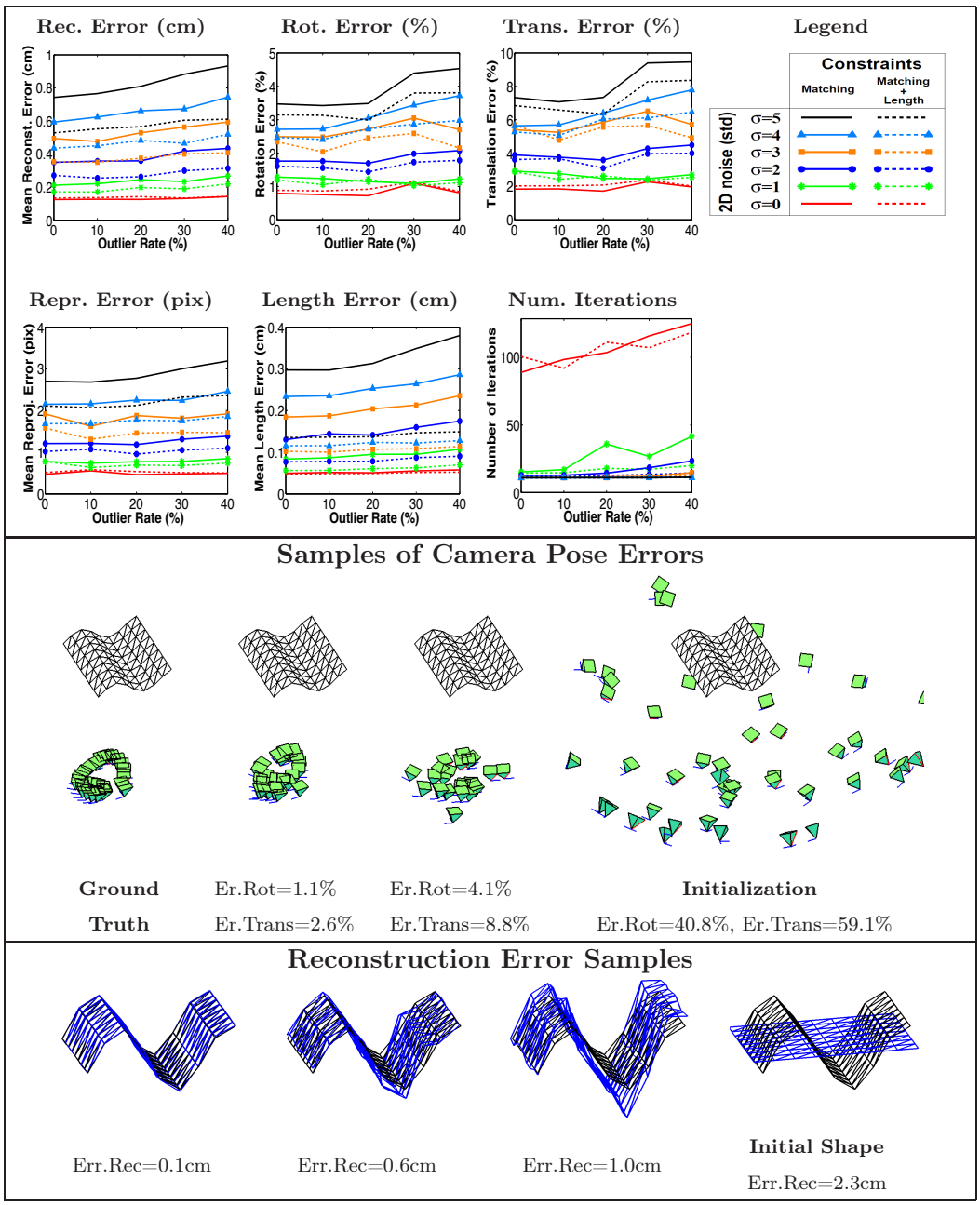

Fig. 6. Results on shape and pose recovery for a sequence of a deforming synthetic mesh. The upper graphs compare the accuracy on the shape and pose estimation when considering uniquely the constraints introduced by the 3D-to-2D matches or when considering additional inextensibility constraints. Dealing with these additional length constraints yields improved results while requiring a similar number of iterations to converge. The second and third rows, show different levels of error to give significance to the errors we obtain. Note that even with quite vague initializations, the proposed algorithm converges to reasonably good solutions.

4.2 Detecting and Removing Outliers

If the input correspondences are corrupted by outliers, the least squares solution we proposed in the previous section may become unreliable, as it simultaneously minimizes the reprojection error of all correspondences. In order to detect and remove outliers we have implemented a weighted least squares procedure that penalizes those matches with large residuals. More specifically, for each 3D-to-2D correspondence we define

$$\lambda_i^k = \frac{\mathbf{d}_i^k}{\text{median}(\mathbf{d}_i^k \,, 1 \leq k \leq n_I \,, 1 \leq i \leq n_c^k)} \,, \tag{23}$$

where $\mathbf{d}_i^k$ is the residual reprojection error of Equation (16). We then reduce the influence of the more noisy correspondences by multiplying the rows of $\mathbf{B}$ and $\mathbf{b}$ associated to the matching constraints with the weight

$$w_i^k = \begin{cases} \exp(-\lambda_i^k) & \text{if } \lambda_i^k < \lambda \\ 1 & \text{otherwise} \end{cases} \tag{24}$$

where the parameter λ is chosen large enough (we set $\lambda = 3$ in all our experiments) to ensure that only those measurements with large errors $\mathbf{d}_i^k$ are penalized. Yet, we initially do not remove these observations, because their gross error might come from a wrong estimate of shape and pose at the current iteration. Instead, following [37], we remove them if after having contributed in the current estimate, their reprojection error remains outside a radius, that is reduced at each iteration. In practice, we start with a 100 pixel radius that progressively reduce until a value of 10 pixels.

Figure 5 shows how outliers are iteratively detected. On the upper graphs, the black horizontal lines indicate the matching constraints that are removed after each iteration because they have large reprojection errors. These constraints correspond to matches that are classified as outliers, as seen in the middle-row figures. Usually, after a few iterations most of the outliers are detected, although the fitting procedure has not yet converged.

5 Experimental Results

In this section we evaluate the performance of our approach against noise in the correspondences, the presence of outliers, or its dependence on the quality of the initialization. We show results on both synthetic and real images. For the synthetic sequences we compare the results of our approach when uniquely considering the constraints introduced by the 3D-to-2D correspondences, and when considering them in conjunction with inextensibility constraints.

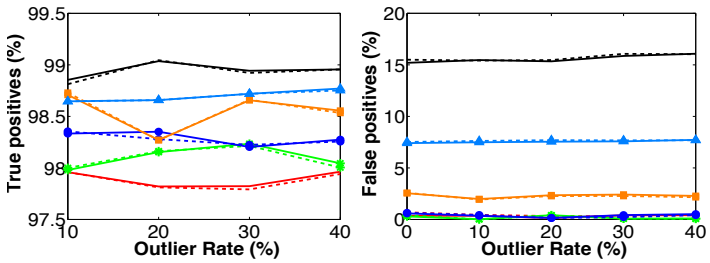

Fig. 7. Detection of outliers. P: true number of outliers. N: true number of inliers. TP: Number of outliers correctly detected. FP: Number of inliers misclassified as outliers. True Positives(%)=$\frac{TP}{P}$. False Positives(%)=$\frac{FP}{N}$. Observe that even for large levels of noise, our algorithm correctly detects most of the outliers. For the description of the colors and line-styles, we refer the reader to the legend on the top-right of Figure 6.

In addition, and in order to position the current approach within the state-of-the-art, we also provide a comparison against the approach proposed by [33], which is representative of the non-rigid shape from motion techniques. Although the two methods are not directly comparable, as they require from different assumptions, we enforce the comparison and show the benefits of using combined priors on the shape deformation and camera dynamics, even when they are very weak and initialized far from the ground truth solution.

5.1 Synthetic Data

We first applied our approach to a 50 frames synthetic sequence of a 9×9 mesh, simulating the deformation of a wave with increasing amplitude. The reference configuration was represented by a 30×30 cm planar shape. The camera was allowed to move according to random Brownian paths on the surface of a 80 cm sphere centered on the mesh, and with the optical axis pointing to the center of the sphere. The left-most graph in the middle row of Figure 6 shows one example of shape and camera poses generated this way.

For each pair of camera pose and mesh shape, we then synthetically produced 150 random 2D-to-3D correspondences, between a 640×480 image acquired for that particular shape and pose, and the reference configuration. Given this setup, we performed two different types of experiments, to evaluate both the robustness and convergence performance of the proposed algorithm.

5.1.1 Robustness to noise and outliers

In the first experiment we analyzed and compared the performance of our approach when using uniquely the constraints introduced by the 3D-to-2D matches, and when additionally enforcing inextensibility. We evaluated both situations against noise in the 2D correspondences and the presence of outliers. More specifically, we performed 10 different experiments by adding noise with standard deviation of $\{0, 1, 2, 3, 4, 5\}$ pixels, and by introducing a percentage of outliers of $\{0, 5, 10, 20, 30, 40\}\%$. In addition, this combination of parameters was repeated for 10 different random camera paths.

In each of these experiments all the shapes in the sequence were initialized with the reference mesh. The poses were initialized by adding random noise to the ground truth poses such that the initial percentage of rotation and translations errors were approximately of 50%. The right-most graphs in the second and third rows of Figure 6 show that these initializations are significantly far from the ground truth solutions.

In all experiments we used the same set of parameters to describe the dynamic and measurement models. As a dynamic model, we used simple Brownian motion, and the function $f(\cdot)$ in Equation (4) was taken to be the identity, that is, $f(\boldsymbol{\phi}) = \boldsymbol{\phi}$. The covariance matrix $\boldsymbol{\Sigma}_\rho$ for the poses was set to a constant diagonal matrix, with a 0.1 radiants of standard deviation for the rotation components, and 3 cm for the translational ones. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ for the modal weights was computed from the training data used to estimate the deformation modes, scaled by a factor of 3 to handle larger deformations and increase the generality of the method. The covariance $\boldsymbol{\Sigma}_{\mathbf{u}}$ of the measurement model, was set to a diagonal matrix with a 3 pixels standard deviation. When considering the additional length constraints, we set the corresponding standard deviation σ_l of Equation (22) to 0.5 cm.

Figure 6 reports the mean results of the experiment. In the upper plots, we depict accuracy of the two configurations of our approach as a function of the percentage of outliers and for different levels of noise in the correspondences. Observe that even for large levels of noise and outliers, the results are within reasonable bounds. We can assess the quality of these results by observing the graphs in the middle and bottom rows, that give significance of the errors we obtain. For instance, observe that a mean reconstruction error of 1.0 cm, still represents a good approximation to the true shape. In addition, note that considering the inextensibility constraints introduced in Section 4.1 yields improved results, and as expected, this improvement is more remarkable in the accuracy of the reconstructed shape. It is also worth to mention that using both approximations the number of iterations remains practically the same, and the algorithm typically converges in less than 50 iterations. One surprising exception is the noiseless case, which requires from a larger number

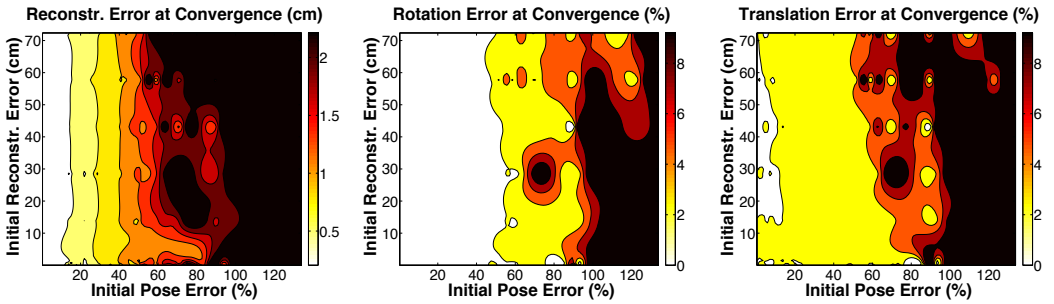

Fig. 8. Shape and pose errors at convergence, as a function of the error in the initialization.

of iterations. This is because our stopping criterion is based on relative, and not absolute, reductions of the error, and for the noiseless case, although the error is very small it decreases relatively slowly.

In terms of computation time, each iteration requires between 3 and 5 seconds, and thus the shape and poses for the 50 frames of the input sequences are computed in about 2 to 3 minutes. All the execution times correspond to a non-optimized implementation in Matlab.

In Figure 7 we evaluate the methodology described in Section 4.2 to detect outliers. Observe that we obtain very large rates of true positives and low rates of false positives. This means that our approach correctly detects most of the outlier correspondences, while only misclassifies a very small percentage of correct correspondences. Of course, the results slightly fall when the noise in the correspondences is increased, because then, correct but very noisy correspondences are classified as outliers. In addition, note that the outlier rejection results when using either matching or matching plus length constraints are virtually the same.

5.1.2 Convergence of the algorithm

In a second experiment with the synthetic data we evaluated the convergence behavior of our approach. With this purpose, we initialized our algorithm with very different poses and shapes, either relatively close to the true solutions or very far away. Figure 8 shows the reconstruction and pose errors at convergence as a function of the errors in the initialization. Errors above specific bounds are saturated and shown in black, such that we can consider the black regions, as non-convergence areas. In fact these non-convergence values are reasonable values for which the retrieved solutions are visually disturbing. Observe that convergence almost does not depend on the quality of the initial shapes, and the initial pose is the dominant factor. That being said, our algorithm tolerates errors in the initial pose of up to 80%, which is relatively large, specially

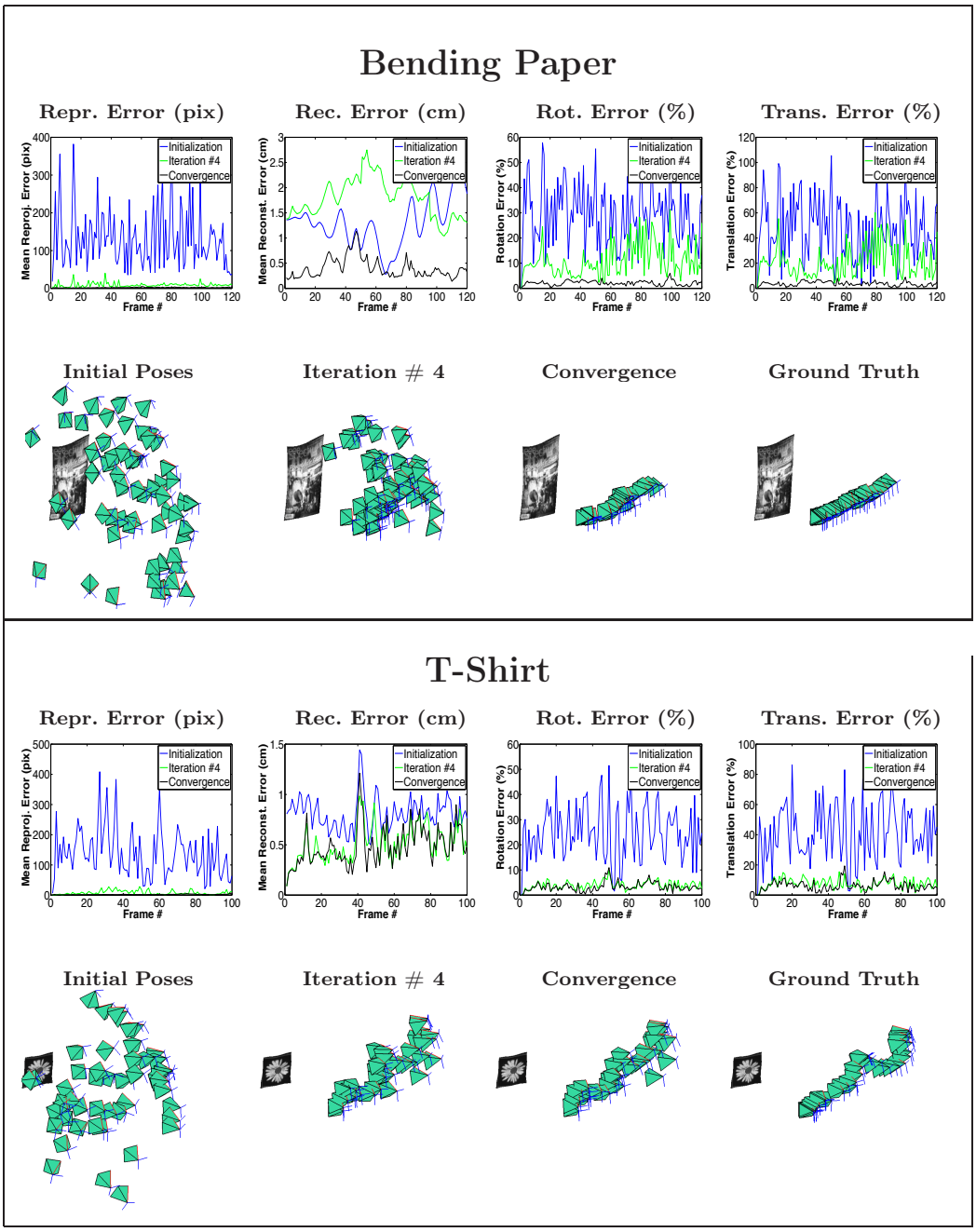

Fig. 9. Results on two real sequences. In each case, the upper plots depict the errors per frame obtained after initialization, 4 iterations and convergence. The bottom plots show the poses corresponding to the previous time instances.

considering that the pose error shown in the middle-right plot of Figure 6 is of about 50%.

We also tested our method on a 120-frames sequence of a bending paper and a 100-frames sequence of a deforming T-shirt, acquired with a Pointgrey Bumblebee stereo camera. In both cases the camera was moved around the deforming shape while capturing the sequence. The upper images in Figure 1 show three different frames of the "bending paper" sequence, where the movement of the camera can be clearly appreciated from the viewpoint change of the calibration box, which does not change its position.

We used the stereo rig to estimate the ground truth shape. For this, we computed the 3D position of a few points of interests on the surface and then inferred the position of the mesh vertices using linear interpolation. Since our algorithm just requires monocular sequences we ran it with only the images from one of the cameras. The ground truth camera pose was computed by applying the Perspective-n-Point (PnP) algorithm [26] over a small set of manually introduced correspondences between points on a 3D model of the calibration box and points in each of the input images. The 3D-to-2D correspondences between the reference configuration and the input images of the mesh were computed using SIFT [21]. Yet, as the reference and input images can be significantly different, these correspondences could have an arbitrarily large amount of outliers. The capability of our approach to handle outliers was therefore important to handle this situation.

Since the distance between the camera and the surface was roughly the same as for the synthetic experiments, and the inter-frame camera displacement was also very similar we used the same dynamic and measurement models defined in the previous section.

Yet, an issue we had to resolve was that the number of frames of the real sequences was much larger than for the synthetic case, and the size of the matrix $\mathbf{B}$ in Equation (18) became very large to be tractable. To handle this situation we implemented an incremental version of our algorithm, in which the sequence was split into several parts, and each part solved independently. However, in order to avoid jumps between the different parts, we allowed certain overlapping of the frames and shared their solution among sub-sequences. For the real experiments we considered 50 frames at a time from which 5 were repeated with the previous set. Note that the number of new frames considered at each iteration and the size of the overlap with previous processing steps allow trading off accuracy for speed, adjusting the efficiency to the requirements of the application at hand.

Figure 9 depicts the results for the two real experiments. In each case, the upper-row graphs plot the errors per frame, at initialization, after 4 iterations,

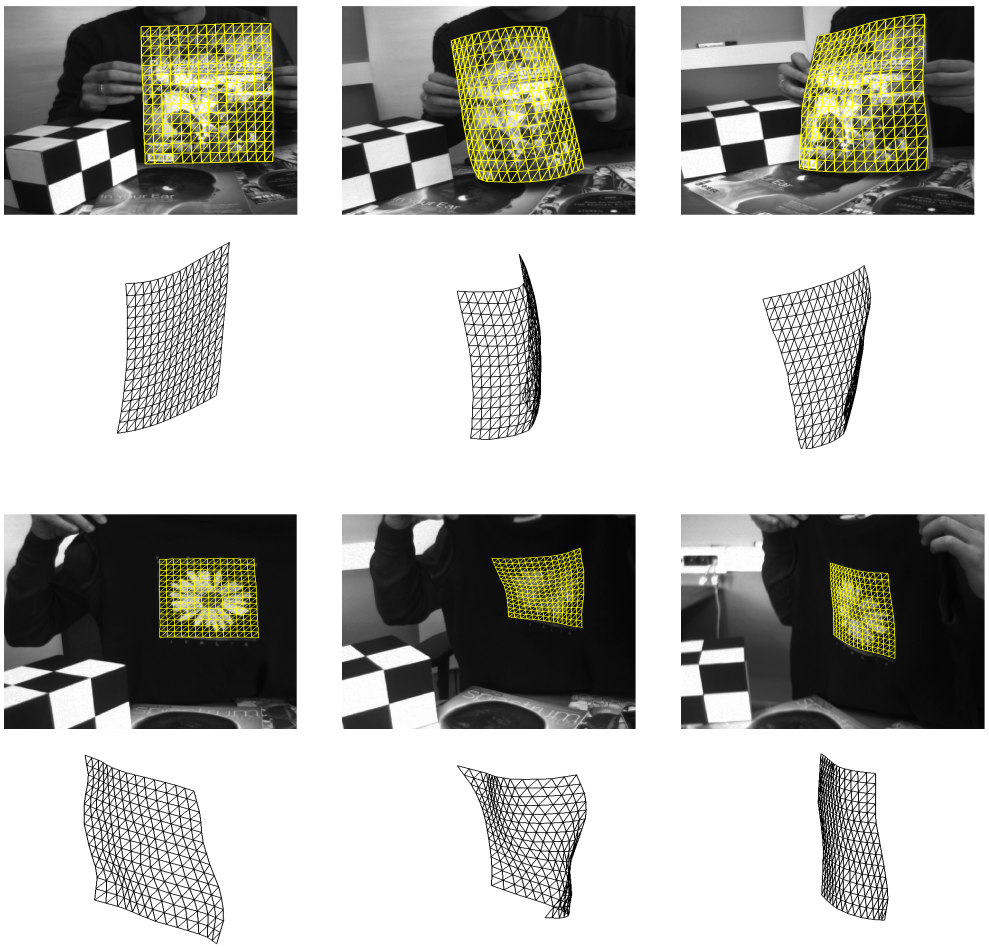

Fig. 10. Shape recovery on two real sequences. For each experiment, the upper figures correspond to the recovered mesh overlaid on the original image and the figures right below are the 3D mesh seen from a constant point of view, after eliminating the camera movement.

and at convergence. Since the results have been obtained by applying our algorithm to several sub-sequences the number of iterations to converge is not unique. However, all sub-sequences converged in about 50 to 70 iterations. The bottom plots show the configuration of retrieved camera poses. Observe that our algorithm yields fairly good results, specially considering the large error of the initial set of poses. Finally, in Figure 10 we show the detail of the recovered shape for different frames of each sequence.

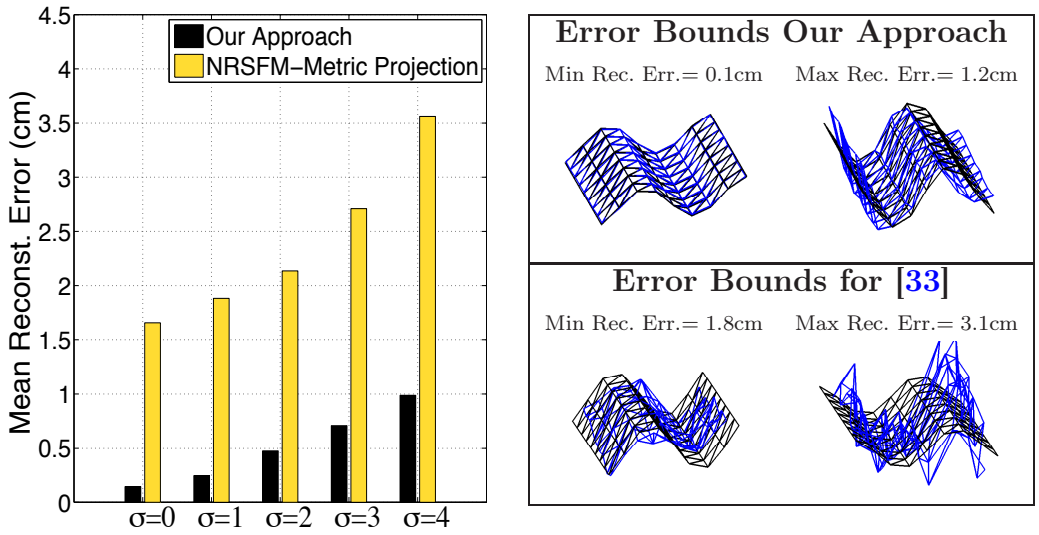

Fig. 11. Comparison with NRSFM aproaches. **Left:** Reconstruction error of our approach and that by [33] for the synthetic sequence in Section 5.1, as a function of the input noise. **Right:** Sample reconstructions showing the error bounds for both methods. Observe that even the solution with largest error of our approach represents a better approximation than the solution with smallest error obtained by [33]. This additional accuracy is a consequence of using known deformation modes.

5.3 Comparison with NRSFM techniques

We finally compare our method with the method proposed by [33], a recent Non-Rigid Shape From Motion algorithm. As said above, there are substantial differences between our approach and NRSFM methods. The most important is that we make use of a deformation model, computed from training data, while NRSFM methods do not assume that training data is available, and simultaneously estimate 3D shape and modes. This obviously makes these algorithms more general, but at the price of being more sensitive to noise and constrained to relatively small deformations.

Figure 11 shows the results of the comparison for the synthetic sequence used in Section 5.1. In order to satisfy the input data requirements of the method by [33], we provided the tracks of all the vertices of the mesh, and projected them onto the image using an orthographic camera model. We then computed the reconstruction error for increasing levels of input noise. As expected, the behavior of the NRSFM methods is quite poor, and becomes specially unstable for large amounts of noise. In contrast, the use of the deformation modes yields a remarkable robustness and stability to our algorithm. In addition, besides shape, our algorithm also provides an accurate estimation of the camera pose, which we did not show here because the method by [33] does not explicitly compute the pose.

6 Conclusions

In this paper we have shown that the problem of simultaneously retrieving pose and non-rigid 3D shape given a set of 3D-to-2D correspondences can be probabilistically formulated as a MAP estimate. We then introduce dynamic and measurement models accounting for noisy data, and reduce the MAP estimate to a non-linear least squares optimization that we solve using standard techniques. In the results section, we have shown that we obtain satisfactory results under situations where current methods are prone to fail, such as, large rates of outliers and noise in the input data, or very poor quality of the initializations.

The formulation of the problem we propose is very general, and allows introducing additional constraints either on the structure of the mesh or on the dynamic models. In particular, we have shown that length constraints on the edges of the mesh naturally fit within the proposed framework, and yield improved results compared to when uniquely using the constraints enforced by the 3D-to-2D correspondences.

As part of future work, we aim at developing online versions of the current approach following [18], which do not require from batch processing all the constraints. We believe that using more accurate dynamic models and the shape estimated in previous frames, will let our approach to converge in very few iterations. This would allow real time applications and even relax the dependence of the method on using deformation modes learnt from training data.

Acknowledgments

This work has been partially funded by Spanish Ministry of Economy and Competitiveness under projects RobInstruct TIN2014-58178-R and RobCab DPI2014-57220-C2-2-P; and by the ERA-net CHISTERA project I-DRESS PCIN-2015-147.

References

[1] A. Agudo, B. Calvo, and J.M.M. Montiel. Finite element based sequential bayesian non-rigid structure from motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1418–1425, 2012.

[2] A. Agudo, F. Moreno-Noguer, B. Calvo, and J.M.M. Montiel. FEM models to code non-rigid EKF monocular SLAM. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 38(5):979–994, 2016.

[3] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[4] K. S. Bhat, C. D. Twigg, J. K. Hodgins, P. K. Khosla, Z. Popovic, and S. M. Seitz. Estimating cloth simulation parameters from video. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pages 37–51, 2003.

[5] P. Biber and T. Duckett. Experimental analysis of sample-based maps for long-term SLAM. *International Journal of Robotics Research*, 28(1):20–33, 2009.

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

[7] M. Brand. Morphable 3D models from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 456–463, 2001.

[8] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 690–696, 2000.

[9] L. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2d and 3d images. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 15(11):1131–1147, 1993.

[10] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498, 1998.

[11] A. Dani, Z.Kan, N. Fischer, and W. Dixon. Structure estimation of a moving object using a moving camera: An unknown input observer approach. In *Conference on Decision and Control and European Control Conference*, pages 5005 –5010, 2011.

[12] T. Davis. Multifrontal multithreaded rank-revealing sparse QR factorization. *ACM Transactions on Mathematical Software*, 38(1), 2009.

[13] F. Dellaert and M. Kaess. Square root SAM: Simultaneous localization and mapping via square root information smoothing. *International Journal of Robotics Research*, 25(12):1181–1203, 2006.

[14] A. Ecker, A. D. Jepson, and K. N. Kutulakos. Semidefinite programming heuristics for surface reconstruction ambiguities. In *European Conference on Computer Vision*, pages 127–14, 2008.

[15] J. Fayad, L. Agapito, and A. Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In *European Conference on Computer Vision*, 2010.

[16] G. Grisetti, C. Stachniss, S. Grzonka, and W. Burgard. A tree parameterization for efficiently computing maximum likelihood maps using gradient descent. In *Robotis: Science and Systems*, 2007.

[17] P. Guan, A. Weiss, A. Balan, and M.J.Black. Estimating human shape and pose from a single image. In *International Conference on Computer Vision*, 2009.

[18] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental smoothing and mapping. *IEEE Transactions on Robotics*, 24(6):1365 –1378, 2008.

[19] A. Kawewong, N. Tongprasit, S. Tangruamsub, and O. Hasegawa. Online and incremental appearance-based SLAM in highly dynamic environments. *International Journal of Robotics Research*, 30(1):33–55, 2010.

[20] J. Leonard, I. Cox, and H. Durrant-Whyte. Dynamic map building for an autonomous mobile robot. *International Journal of Robotics Research*, 11(4):286–289, 1992.

[21] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[22] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60:135–164, 2004.

[23] T. McInerney and D. Terzopoulos. A finite element model for 3D shape reconstruction and nonrigid motion tracking. In *International Conference on Computer Vision*, pages 518–523, 1993.

[24] D. Metaxas and D. Terzopoulos. Constrained deformable superquadrics and nonrigid motion tracking. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 15(6):580–591, 1993.

[25] F. Moreno-Noguer and P. Fua. Stochastic exploration of ambiguities for non-rigid shape recovery. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 35(2):463–475, 2013.

[26] F. Moreno-Noguer, V. Lepetit, and P. Fua. Accurate non-iterative O(n) solution to the PnP problem. In *International Conference on Computer Vision*, 2007.

[27] F. Moreno-Noguer and J. M. Porta. Probabilistic simultaneous pose and non-rigid shape recovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1289–1296, 2011.

[28] F. Moreno-Noguer, J. M. Porta, and P. Fua. Exploring ambiguities for monocular non-rigid shape estimation. In *European Conference on Computer Vision*, pages 370–383, 2010.

[29] F. Moreno-Noguer, M. Salzmann, V. Lepetit, and P. Fua. Capturing 3D stretchable surfaces from single images in closed form. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1842–1849, 2009.

[30] E. Munoz, J. Buenaposada, and L. Baumela. A direct approach for efficiently tracking with 3D morphable models. In *International Conference on Computer Vision*, pages 1615–1622, 2009.

[31] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: dense tracking and mapping in real-time. In *International Conference on Computer Vision*, pages 2320–2327, 2011.

[32] E. Olson, J. Leonard, and S. Teller. Fast iterative optimization of pose graphs with poor initial estimates. In *International Conference on Robotics and Automation*, pages 2262–2269, 2006.

[33] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito. Factorization for non-rigid and articulated structure using metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2898–2905, 2009.

[34] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 13:715–729, 1991.

[35] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. In *British Machine Vision Conference*, 2008.

[36] V. Rabaud and S. Belongie. Linear embeddings in non-rigid structure from motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 124–137, 2009.

[37] M. Salzmann and P. Fua. Reconstructing sharply folding surfaces: A convex formulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1054–1061, 2009.

[38] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-form solution to non-rigid 3D surface registration. In *European Conference on Computer Vision*, volume 4, pages 581–594, 2008.

[39] J. Sanchez, J. Ostlund, P. Fua, and F. Moreno-Noguer. Simultaneous pose, correspondence and non-rigid shape. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1189–1196, 2010.

[40] R. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *International Journal of Robotics Research*, 5(4):56–68, 1986.

[41] D. Tardós, J. Neira, P. Newman, and J. Leonard. Robust mapping and localization in indoor environments using sonar data. *International Journal of Robotics Research*, 21(4):311–330, 2002.

[42] J. Taylor, A. Jepson, and K. Kutulakos. Non-rigid structure from locally-rigid motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2761–2768, 2010.

[43] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer. Elastically deformable models. In *Proc. ACM SIGGRAPH*, pages 205–214, 1987.

[44] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 30(5):878–892, 2008.

[45] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms*, pages 298–372, 1999.

[46] L. Tsap, D. Goldgof, and S. Sarkar. Nonrigid motion analysis based on dynamic refinement of finite element models. *IEEE Transactions Pattern Analylis and Machine Intelligence*, 22(5):526–543, 2000.

[47] R. Vidal and R. Hartley. Perspective nonrigid shape and motion recovery. In *European Conference on Computer Vision*, pages 276–289, 2008.

[48] M. R. Walter, R. M. Eustice, and J. Leonard. Exactly sparse extended information filters for feature-based SLAM. *International Journal of Robotics Research*, 26(4):335–359, 2007.

[49] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *International Conference on Computer Vision*, pages 1075–1082, 2005.

[50] W. Zhang, Q. Wang, and X. Tang. Real time feature based 3D deformable face tracking. In *European Conference on Computer Vision*, pages 720–732, 2008.