# Simultaneous Pose, Correspondence and Non-Rigid Shape

Jordi Sanchez-Riera, Jonas Östlund, Pascal Fua, Francesc Moreno-Noguer

# Simultaneous Pose, Correspondence and Non-Rigid Shape*


Jordi Sánchez-Riera
Perception Team
INRIA
Grenoble, France

Jonas Öslund
EPFL - CVLab
Lausanne
Switzerland

Pascal Fua
EPFL - CVLab
Lausanne
Switzerland

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



## Abstract

*Recent works have shown that 3D shape of non-rigid surfaces can be accurately retrieved from a single image given a set of 3D-to-2D correspondences between that image and another one for which the shape is known. However, existing approaches assume that such correspondences can be readily established, which is not necessarily true when large deformations produce significant appearance changes between the input and the reference images. Furthermore, it is either assumed that the pose of the camera is known, or the estimated solution is pose-ambiguous.*

*In this paper we relax all these assumptions and, given a set of 3D and 2D unmatched points, we present an approach to simultaneously solve their correspondences, compute the camera pose and retrieve the shape of the surface in the input image. This is achieved by introducing weak priors on the pose and shape that we model as Gaussian Mixtures. By combining them into a Kalman filter we can progressively reduce the number of 2D candidates that can be potentially matched to each 3D point, while pose and shape are refined. This lets us to perform a complete and efficient exploration of the solution space and retain the best solution.*


## 1. Introduction

Reconstructing 3D deformable surfaces from single images is one of the central goals in computer vision with a large number of applications in related fields such as robotics, computer graphics or augmented reality. When 3D-to-2D correspondences between an input image and another one for which the shape is known can be established, monocular 3D non-rigid reconstruction is a well understood problem effectively addressed by many recent works [7, 19, 21, 22, 23]. However all these methods rely on the quality of the matches, which are usually established by means of local image descriptors that may become unreliable when the shape deformation in the reference and input images is significantly different, or when the surface texture contains repetitive patterns, such as the example shown in Fig 1. In addition, existing approaches assume the pose of the camera with respect to the reference shape to be known and that does not change when capturing the input image.

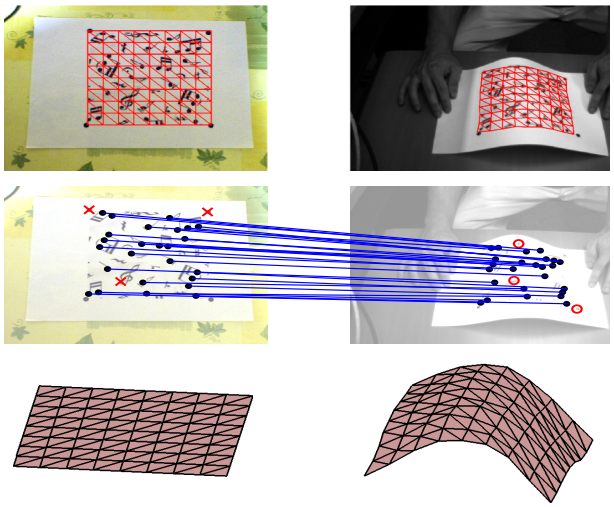

Figure 1. Simultaneous Correspondence and Non-Rigid Shape Reconstruction. **Left column:** Reference configuration for which we know the shape and a set of 3D points lying on it. **Right Column:** Given a set of 2D points in an input image in which the shape is unknown and the camera pose is different from that in the reference configuration, we seek to simultaneously establish the 3D-to-2D matches (second row) and retrieve the pose of the camera and the shape of the surface in the input image (third row, right image). Our approach solves the matching problem without considering texture information, which in this case is unreliable because the scene contains many repetitive patterns. In addition, we can handle certain amount of outliers and clutter points, shown as red crosses and circles, respectively, in the second row images.

In order to obtain a solution that performs robustly when texture is not a reliable cue and when the camera can freely move, we propose an approach that simultaneously solves for the correspondences, pose and 3D shape without using image appearance information, that is, just considering the


*This work has been partially funded by the Spanish Ministry of Science and Innovation under projects 200850I055, DPI2008-06022, and Consolider Ingenio 2010 CSD2007-00018, by EU GARNICS project FP7-247947, and by the Swiss National Science Foundation.

3D position and the 2D location of two unmatched sets of points. Inspired on a recent paper on rigid object pose estimation [18], we reduce the complexity of this problem using weak priors on pose and shape, that we learn from training data, and model as Gaussian Mixture Models. These priors let us to define a region in the image where to seek for the potential 2D candidates that may be assigned to each 3D point. Using a Kalman filter strategy this search region is progressively shrunk while the estimations of the pose and shape are refined. This is repeated for different combinations of pose and shape priors, in order to guarantee a complete exploration of the solution space. As shown in Fig. 1, we can recover shape and correspondences even under the presence of outliers and clutter.

To the best of our knowledge, this is the first approach addressing the *simultaneous pose and correspondence problem for non-rigid objects*. Indeed, several works have already been proposed for retrieving matches and the six degrees of freedom of the pose in rigid objects [6, 10, 18, 20]. However, when reconstructing non-rigid surfaces many more variables need to be considered to account for the deformation degrees of freedom, which require from a different solution to make the problem tractable.

## 2. Related Work

Monocular 3D reconstruction of non-rigid surfaces is known to be a highly under-constrained problem that requires from prior information to solve it.

The most common approach to limit the set of possible solutions is to represent the shape as a weighted sum of modes, either physically-based ones [4, 15, 16] or learned from training data [2, 5]. Estimating shape, then amounts to retrieving the weights of this linear combination, by minimizing an image-based objective function. However, since such functions usually are non-convex and have many local minima, these methods require from good initializations.

Several approaches have shown that shape may be recovered from a set of 3D-to-2D correspondences, between the 3D points of a reference shape and the 2D points on the input image [7, 19, 21, 22, 23]. These point correspondences use to be computed by SIFT-like local image descriptors [1, 12, 14, 17], which have been proved to be robust to certain transformations of the image, although they are prone to fail when dealing with the large nonlinearities produced in the deformation of non-rigid surfaces. Recent works have addressed the problem of general deformations [3, 13]. While these are promising approaches, they are still based on assumptions that can only rarely be satisfied, such as that intensity variations due to deformations are minimal.

This brings us to the situation where image intensity becomes an unreliable cue to establish correspondences, for example because the deformation itself creates artifacts on the image, such as self-shadowing or strong occlusions, or because the surface texture contains repetitive and undistinguishable patterns. In these cases it becomes necessary to simultaneously retrieve the shape and establish the correspondences. This problem has been deeply studied for rigid objects, in which case is then necessary to simultaneously estimate pose and correspondences. The most traditional approach in doing so is the RANSAC algorithm [9], which, given a set of 3D and 2D points, iteratively hypothesizes and validates small subsets of correspondences until a good solution is found. Many variants of this strategy have been proposed over the years to reduce the computational cost [6, 10, 18, 20].

In fact, our approach is inspired in [18], a recent work that uses priors on the camera pose to register rigid objects. However, in order to address the problem for non-rigid surfaces, besides pose and correspondences, we will have to estimate additional variables accounting for the degrees of freedom of the surface deformation. This, considerably increases the complexity of the problem and we believe that our approach is the first one to simultaneously solve for correspondences, pose and non-rigid shape from single images.

## 3. Shape and Pose Priors for Non-Rigid Surface Reconstruction

In this section we first show that the solution of our problem can be expressed as a minimization of an error function depending on the correspondences, pose, and shape parameters. We then show that weak priors on both pose and shape can be effectively used to retrieve the minimum of the error function, in spite of having to explore a high dimensional solution space.

### 3.1. Problem Statement and Notation

We assume we are given a set of 3D points $\mathcal{P} = \{\mathbf{p}_1^{ref}, \ldots, \mathbf{p}_M^{ref}\}$ on a *reference configuration* with known shape, and a set of 2D points $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ on an *input image* of the same surface but with a different and unknown deformation. The correspondence between these two sets of points is also unknown.

We represent the surface as a triangulated 3D mesh with $n_v$ vertices $\mathbf{v}_i$ concatenated in a vector $\mathbf{x} = [\mathbf{v}_1^\top, \ldots, \mathbf{v}_{n_v}^\top]^\top$, and denote by $\mathbf{x}^{ref}$ the reference mesh, and $\mathbf{x}$ the mesh we seek to recover.

Let $\mathbf{p}_i$ be a point on the mesh $\mathbf{x}$ corresponding to the point $\mathbf{p}_i^{ref}$ in the reference configuration. We can express $\mathbf{p}_i$ in terms of the barycentric coordinates of the face it belongs:

$$\mathbf{p}_i = \sum_{j=1}^{3} a_{ij} \mathbf{v}_j^{[i]} \,, \tag{1}$$

where the $a_{ij}$ are the barycentric coordinates and $\mathbf{v}_j^{[i]}$ are the vertices of the face containing the point $\mathbf{p}_i$. Since we

assume the mesh does not stretch while it deforms, these barycentric coordinates remain constant for each point and can be easily computed from points $\mathbf{p}_i^{ref}$ and the mesh $\mathbf{x}^{ref}$. Let us denote by $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$ the set of barycentric coordinates associated to the 3D points, where $\mathbf{a}_i = [a_{i1}, a_{i2}, a_{i3}]^\top$.

Additionally, we assume we can model the mesh deformation as a linear combination of a mean shape $\mathbf{x}_0$ and $n_m$ deformation modes $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{n_m}]$

$$\mathbf{x} = \mathbf{x}_0 + \sum_{i=1}^{n_m} \alpha_i \mathbf{q}_i = \mathbf{x}_0 + \mathbf{Q}\boldsymbol{\alpha} \ , \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{n_m}]^\top$ are the unknown weights of the basis shapes. In our implementation, these modes were obtained by applying Principal Component Analysis (PCA) over a set of synthetic inextensible meshes generated with the modelling software Blender [11].

And finally, we assume we know the calibration matrix $\mathbf{A}$ of the camera, although we do not know its pose. As shown in Fig. 1, this may occur if the reference and input images were captured with different camera positions.

Given all these initial assumptions, our goal is to simultaneously retrieve the pose of the camera –parameterized by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$– , the shape of the mesh $\mathbf{x}$ –parameterized by the vector $\boldsymbol{\alpha}$–, and as many 3D-to-2D correspondences as possible, considering that not all the 3D points must have a 2D match and viceversa.

Let us consider $\tilde{\mathbf{u}}_i$ to be the projection of a point $\mathbf{p}_i$ given a pose and shape estimates $\{\mathbf{R}, \mathbf{t}, \boldsymbol{\alpha}\}$:

$$\begin{aligned} w_i \begin{bmatrix} \tilde{\mathbf{u}}_i \\ 1 \end{bmatrix} &= \mathbf{A}\,[\mathbf{R}|\mathbf{t}] \begin{bmatrix} \mathbf{p}_i \\ 1 \end{bmatrix} \\ &= \mathbf{A}\mathbf{R} \sum_{j=1}^{3} a_{ij} \left( \mathbf{x}_{0j}^{[i]} + \mathbf{Q}_j^{[i]} \boldsymbol{\alpha} \right) + \mathbf{A}\mathbf{t} \end{aligned} \tag{3}$$

where $w_i$ is a scalar projective parameter and $\mathbf{x}_{0j}^{[i]}$ and $\mathbf{Q}_j^{[i]}$ are the subvector of $\mathbf{x}_0$ and submatrix of $\mathbf{Q}$ corresponding to the coordinates of the vertex $\mathbf{v}_j^{[i]}$. Note that in the second step we have used Eq. 1 and 2 to write the point $\mathbf{p}_i$ in terms of the modal weights.

We can now formulate our problem as an optimization one, where we seek to retrieve the parameters $\mathbf{R}$, $\mathbf{t}$ and $\boldsymbol{\alpha}$ such that minimize the reprojection error between points $\mathbf{p}_i$ projected on the image and their corresponding matches. This can be written as

$$\underset{\mathbf{R},\mathbf{t},\boldsymbol{\alpha}}{\text{minimize}} \sum_{i=1}^{M} \mathsf{Detected}\left( \|\tilde{\mathbf{u}}_i - \mathsf{Match}(\tilde{\mathbf{u}}_i, \mathcal{U})\| \right) \tag{4}$$

where $\mathsf{Match}(\tilde{\mathbf{u}}_i, \mathcal{U})$ returns the 2D point of $\mathcal{U}$ that is closest to $\tilde{\mathbf{u}}_i$, and $\mathsf{Detected}(d) = \min(d, \mathcal{T})$ is used to penalize points that are not correctly matched and avoid trivial solutions with just a few, but accurate, matches. In all our experiments we set $\mathcal{T} = 10$ pixels.

### 3.2. Estimating Pose and Shape Priors

Trying to minimize Eq. 4 with no other information besides the deformation modes, becomes computationally prohibitive, even when using a small number of points. For example, the approach in [6] is one of the most efficient algorithms in solving pose and correspondences in the rigid case and has a $\mathcal{O}(MN^2)$ complexity, which allows to handle problems with about a $M \approx 100$ points in a reasonable amount of time. However, if on top of this complexity one has to consider the additional $n_m$ degrees of freedom introduced by the deformable model, the problem seems to be unsolvable in practice.

In this paper, we show that using weak priors on the camera pose and on the types of deformations the surface can have, we can turn our problem into a tractable one. To this end, we first compute pose and shape priors and parameterize them in terms of Gaussian Mixture Models. We then combine these priors, and using a Kalman filter strategy we guide the matching process for each 3D point. We show that the search space for potential 2D candidates is drastically reduced from the entire image, to a small region. We can then efficiently explore the space of possible solutions and keep the solution that minimizes Eq. 4.

#### 3.2.1 Priors Computation

Pose priors are computed following a similar methodology as in [18]. Given the reference mesh, we initially generate sample positions on a large region where the camera center is expected to be. For example, for the $40 \times 40$ cm mesh shown in Fig. 2(Left), the region of possible camera positions we define is a $140 \times 140 \times 20$ cm$^3$ volume above the mesh. Then, for each camera center sample, we consider a random direction of the optical axis by allowing the camera to point anywhere on the reference mesh. We also allow a random rotation of the camera around its optical axis. All these pose samples are then represented by 6-dimensional vectors accounting for the three degrees of freedom of the rotation and the three of translation. Using Expectation-Maximization (EM) [8] we model them as a Gaussian Mixture Model, with $\{\boldsymbol{\rho}_1^p, \ldots, \boldsymbol{\rho}_{G_p}^p\}$ mean poses with associated $6 \times 6$ covariances $\{\boldsymbol{\Sigma}_1^p, \ldots, \boldsymbol{\Sigma}_{G_p}^p\}$, and probabilities $\{\mathrm{p}_1^p, \ldots, \mathrm{p}_{G_p}^p\}$. Fig. 2(Left) shows the part of the pose priors corresponding to the camera centers, with their associated $3 \times 3$ covariance matrices.

As mentioned above we represent the surface as a linear combination of $n_m$ basis shapes computed by applying PCA on a large training set of synthetically deformed meshes. Therefore, each shape is indeed represented by a $n_m$-vector of weights $\boldsymbol{\alpha}$. Shape priors are then computed by applying EM over the weights of all training meshes. This results in $G_s$ mean shape vectors $\{\boldsymbol{\rho}_1^s, \ldots, \boldsymbol{\rho}_{G_s}^s\}$ with associated $n_m \times n_m$ covariances $\{\boldsymbol{\Sigma}_1^s, \ldots, \boldsymbol{\Sigma}_{G_s}^s\}$ and probabilities $\{\mathrm{p}_1^s, \ldots, \mathrm{p}_{G_s}^s\}$. Fig. 2(Middle) shows the region

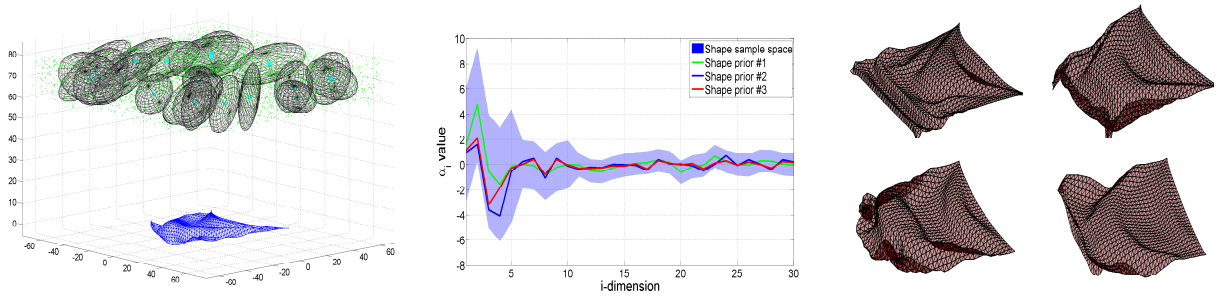


Figure 2. **Left**: Pose Priors. The small green dots are the samples of camera centers we use to compute the pose priors. The ellipsoids represent the covariance of the priors on the translational space. We obtain similar covariances for the three-dimensional rotational space, although we do not plot them here. **Middle-Right**: Shape Priors. We represent the shape by a 30-dimensional vector of modal weights, and compute the shape priors in this space. The blue shaded area represents the region where shape samples are generated and the meshes on the right side are four of such samples. Expectation-Maximization is then applied over these samples to compute the shape priors. The color lines in the middle image represent the mean vector for three of these priors.

of the shape space where samples were generated and the mean vectors of three shape priors.

It is worth to mention that the kind of priors on the modal weights we use are more restrictive than the regularization terms introduced by current techniques to penalize the less meaningful modes. This lets us to retrieve the shape without imposing the additional constraints that these methods consider, based on local inextensibility [7, 21, 22, 23] or shading information [19].

#### 3.2.2 Weighting Joint Priors

In the next subsection we will use different combinations of pose and shape priors to simultaneously solve for the correspondences while pose and shape are refined. We could, in principle, explore all possible combinations of $\{(\boldsymbol{\rho}_i^p, \boldsymbol{\Sigma}_i^p), (\boldsymbol{\rho}_j^s, \boldsymbol{\Sigma}_j^s)\}$ for $i = 1, \ldots, G_p$ and $j = 1, \ldots, G_s$. However, we prefer not doing so because this has a high computational cost, and to avoid having to explore all the $G_p \cdot G_s$ prior combinations, the algorithm will be stopped when Eq. 4 drops below a certain threshold. In order to further accelerate this convergence we will not explore the joint priors following a random ordering. Instead, we will explore them according to the ordering determined by a weight computed as follows.

Given the 2D image points $\mathcal{U}$, a combination of the $i$-th pose and $j$-th shape priors will be assigned a weight proportional to the joint probability

$$\mathrm{p}(i, j|\mathcal{U}) \propto \mathrm{p}(\mathcal{U}|i, j)\mathrm{p}_i^p\mathrm{p}_j^s \tag{5}$$

where $\mathrm{p}(\mathcal{U}|i, j)$ is computed by projecting the points $\{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$ onto the image assuming a pose $\boldsymbol{\rho}_i^p$ and a shape $\boldsymbol{\rho}_j^s$, and comparing these projected points with the actual distribution $\mathcal{U}$. The comparison is done in terms of point cloud Hausdorff distance and, since 3D-to-2D correspondences are not explicitly solved, it can be computed very fast. Although this is just an approximate measurement, in the Results Section we will show that this ordering lets us to terminate the algorithm when only a small percentage of the joint priors are explored.

### 3.3. Using Priors to Simultaneously Solve for Pose, Shape and Correspondences

We will now use the joint priors to guide the 3D-to-2D matching process while pose and shape are progressively refined. Starting with the pair of priors $\{(\boldsymbol{\rho}^p, \boldsymbol{\Sigma}^p), (\boldsymbol{\rho}^s, \boldsymbol{\Sigma}^s)\}$ that was scored the most probable according to Eq. 5, our method is based on the following iterative algorithm.

#### 3.3.1 Computing initial set of potential matches

We denote by $\boldsymbol{\rho}^{p,s} = [\boldsymbol{\rho}^{p\top}, \boldsymbol{\rho}^{s\top}]^\top$ the initial value of our state vector accounting for pose and shape. Given this initial estimate we use Eq. 3 to project the set of 3D points $\mathcal{P}$, represented by their barycentric coordinates $\mathcal{A}$, onto the image plane at positions $\tilde{\mathcal{U}} = \{\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_M\}$.

In order to define a search region for each point $\tilde{\mathbf{u}}_i$ in the image plane, we propagate the covariance matrices according to

$$\boldsymbol{\Sigma}_i^{\mathbf{u}} = \mathbf{J}(\mathbf{a}_i)\boldsymbol{\Sigma}^{p,s}\mathbf{J}(\mathbf{a}_i)^\top \tag{6}$$

where $\mathbf{J}(\mathbf{a}_i)$ is the $2 \times (6 + n_m)$ Jacobian of Eq. 3 with respect to the 2D coordinates evaluated on the barycentric coordinates of the point $\mathbf{p}_i$, and $\boldsymbol{\Sigma}^{p,s}$ is the $(6 + n_m) \times (6 + n_m)$ block diagonal covariance matrix, in which the diagonal elements are $\boldsymbol{\Sigma}^p$ and $\boldsymbol{\Sigma}^s$.

We can then define an elliptical search region on the image plane by considering the potential matches for $\tilde{\mathbf{u}}_i$, as the $\mathbf{u}_j \in \mathcal{U}$ such that

$$(\mathbf{u}_j - \tilde{\mathbf{u}}_i)^\top(\boldsymbol{\Sigma}_i^{\mathbf{u}})^{-1}(\mathbf{u}_j - \tilde{\mathbf{u}}_i) \leq \mathcal{M}^2 \tag{7}$$

where $\mathcal{M}$ is a threshold chosen to guarantee a specified level of confidence, and its value can be computed using the cumulative chi-squared distribution. In our experiments we set $\mathcal{M} = 3$ to achieve a $99\%$ degree of confidence.

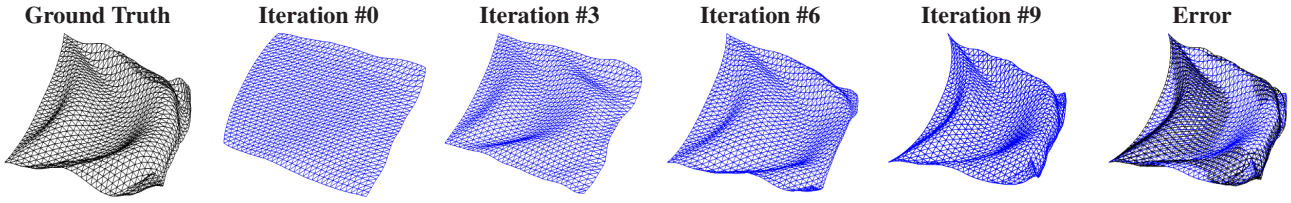


Figure 3. Shape Convergence. When correct matches are iteratively assigned the shape estimated by the Kalman Filter converges to the ground truth solution. Observe that at iteration #9 only slight differences remain between the estimated shape and the ground truth one.

#### 3.3.2 Iteratively refining pose and shape

Given the set of 3D points and their potential 2D candidates, we could now follow a RANSAC-based approach and hypothesize sets of 3D-to-2D correspondences and validate them with any of the current techniques to retrieve shape from such correspondences [7, 19, 21, 22, 23]. However, these methods are based on a relatively large number of correspondences –about $n_m$–, which would require from an excessively large number of hypotheses to guarantee retrieving a correct solution.

More specifically, let us assume that all the 3D points have the same number $n << N$ of potential 2D matches. If $p_d$ is the probability that a 3D point is detected and is not an outlier, the probability of $n_m$ correct matches is $(p_d/n)^{n_m}$. Hence, the number of hypotheses-and-validations that are needed to ensure with a probability $R$ that at least one of them is correct will be:

$$\underset{\text{Constant Search Region}}{\text{Number Trials}} = \left(\frac{n}{p_d}\right)^{n_m} \log\left(\frac{1}{1-R}\right) \qquad (8)$$

In our experiments we use $n_m = 30$ deformation modes, and hence, even for relatively small search regions yielding only a few 2D potential candidates $n$ for each 3D point, the number of trials would become prohibitive.

In order to reduce this theoretical high number of trials, we will use a Kalman filter formulation, where correspondences are progressively done for different 3D points, and after each match the number of potential 2D candidates for the rest of 3D points is reduced.

We start considering the 3D point with less potential candidates. Let $\tilde{\mathbf{u}}_i \in \tilde{\mathcal{U}}$ be the image projection of such a point, and $\mathbf{u}_j \in \mathcal{U}$ the 2D point closest to $\tilde{\mathbf{u}}_i$ in terms of the Mahalanobis distance of Eq. 7. By establishing an hypothetical match between $\tilde{\mathbf{u}}_i$ and $\mathbf{u}_j$, we can use the Kalman filter equations to update the state vector $\boldsymbol{\rho}^{p,s}$ and reduce the covariance matrix $\boldsymbol{\Sigma}^{p,s}$ in the pose-shape space:

$$\boldsymbol{\rho}^{p,s+} = \boldsymbol{\rho}^{p,s} + \mathbf{K}(\mathbf{u}_j - \tilde{\mathbf{u}}_i) \qquad (9)$$

$$\boldsymbol{\Sigma}^{p,s+} = (\mathbf{I} - \mathbf{K}\mathbf{J}(\mathbf{a}_i))\,\boldsymbol{\Sigma}^{p,s} \qquad (10)$$

where $\mathbf{K}$ is the Kalman gain and $\mathbf{I}$ is the $(6+n_m)\times(6+n_m)$ identity matrix.

The new pose, shape and covariance matrices are used to project again the 3D points onto the image, with the renewed search regions which are considerably smaller than those in the previous iteration. This allows to reduce the number $n$ of potential 2D matches that can be associated to each 3D point. A second match is then established and the Kalman update equations applied again. After a few iterations –less than 10 in practice– of repeating this process the Kalman filter converges to a solution, although we do not know yet if it corresponds to the correct one.

To assess the accuracy of the solution we match the remaining 3D points. For that, we project them onto the image plane and match to their nearest neighbor from $\mathcal{U}$. We then compute the error of Eq. 4, and if it is below a given threshold the algorithm is stopped. Otherwise we repeat the whole process of establishing correspondences and updating shape and pose with different combinations of the potential candidates, and by iterating over the various pose and shape joint priors according to the ordering determined in Section 3.2.2. Fig. 3 shows how the shape estimation converges to the true solution when correct matches are established.

Let us now analyze the computational complexity of the solution we propose. If we denote by $n_{it} << n_m$ the number of iterations until convergence of the Kalman filter, and by $\mathsf{Shrink}(\ )$ a monotonically decreasing function accounting for the reduction of the search region size, we can rewrite Eq. 8 by

$$\underset{\text{Shrunk Search Region}}{\text{Number Trials}} = \left(\frac{\prod_{i=1}^{n_{it}} \mathsf{Shrink}(i)\cdot n}{p_d^{n_m}}\right) \log\left(\frac{1}{1-R}\right) \qquad (11)$$

Since the number of iterations $n_{it} \approx 10$ is considerably smaller than the number of modes $n_m = 30$ and the product $\mathsf{Shrink}(i)\cdot n$ rapidly drops to 1, the final amount of theoretical trials that our approach requires is much smaller than when considering search regions of constant size.

However, still one critical element remains in Eq. 11, which is the increase in complexity produced by the number of outliers, represented by the percentage $p_d$ of detected 3D points. For example, if we had $20\%$ of outliers, that is $p_d = 0.8$, the term $(1/pd)^{n_m}$ would represent having to evaluate $870$ times more hypotheses than when all the points are detected. And for $p_d = 0.6$, this factor would grow up to $4.5 \times 10^6$.

In order to handle this situation, we consider that each point $\tilde{\mathbf{u}}_i$ can be either matched to any of its potential

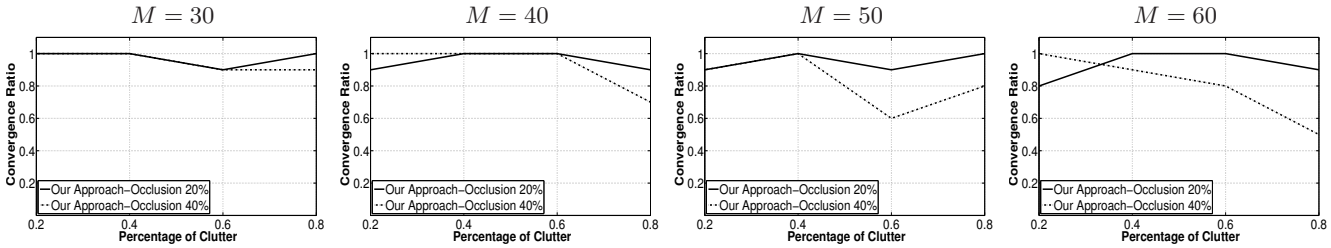


Figure 4. Convergence ratios in the synthetic experiments.

matches within the search region of Eq. 7, or it can be an outlier. Since correspondences are iteratively established while pose and shape are refined, the probability that $O$ consecutive 3D points are considered as outliers, decreases with $O$. Thus, we limit the maximum number $n_{co}$ of consecutive 3D points labelled as "outliers". In all our experiments we fixed this maximum value to $n_{co} = 0.25Mp_d$. For $M = 50$ 3D points, and $p_d = 0.6$ this would mean that we only should have to evaluate $(1/pd)^{n_{co}} = 46$ times more hypotheses than for the case $p_d = 1$.

## 4. Results

We now present the results on both synthetic and real data. In the synthetic results we compare our approach to [18], which we denote by *BPnP*. As discussed in Section 2 this work uses priors on the camera pose to simultaneously retrieve pose and correspondences for rigid objects.

### 4.1. Synthetic Experiments

Using the Blender software [11], we synthesized 25 frames of a deforming $40 \times 40$ cm flag approximated by a $32 \times 32$ mesh such as the one shown in Fig. 3(Left). In order to generate the input data we randomly distributed $\mathcal{M}$ 3D points on the surface of a reference planar mesh. Assuming a virtual camera placed approximately 70 cm upon the reference mesh, a percentage $p_d$ of the 3D points – accounting for the number of detected points– were projected onto the input image of a deformed mesh. A 2 pixel variance Gaussian noise was added to these projections. Furthermore, to account for clutter, a percentage $p_c$ of 2D points were added at random positions in the image plane.

Fig. 2 shows the kind of priors we used. For the pose, we generated $G_p$=20 Gaussian priors distributed in a relatively large volume over the mesh. The shape priors were approximated by $G_s = 5$ Gaussian distributions in the $n_m = 30$-dimensional space of the modal weights.

In order to evaluate our approach we proceeded similarly as in [18]. For each one of the 25 meshes, we performed several experiments by changing the number of 3D model points $M = \{30, 40, 50, 60\}$, the fraction $p_d = \{0.8, 0.6\}$ of detected points –corresponding to 20% and 40% occlusion rates, respectively–, and the percentage of clutter $p_c = \{0.2, 0.4, 0.6, 0.8\}$. In addition, we did 10 different trials for each combination of these parameters, yielding a total of $25 \times 4 \times 2 \times 4 \times 10$=8000 experiments.

We report the accuracy of our method in terms of a convergence ratio we define by comparing the pose and shape we retrieve with their ground truth values. More specifically, let $\boldsymbol{\rho}^p_{true}$, $\boldsymbol{\rho}^s_{true}$ be the true pose and shape, and let $\boldsymbol{\rho}^p$, $\boldsymbol{\rho}^s$ be our respective estimates. We computed the relative errors by $E^p = \|\boldsymbol{\rho}^p_{true} - \boldsymbol{\rho}^p\|/\|\boldsymbol{\rho}^p_{true}\|$ and $E^s = \|\boldsymbol{\rho}^s_{true} - \boldsymbol{\rho}^s\|/\|\boldsymbol{\rho}^s_{true}\|$, and considered a solution did converge if both $E^p$ and $E^s$ were below 0.1. The convergence of the BPnP was computed by just considering the percentage of error in the pose.

The results of all experiments are summarized in Fig. 4. Each graph plots the mean convergence as a function of $M$, $p_c$ and $p_d$. Observe that for most experimental conditions our approach guarantees a convergence ratio above 0.8. Only slightly lower values are obtained for large percentages of clutter and occlusions.

In Fig. 5 we compare the performance of our approach to that of the BPnP, in two meshes with different levels of deformation. As shown in the left column of Fig. 5, when the deformation of the input mesh is relatively similar to the planar reference configuration, BPnP still yields valid solutions with large percentages of convergence, almost comparable to those obtained with our approach. However, as shown in the right column of Fig. 5, when dealing with meshes whose deformation significantly differs from that of the reference configuration, the accuracy of the BPnP rapidly deteriorates, while we still obtain good results.

In terms of computation time our approach needs approximately 10, 15, 20 and 25 minutes per frame to compute a solution for $M = 30, 40, 50$ and 60 model points, respectively, when all the joint priors on pose and shape need to be explored. This is about 10 times more than the time required by BPnP, although this was indeed expected because we are seeking the solution in a space of much higher dimensionality and we are considering a much larger number of priors. Nevertheless, when exploring the priors combination following the ordering defined in Section 3.2.2, in many of the experiments we obtained a good solution without having to consider all of the joint priors. This let us terminate the algorithm much faster. Fig. 6 plots a histogram of the number of joint priors explored in 200 experiments with $G_p = 20$

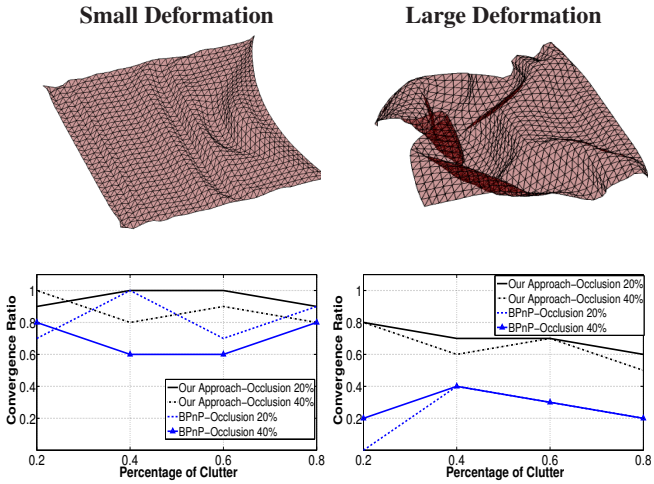


Figure 5. Comparing the convergence rate of our approach and BPnP [18]. **Left Column:** Results over a mesh with small deformation compared to a planar reference shape. **Right Column:** Results over a mesh with large deformation.

and $G_s = 5$. Note that although this yields a total of 100 possible joint priors, for most of the experiments the algorithm converged to a solution after exploring less than 20 of them.

### 4.2. Real Experiments

We also evaluated our approach on a real 150 frames-sequence of the 16×16 cm bending paper already introduced in Fig. 1, which is textured with repetitive patterns of musical notes. This is a clear example where texture is not a reliable cue for establishing matches, and we can take advantage of a technique like ours that uses geometry alone.

We initially acquired a reference image under a planar configuration, as shown in the top-left image of Fig. 1, and using SIFT [14] we detected a set of interest points, and computed their 3D barycentric coordinates with respect to the mesh vertices. For each input image the set of 2D points was computed using the same detector. Since SIFT usually returns several hundreds of interest points per image but our algorithm only can handle in a reasonable amount of time about $M = 60$ model points, we just kept those with larger gradient values. Note however, that this is the typical amount of points that state-of-the-art algorithms solving the simultaneous pose and correspondence problem for rigid objects can handle.

Pose priors were defined in a region of about $1\text{m}^3$ above the mesh, large enough to ensure it contained the real camera pose. Shape priors, were computed applying the methodology explained in Section 3.2.1 over synthetic sequences of meshes resembling paper deformations.

Figure 7 shows the reconstruction obtained in several frames of the sequence. Observe that with the proposed approach we are able to retrieve shape in situations where non-rigid deformations and repetitive patterns, might lead to failure the algorithms that just rely on texture for establishing correspondences.

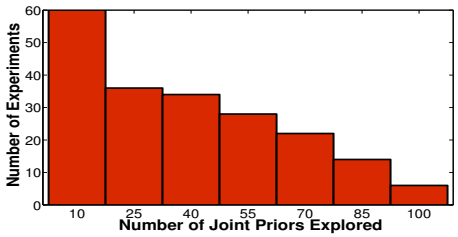


Figure 6. Number of Joint Priors Explored. If the combination of pose and shape priors are evaluated following the ordering defined in Section 3.2.2, our algorithm converges to a solution without having to explore all of the joint priors. The figure plots the distribution of number of explored priors over a maximum of 100, for 200 different experiments.

Finally, we also applied our approach to simultaneously estimate the camera pose and shape of a sail. Fig. 8-Left shows the priors we used on the translational component of the pose. Fig. 8-Right shows the results obtained in two different frames, where the camera pose we retrieve is represented on a coordinate system fixed on the sail.

## 5. Conclusion

Many recent approaches to monocular 3D shape reconstruction rely on the fact that point correspondences can be readily established between the input image and a reference image in which the shape and pose are known. However, there are cases where it is difficult to compute such matches, for example, when the surface deformation in the input image highly differs from that in the reference one, or when the surface texture contains repetitive patterns. Furthermore, the camera may move, and hence, its pose can no longer be considered to be known.

In such cases it is necessary to simultaneously estimate the camera pose, shape and correspondences, which is a computationally prohibitive problem unless additional constraints are considered. In this paper we have presented an approach to turn this problem into a tractable one, by using weak priors on both pose and shape, and modelling them using Gaussian Mixture Models. We have shown that using a Kalman filter strategy such priors can be progressively refined while solving for the correspondences to an ever smaller number of possible matches.

In future work, we plan to integrate additional sources of information, such as texture and motion. For instance, although we have shown that texture may be in some occasions an unreliable cue, we could nevertheless use its information as additional prior with an associated uncertainty. We believe that this would drastically reduce the number of potential 2D matches for each 3D point, yielding faster and more accurate solutions.

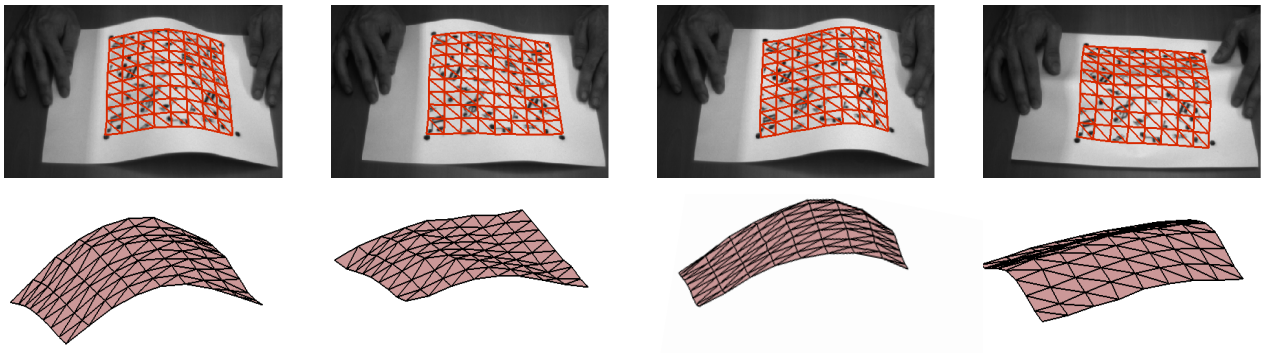

Figure 7. Reconstruction of a bending paper. Top: Mesh recovered using the proposed approach overlaid on the original image. Botton: Reconstructed mesh seen from a different viewpoint.

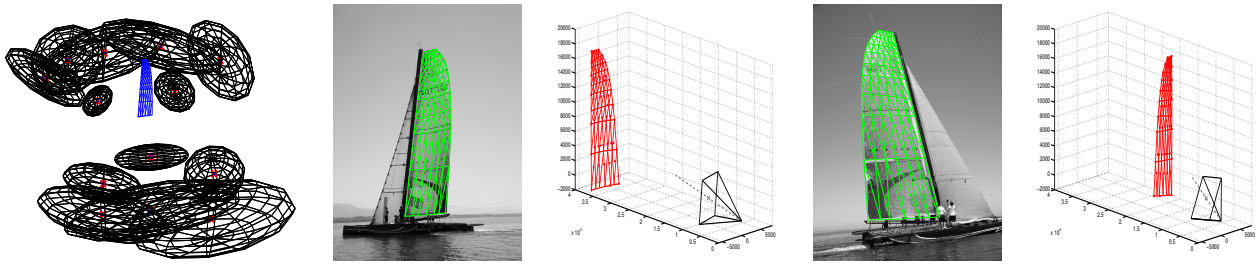

Figure 8. Reconstructing a sail and retrieving the camera pose. Left: Pose priors placed all around the mean shape. Right: Detection of the sail and the camera pose. For each pair of images we plot the recovered mesh overlaid on the original image and a 3D plot of the retrieved shape and the estimated camera pose.

## References


[1] S. Belongie, M. Jitendra. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

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

[3] H. Cheng, Z. Liu, N. Zheng, and J. Yang. A deformable local image descriptor. In *CVPR*, 2008.

[4] L. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2D and 3D images. *PAMI*, 15(11):1131–1147, 1993.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998.

[6] P. David, D. DeMenthon, R. Duraiswami, and H. Samet. Softposit: Simultaneous pose and correspondence determination. *IJCV*, 59(3):259–284, 2004.

[7] A. Ecker, A. D. Jepson, and K. N. Kutulakos. Semidefinite programming heuristics for surface reconstruction ambiguities. In *ECCV*, 2008.

[8] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002.

[9] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981.

[10] W. Grimson and T. Lozano-Pérez. Localizing overlapping parts by searching the interpretation tree. *PAMI*, 1987.

[11] B. http://www.blender.org/.

[12] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 28(9):1465–1479, 2006.

[13] H. Ling and D. Jacobs. Deformation invariant image matching. In *ICCV*, 2005.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] T. McInerney and D. Terzopoulos. A finite element model for 3D shape reconstruction and nonrigid motion tracking. In *ICCV*, 1993.

[16] D. Metaxas and D. Terzopoulos. Constrained deformable superquadrics and nonrigid motion tracking. *PAMI*, 1993.

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[18] F. Moreno-Noguer, V. Lepetit, and P. Fua. Pose priors for simultaneously solving alignment and correspondence. In *ECCV*, 2008.

[19] F. Moreno-Noguer, M.Salzmann, V. Lepetit, and P. Fua. Capturing 3D stretchable surfaces from single images in closed form. In *CVPR*, 2009.

[20] C. Olson. Efficient pose clustering using a randomized algorithm. *IJCV*, 23(2):131–147, 1997.

[21] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. In *BMVC*, 2008.

[22] M. Salzmann and P. Fua. Reconstructing sharply folding surfaces: A convex formulation. In *CVPR*, 2009.

[23] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-form solution to non-rigid 3D surface registration. In *ECCV*, 2008.

[24] J. Zhu, S. Hoi, Z. Xu, and M. Lyu. An effective approach to 3D deformable surface tracking. In *ECCV*, 2008.