Exploring Ambiguities for Monocular Non-Rigid Shape Estimation

Francesc Moreno-Noguer[1], Josep M. Porta[1], and Pascal Fua[2]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2] Computer Vision Laboratory, EPFL, Lausanne, Switzerland
fmoreno@iri.upc.edu, porta@iri.upc.edu, pascal.fua@epfl.ch

Abstract.
Recovering the 3D shape of deformable surfaces from single images is difficult because many different shapes have very similar projections. This is commonly addressed by restricting the set of possible shapes to linear combinations of deformation modes and by imposing additional geometric constraints. Unfortunately, because image measurements are noisy, such constraints do not always guarantee that the correct shape will be recovered. To overcome this limitation, we introduce an efficient approach to exploring the set of solutions of an objective function based on point-correspondences and to proposing a small set of candidate 3D shapes. This allows the use of additional image information to choose the best one. As a proof of concept, we use either motion or shading cues to this end and show that we can handle a complex objective function without having to solve a difficult non-linear minimization problem.

Key words: 3D shape recovery, deformation model, nonrigid surfaces.

1 Introduction

It has been shown that the 3D shape of deformable surfaces can be recovered from even single images provided that enough correspondences can be established between that image and one in which the surface's shape is already known [1–3]. While effective, these techniques return one single reconstruction without accounting for the fact that several plausible shapes could produce virtually the same projection and therefore be indistinguishable on the basis of correspondences and geometry alone. In practice, as shown in Fig. 1 disambiguation is only possible using additional image information.

In this paper, we introduce an efficient way to sample the space of all plausible solutions. We achieve this by representing shape deformations in terms of a weighted sum of deformation modes and relating uncertainties on the location of point correspondences to uncertainties on the mode weights. This lets us

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

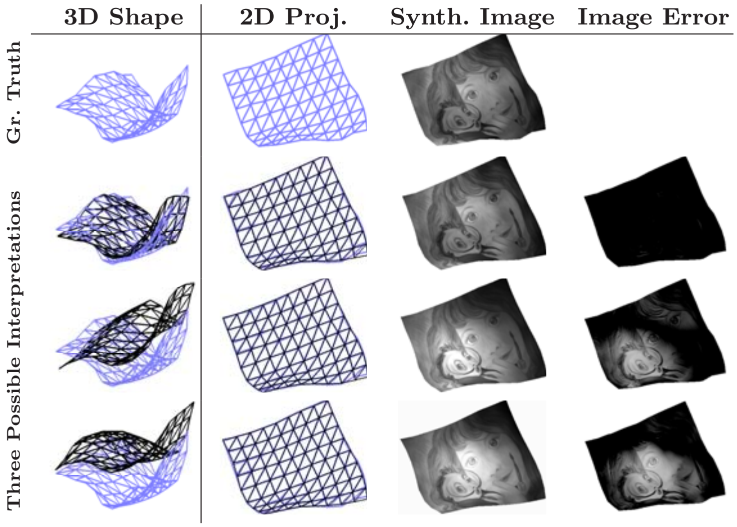


Fig. 1. Handling 3D shape ambiguities. **First Row.** Image of a surface lit by a nearby light source and the corresponding ground truth surface. **Three other Rows.** In each one, a different candidate surface proposed by our algorithm is shown in black. The corresponding projection and synthesized image given automatically estimated lighting parameters are shown in the middle columns. As can be seen, its projection is very similar, even though its shape may be very different from the original one. However, when comparing the true and synthesized images, it becomes clear that the correct shape is the one at the second row.

explore the space of modes and select a very small number of likely ones, which correspond to 3D shapes such as those shown in the left column of Fig. 1.

In this paper, as a proof of concept, we use either shading or motion information to select the best 3D shape among the candidates generated in this manner. When using shading, we show that we can exploit it both when the light sources are distant and when they are nearby. The latter is particularly significant because exploiting nearby light sources would involve solving a difficult non linear minimization problem if we did not have a reliable way to generate 3D shape hypotheses. In our examples, this is all the more true since the lighting parameters are initially unknown and must be estimated from the images. Alternatively, when a video is available, we can exploit three-frame sequences to pick the set of candidate 3D shapes that provides the most temporally consistent motion. We show that both these approaches outperform state-of-the-art methods [4, 5].

Summarizing, our contribution is an approach to avoiding being trapped in the local minima of a potentially complicated objective function by efficiently exploring the solution space of a simpler one. As a result, we only need to evaluate the full objective function for a few selected shapes, which implies we could use a very discriminating and expensive one if necessary.

2 Related Work

Single-view 3D reconstruction of non-rigid surfaces is known to be a highly under-constrained problem that cannot be solved without *a priori* knowledge. A typical approach to introducing such knowledge and reducing the space of possible shapes is to use deformation models [7–11]. Surface deformations are expressed as weighted sums of modes and retrieving shape entails estimating the modal weights by minimizing an image-based objective function. Since such functions usually have many local minima, a good initialization is required.

Several recent methods propose to recover the shape of inextensible surfaces without an explicit deformation model. Some are specifically designed for applicable surfaces, such as sheets of paper [12, 13]. Others constrain the distances between surface points to remain constant [6, 1]. This is generally applicable to many materials that do not perceptibly shrink or stretch as they deform.

Other approaches achieve shape-recovery either in closed form [4] or by solving a convex optimization problem [2], and thus, eliminate the need for an initialization. To this end, they require 2D point correspondences between the image in which one wishes to compute the shape and one in which it is already known. However, as will be shown in the results section, small inaccuracies in the correspondences can result in erroneous reconstructions.

The method proposed in this paper builds on the formalism introduced in [4] to return not a single solution but a representative set of *all* possible solutions and then uses additional information to decide which one is best. In this paper, we use shading or motion but any image cue could have been used instead.

Of course, many methods, such as [14, 15], have been proposed to merge geometric and shading cues into a common framework. However, these techniques, unlike ours, involve multiple iterative processes that require good initial estimates. An exception is the algorithm of [5] that solves for shape in closed form but is only applicable for Lambertian surfaces lit by a distant point light source.

3 Exploring the Space of Potential 3D Shapes

Let us assume that we are given a *reference image* in which the shape of a 3D deformable surface represented by a triangulated mesh is known and a set of 2D point correspondences between this reference image and an *input image* in which the shape is unknown. In [4], it was shown that this unknown 3D shape could be computed in closed form by representing the surface deformations in terms of a weighted sum of modes and picking the weights that minimize the reprojection errors while preserving the length of the mesh edges. However, the resulting shape is not always the right one, as shown in Table 1. This is because the correspondences are not infinitely accurate and the algorithm can trade a small amount of reprojection error against similarly small violations of the length constraints. As it turns out, this is enough to result in large changes in 3D shape since, as discussed earlier, very different shapes can have very similar projections.

To avoid this problem, we also represent the shape as a weighted sum of modes. But, instead of picking the best set of weights according to a geometric

	Shape # 1	**Shape # 2**	**Shape # 3**
Reconst. Error (mm)	0.82	4.25	5.35
Reproj. Error (pix)	1.92	1.87	1.93
Inextens. Error (mm)	4.00	4.27	3.97

Table 1. Mean reconstruction, reprojection and inextensibility errors for the candidate shapes of Fig.1. Note that, although shape#1 violates edge-length constraints slightly more than shape#3, it still is the reconstruction closest to the ground truth by far.

criterion, we fit a Gaussian distribution to those that correspond to acceptable projections. This lets us exhaustively sample the sets of weights that also preserve the length of the mesh edges. This typically results in approximately one hundred candidate shapes per image, among which the best can be picked using additional sources of shape information. In Section 4, we show that either shading or motion cues can be used for this purpose.

3.1 Problem Formulation

We represent our surface as a triangulated 3D mesh with n_v vertices $\mathbf{v}_i$ concatenated in a vector $\mathbf{x}$=$[\mathbf{v}_1^\top, \ldots, \mathbf{v}_{n_v}^\top]^\top$. We model surface deformations as weighted sums of n_m deformation modes $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{n_m}]$, obtained by applying Principal Component Analysis over a set of training meshes. We write

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

where $\mathbf{x}_0$ is a mean shape and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{n_m}]^\top$ are unknown weights that define the current surface shape.

As in [4, 5], we treat a correspondence between a 2D point $\mathbf{r}_i$ in the reference image and a 2D point $\mathbf{u}_i$ in the input image as a 2D-to-3D correspondence between $\mathbf{u}_i$ and $\mathbf{p}_i$, the 3D point on the mesh in its reference configuration that projects at $\mathbf{r}_i$. We express the coordinates of $\mathbf{p}_i$ in terms of the barycentric coordinates of the face to which belongs as $\mathbf{p}_i = \sum_{j=1}^{3} a_{ij} \mathbf{v}_j^{[i]}$, where the a_{ij} are the barycentric coordinates and the $\mathbf{v}_j^{[i]}$ are the vertices.

Assuming the matrix $\mathbf{A}$ of internal camera parameters to be known and that the 3D points are expressed in the camera referencial, the fact that $\mathbf{p}_i$ projects at $\mathbf{u}_i$ implies that

$$w_i \begin{bmatrix} \mathbf{u}_i \\ 1 \end{bmatrix} = \mathbf{A}\mathbf{p}_i = \begin{bmatrix} \mathbf{A}_{2\times3} \\ \mathbf{a}_3^\top \end{bmatrix} \mathbf{p}_i \ , \tag{2}$$

where w_i is a scalar, $\mathbf{A}_{2\times3}$ are the first two rows of $\mathbf{A}$ and $\mathbf{a}_3^\top$ the last one. Since $w_i = \mathbf{a}_3^\top \mathbf{p}_i$, we can write $\left(\mathbf{u}_i \mathbf{a}_3^\top - \mathbf{A}_{2\times3}\right) \mathbf{p}_i = \mathbf{0}$. By representing $\mathbf{p}_i$ with its barycentric coordinates, we then have

$$\sum_{j=1}^{3} a_{ij} \left(\mathbf{u}_i \mathbf{a}_3^\top - \mathbf{A}_{2\times3}\right) \mathbf{v}_j^{[i]} = \mathbf{0} \ \ . \tag{3}$$

In short, for each 3D-to-2D correspondence, Eq. 3 provides 2 linear constraints on $\mathbf{x}$. n_c such correspondences yield $2n_c$ constraints which can be written as a linear system $\mathbf{Mx} = \mathbf{0}$, where $\mathbf{M}$ is a $2n_c \times 3n_v$ matrix obtained from the known values a_{ij}, $\mathbf{u}_i$ and $\mathbf{A}$. Injecting the modal description of Eq. 1 then yields

$$\mathbf{MQ}\boldsymbol{\alpha} + \mathbf{Mx}_0 = \mathbf{0} \ , \tag{4}$$

such that any set of weights $\boldsymbol{\alpha}$ that is a solution of it corresponds to a surface that projects at the right place.

3.2 Proposing Candidate Shapes

Since correspondences $\{\mathbf{p}_i, \mathbf{u}_i\}$ are potentially noisy, the simplest way to solve Eq. 4 is in the least-squares sense. This, however, may not be satisfactory because $\mathbf{MQ}$ is an ill-conditioned matrix with several small eigenvalues [4, 5]. As a result, even when there are many correspondences, small changes in the exact correspondence locations, and therefore in the coefficients of $\mathbf{M}$, can result in very large changes of the resulting $\boldsymbol{\alpha}$ values. In other words, many different sets of $\boldsymbol{\alpha}$ weights can result in virtually the same projection. In [4], this is addressed by choosing the weights that best preserve the lengths of the mesh edges. However, as shown by Table 1, this does not necessarily yield the best answer.

In this paper, instead of choosing the best set of weights based on geometric considerations alone we have devised a way to quickly propose a restricted set of candidate solutions among which the best can be chosen using additional sources of image information, as will be done in Sections 4.1 and 4.2. To this end, we first derive an analytical expression of the solution space as a function of the 2D input data statistics. We then efficiently sample this space and keep the best samples in terms of both minimizing reprojection errors and preserving edge lengths.

Gaussian Representation of the Solution Space. The $\boldsymbol{\alpha}$ weights we seek can be computed as the least-squares solution of Eq. 4:

$$\boldsymbol{\alpha} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{b} \ , \tag{5}$$

where $\mathbf{B}$=$\mathbf{MQ}$ is a $2n_c{\times}n_m$ matrix, and $\mathbf{b}$=$-\mathbf{Mx}_0$ is a $2n_c$ vector. The components of $\mathbf{B}$ and $\mathbf{b}$ are linear functions of the known parameters a_{ij}, $\mathbf{u}_i$, $\mathbf{Q}$ and $\mathbf{A}$. We have seen that this solution may not, in fact, be the right one because $\mathbf{B}$ is ill-conditioned and solving the system in the least-squares sense magnifies small inaccuracies in the correspondences. We can nevertheless exploit the expression of Eq. 5 to model where to look for other potential solutions as follows.

Let us assume that the estimated correspondence locations are normally distributed around their true locations. Injecting the corresponding $2n_c{\times}2n_c$ diagonal covariance matrix $\boldsymbol{\Sigma}_\mathbf{u}$ into Eq. 5 means that the $n_m{\times}n_m$ covariance matrix for the $\boldsymbol{\alpha}$ weights can be written as $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \mathbf{J}_\beta \boldsymbol{\Sigma}_\mathbf{u} \mathbf{J}_\beta^\top$, where $\mathbf{J}_\beta$ is the $n_m{\times}2n_c$ Jacobian of $(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{b}$ with respect to the 2D correspondence coordinates:

$$\mathbf{J}_\beta = \frac{\partial(\mathbf{B}^\top\mathbf{B})^{-1}}{\partial\mathbf{u}} \mathbf{B}^\top\mathbf{b} + (\mathbf{B}^\top\mathbf{B})^{-1} \frac{\partial\mathbf{B}^\top\mathbf{b}}{\partial\mathbf{u}} \ . \tag{6}$$

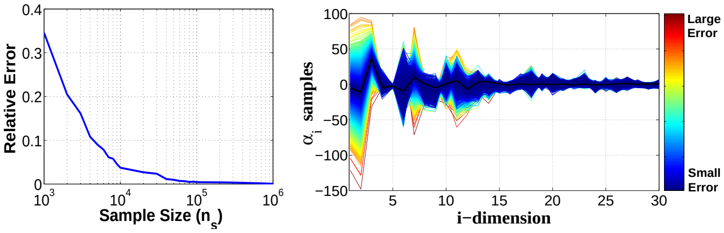


Fig. 2. Efficient exploration of the solution space. **Left:** Number of samples n_s needed to correctly approximate $\mathcal{R}_{\boldsymbol{\alpha}}$. We plot $\frac{\det(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}})-\det(\mathcal{M}^2\boldsymbol{\Sigma}_{\boldsymbol{\alpha}})}{\det(\mathcal{M}^2\boldsymbol{\Sigma}_{\boldsymbol{\alpha}})}$, an estimate of the distance between the theoretical covariance matrix and its empirical estimate from the samples. It diminishes quickly and becomes negligible for $n_s = 10^5$. **Right:** We represent each set of 30-dimensional $\boldsymbol{\alpha}$ weights by a line whose color encodes the value of the error of Eq. 9, according to the color-code at the right. The black line represents the ground truth. Note how well distributed the samples are around it.

We can therefore represent the family of 3D surfaces whose projections are close to the one that minimizes the reprojection error as being normally distributed around $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$, the least squares solution of Eq. 4, with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$. Note that, because $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is the solution of an ill-conditioned system, it is an unreliable estimate of the distribution's center. We could have improved the system's conditioning by adding a damping term, but this would have amounted to arbitrarily constraining the norm of $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$. Instead, as discussed in the next section, we use a sampling mechanism to explore different possible values of $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$.

Efficiently Exploring the Solution Space. To create a set of plausible 3D shapes whose projection are acceptably close to the correct one, we first define a search region $\mathcal{R}_{\boldsymbol{\alpha}}$ in n_m-dimensional space. We then sample it using a standard numerical technique and progressively apply more stringent constrains to an ever decreasing number of samples.

Given the normal distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ introduced above, we take $\mathcal{R}_{\boldsymbol{\alpha}}$ to be made of the $\boldsymbol{\alpha}_i$ such that

$$(\boldsymbol{\alpha}_i - \boldsymbol{\mu}_{\boldsymbol{\alpha}})^\top \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}(\boldsymbol{\alpha}_i - \boldsymbol{\mu}_{\boldsymbol{\alpha}}) \le \mathcal{M}^2 \quad , \tag{7}$$

where $\mathcal{M}$ is a threshold chosen to achieve a specified degree of confidence. To compute its value we use the cumulative chi-squared distribution, which depends on the dimensionality of the problem . In our experiments, we use $n_m = 30$ modes and $\mathcal{M} = 7$ yields a 98% level of confidence.

To sample $\mathcal{R}_{\boldsymbol{\alpha}}$, we draw n_s random samples $\{\tilde{\boldsymbol{\alpha}}_i\}_{i=1}^{n_s}$ from the distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \mathcal{M}^2\boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$. Let $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}$ be the mean and covariance matrix of these samples. The technique we use guarantees that $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}}$ and that the difference between $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}$ and $\mathcal{M}^2\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ approaches zero as n_s increases [16].

In practice, as the $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ we use is unreliable, we do not draw all n_s samples at once. Instead, we draw successive batches and, having drawn batch k, we draw the next one by sampling from the distribution centered around

$$\boldsymbol{\mu}_{\boldsymbol{\alpha}}^{k} = \frac{\sum_{i=1}^{n_s^k} \pi_i^k \tilde{\boldsymbol{\alpha}}_i^k}{\sum_{i=1}^{n_s^k} \pi_i^k} \quad , \tag{8}$$

where the π_i^k are weights associated to individual samples, computed as follows.

Let $\tilde{\mathbf{x}} = [\tilde{\mathbf{v}}_1^\top, \ldots, \tilde{\mathbf{v}}_{n_v}^\top]^\top$ be the mesh computed using sample $\tilde{\boldsymbol{\alpha}}$, and let $\{\tilde{\mathbf{u}}_i\}_{i=1}^{n_c}$ be the 2D projections of the 3D points for which correspondences $\mathbf{u}_i$ are available. $\tilde{\boldsymbol{\alpha}}$ is assigned the weight π such that

$$1/\pi \sim \lambda_1 \sum_{i}^{n_c} \|\tilde{\mathbf{u}}_i - \mathbf{u}_i\| + \lambda_2 \sum_{\{i,j\} \in \mathcal{N}} \|\tilde{l}_{ij} - l_{ij}^{ref}\| \quad , \tag{9}$$

where the two terms account for the reprojection and inextensibility errors, respectively. Since these errors are expressed in different units of measurement, we use λ_1 and λ_2 to give them similar orders of magnitude. In addition, $\tilde{l}_{ij}$ is the distance between two neighboring vertices $\tilde{\mathbf{v}}_i$ and $\tilde{\mathbf{v}}_j$, l_{ij}^{ref} is the distance between the same vertices in the reference configuration, and $\mathcal{N}$ represents the indices of neighboring vertices.

In our experiments we used $n_s = 10^5$ random samples, which as shown in Fig. 2(Left), approximate $\mathcal{R}_{\boldsymbol{\alpha}}$ with an error below 0.5%. These samples were drawn in 10 consecutive batches of 10^4 samples each. As depicted by Fig. 2(Right) the samples generated in this way densely cover a large region of the solution space around the true one. To reduce their number and speed up further processing, we only keep the 10% of the samples with highest weight.

By construction, all these samples represent shapes that yield similar projections and only small violations of the length constraints. Furthermore, many of them yield almost undistiguinshable 3D shapes. To further reduce their number, we therefore run a Gaussian-means clustering algorithm over all the remaining samples in the space of the 3D coordinates [17]. This is a variant of the k-means algorithm that automatically identifies the optimal number of clusters based on statistical tests designed to check whether all the clusters follow a Gaussian distribution. These tests are controlled by means of a significance level parameter which we set to a very low value to favor over-segmentation, that is, to produce more clusters than absolutely necessary to avoid grouping shapes whose difference is statistically significant.

Finally, we take our candidates set of shapes to be the cluster centers. This whole process typically reduces the initial 10^5 samples to about one hundred.

4 Using Additional Cues to Select the Best Candidate

Given correspondences between the reference image and the input image, the algorithm discussed in the previous section returns about 100 candidate 3D meshes that all project correctly in the input image and whose edges have retained their original length. In this section, we show how to use either lighting or motion cues to disambiguate and pick the best one.

4.1 Shading Cues

We consider two different cases. First, we assume the surface is lit by a *distant light source*, which is the situation envisioned in earlier works on monocular deformable surface reconstruction that use shading clues [14, 15, 5]. Second, we address the situation in which the surface is lit by a *nearby light source*. This is more difficult because the inverse of the changing distance to the light source has to be taken into account, which rules out approaches based on simple linear or quadratic constraints. In both cases, we do not assume the lighting parameters known *a priori* and estimate them from the candidate 3D shapes. As shown in Fig. 1, this lets us render the image we would see for any candidate shape, compare it to the real one, and select the best. To perform the rendering, we use ray-tracing and take into account visibility effects and shadows cast by the object on itself. Such non-local and non-linear phenomena are rarely taken into account by continuous optimization-based schemes because they result in highly complex energy landscapes and poor convergence. We now turn to the estimation of the lighting parameters in these two cases.

Light Source at Infinity. Recall from Section 3.1, that we start from a set of correspondences between 3D surface points $\mathbf{p}_i$ and 2D image points $\mathbf{u}_i$ in the input image with intensity I_i. For each i, we also know that $\mathbf{p}_i$ projects at $\mathbf{r}_i$ in the reference image and has intensity I_i^{ref}. In practice, we acquire the reference image under diffuse lighting so that, assuming the surface to be Lambertian, we can take the albedo ρ_i of $\mathbf{p}_i$ to be I_i^{ref}. In the remainder of this Section, let $\mathbf{p}_i$ denote the 3D coordinates of the 3D surface points in the candidate shapes. For each candidate shape, these $\mathbf{p}_i$ are recomputed using the barycentric coordinates, which are the same for all candidates, to average the 3D vertex coordinates of the facets they belong to.

Assuming a distant light source parameterized by its unit direction $\mathbf{l}$ and power L, we can write $I_i = \rho_i L(\mathbf{l} \cdot \mathbf{n}_i)$, where $\mathbf{n}_i$ is the surface normal at $\mathbf{p}_i$, which may be estimated from the $\mathbf{v}_i$ vertex coordinates. Grouping these equations for all n_c correspondences yields

$$\mathbf{I}_\rho = \mathbf{N}\mathbf{L} \quad , \tag{10}$$

where $\mathbf{I}_\rho = [I_1/\rho_1, \dots, I_{n_c}/\rho_{n_c}]^\top$, $\mathbf{N} = [\mathbf{n}_1, \dots, \mathbf{n}_{n_c}]^\top$, and $\mathbf{L} = L \cdot \mathbf{l}$. Solving this system in the least-squares sense yields an estimation of $\mathbf{L}$, from which the light intensity and direction can be taken to be $L = \|\mathbf{L}\|$ and $\mathbf{l} = \mathbf{L}/L$.

Nearby Light Source. When considering light sources that are not located at infinity, the fact that the radiosity due to individual light sources decreases with the square of the distance must be taken into account. The image irradiance at $\mathbf{p}_i$ therefore becomes

$$I_i = \rho_i L \frac{\mathbf{l}_i \cdot \mathbf{n}_i}{\|\mathbf{p}_i - \mathbf{s}\|^2} \tag{11}$$

where $\mathbf{s}$ is the position of the light source and $\mathbf{l}_i = \frac{1}{\|\mathbf{p}_i - \mathbf{s}\|}(\mathbf{p}_i - \mathbf{s})$. $\mathbf{s}$ and L are estimated by minimizing

$$\sum_{i=1}^{n_c} \left| I_i - \rho_i L \frac{\mathbf{l}_i \cdot \mathbf{n}_i}{\|\mathbf{p}_i - \mathbf{s}\|^2} \right| \ , \tag{12}$$

with respect to L and $\mathbf{s}$ using the nonlinear least-squares matlab routine `lsqnonlin`. To avoid local minima, we define a sparse set of light positions $\{\tilde{\mathbf{s}}_j\}_{j=1}^{n_l}$ and use each one in turn to initialize the optimization. In our experiments, we used $n_l = 125$ light positions uniformly distributed within a hemisphere on top of the reference mesh. Its radius was taken to be sufficiently large to include all distances for which the nearby light assumption holds.

Note that what makes this approach computationally feasible is the fact that we are only attempting to recover the lighting parameters, while fixing the shape parameters. Otherwise, the problem would be massively underconstrained. This should also allow the use of more sophisticated lighting models [18] to relax the single light and Lambertian assumptions.

4.2 Motion Cues

When video sequences are available, we can rely on temporal consistency between consecutive shapes to select the most likely ones. Let us assume that a second order autoregressive model [19] has been learned from training data. Given such a model, the shape at time t, $\mathbf{x}^t$, can be expressed as function of the shapes at times $t-1$ and $t-2$ as

$$\mathbf{x}^t = \hat{\mathbf{A}}_2 \mathbf{x}^{t-2} + \hat{\mathbf{A}}_1 \mathbf{x}^{t-1} + \hat{\mathbf{B}} \mathbf{w}^t \ , \tag{13}$$

where $\hat{\mathbf{A}}_2$, $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{B}}$ are $3n_v \times 3n_v$ matrices learned offline, and $\mathbf{w}^t$ is an n_v Gaussian noise vector.

For any three consecutive images and the corresponding shape samples, the most plausible shape in the third one can be found by considering all $\{\tilde{\mathbf{x}}_i^{t-2}, \tilde{\mathbf{x}}_j^{t-1}, \tilde{\mathbf{x}}_k^{t}\}$ triplets and picking the $\tilde{\mathbf{x}}_k^t$ belonging to the one that best satisfies Eq. 13. Since this is done independently at each time step t, we are not imposing temporal consistency beyond our three consecutive frames windows.

5 Results

We compare the performance of our approach on synthetic and real sequences against that of two state-of-the-art techniques [4, 5], which we refer to as *Salzm08* and *Moreno09*, respectively. As discussed in Section 2, the first essentially returns the approximate solution of Eq. 4 that minimizes the variations in edge-length from the reference shape while the second returns the solution that best fits a shading model involving a point light source at infinity. Note that all three methods compute the 3D shape from either individual images or consecutive triplets, without enforcing temporal consistency across the sequence. We can therefore treat their results as independent and compute their statistics.

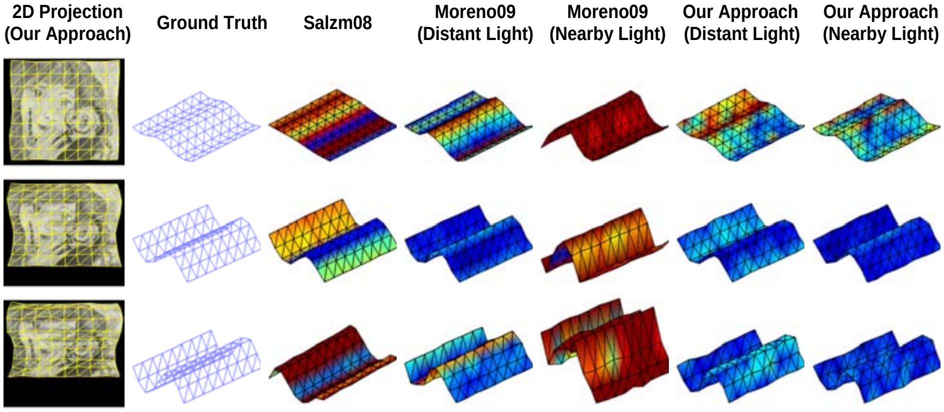

2D Projection (Our Approach)
Ground Truth
Salzm08
Moreno09 (Distant Light)
Moreno09 (Nearby Light)
Our Approach (Distant Light)
Our Approach (Nearby Light)

Fig. 3. Results for the synthetic wave sequence. They are best viewed in color as deviations from the ground truth are encoded according the color-code of Fig. 2. Errors of more than 75% of the maximum amplitude of the ground truth shape appear in red.

5.1 Synthetic Experiments

We created two synthetic data sets by deforming an initially planar 9×9 mesh of 30×30 cm. In the first case, we created 500 meshes such as the one of Fig. 1 by randomly changing the angles between neighboring facets. In the second case, we built 250 meshes by giving the surface a wave-like shape, as shown in Fig. 3. In both cases, the virtual camera was placed approximately 75 cm above the mesh and we used a real image as a texture-map. We synthesized a shaded image by selecting a random light-source direction in the hemisphere above the mesh. The light was located either infinitely far or within 30cm of the mesh center. We then produced 100 random 3D-to-2D correspondences between the reference configuration and individual deformed meshes and added a 2-pixel standard deviation Gaussian noise to the 2D coordinates. To compare the sensitivity of Moreno09 and of our approach to lighting conditions, for each synthetic shape we computed two different estimates, one using the image rendered using the distant light and the other using the nearby light.

Fig. 3 depicts results on the synthetic wave sequence using Salzm08, Moreno09, and our own approach in conjunction with either the distant or the nearby lighting. In Fig. 4, we use boxplots[1] to summarize them. We also include the output of an hypothetical algorithm that would be able to select the best candidate shape among all the samples produced by the sampling mechanism of Section 3, which represents the theoretical optimum an algorithm like ours could achieve by using the image information as effectively as possible. Our method consistently returns a lower 3D reconstruction error. This is true even though the

[1] Box denoting the first $\mathcal{Q}1$ and third $\mathcal{Q}3$ quartiles, a horizontal line indicating the median, and a dashed vertical line representing the data extent taken to be $\mathcal{Q}3 + 1.5(\mathcal{Q}3 - \mathcal{Q}1)$. The red crosses denote points lying outside of this range.

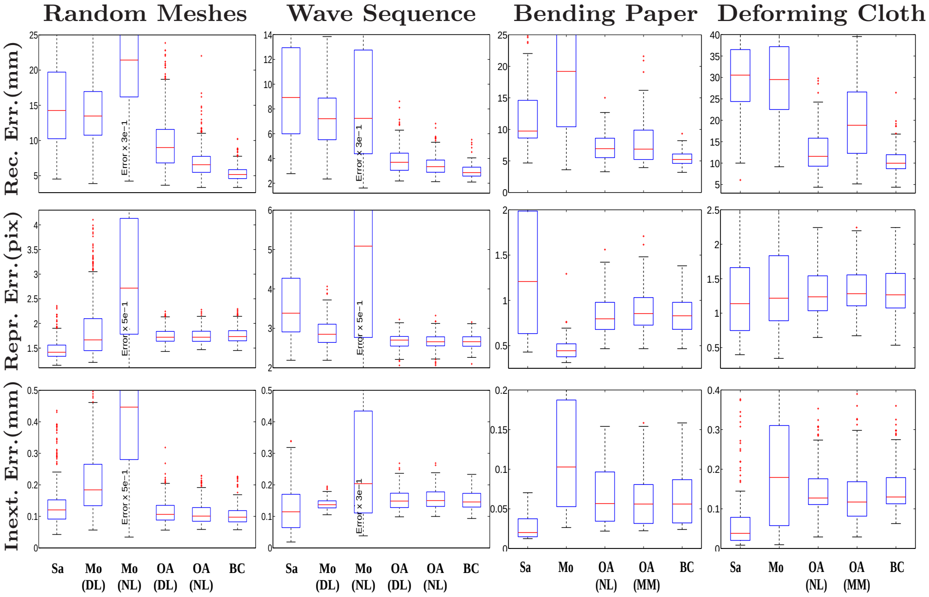


Fig. 4. In each column, reconstruction, reprojection, and inextensibility errors for each of the two synthetic and the two real sequences. Sa: Salzmann08. Mo: Moreno09. OA: Our Approach. BC: An hypothetical algorithm that would always choose the Best Candidate. DL: Distant Light. NL: Nearby Light. MM: Motion Model.

reprojection and inextensibility errors are very similar for all three methods, which confirms that minimizing these is not sufficient by itself to retrieve the correct 3D shape. Both Moreno09 and our approach address this issue by taking advantage of shading cues. Since we explicitly model a nearby light, we clearly outperform Moreno09 in that case.

Another measure of success is the *Percentage of correct solutions* of Table 2. Given the ground truth solution, a 3D sample mesh is considered to be correct if at least 75% of its vertices have a reconstruction error smaller than 0.5×Height, where *Height* refers to the maximum amplitude of the ground truth shape. Again, our approach clearly yields the best numbers. The specific ratios –75% and 0.5×Height– are of course *ad hoc* and have been chosen so that 3D meshes that are deemed incorrect produce disturbing effects when viewed in sequence. To provide the reader with an intuitive understanding of what this measure actually represents, in Fig. 3 facets with reconstruction errors of more than 75% are color-coded in red.

Finally, the table at the top of Fig. 5 depicts the accuracy of the estimated lighting parameters. Note that we estimate the position and direction of a light source that was allowed to move freely within a 30 cm radius hemisphere with an accuracy of less than a 1 cm and 10 degrees.

	Salzm08	**Moreno09**		**Our Method**			**Best Cand.**
		DL	NL	DL	NL	MM	
Random Meshes	84	81	15	91	99	–	100
Wave Sequence	78	95	31	100	100	–	100
Bending Paper	80	–	43	–	99	96	100
Deforming Cloth	59	–	57	–	97	81	99

Table 2. Percentages of correct solutions for all four set of experiments. DL: Distant Light. NL: Nearby Light. MM: Motion Model.

5.2 Real Images

We tested our approach on a 120-frames sequence of bending paper and a 150-frame sequence of a deforming T-shirt, both acquired with a Pointgrey BumBlebee stereo rig. The surfaces were lit by a dim ambient lighting and a light source located about 30 cm from the surface. We used the stereo pairs to estimate the ground truth shape and then ran our algorithms using the output of a single camera. We used SIFT [20] to establish correspondences between the reference and input images. In both experiments we used the algorithm described in Sect. 3 to initially produce a set of candidate 3D shapes in each individual frame. We then chose the best using either shading or motion information.

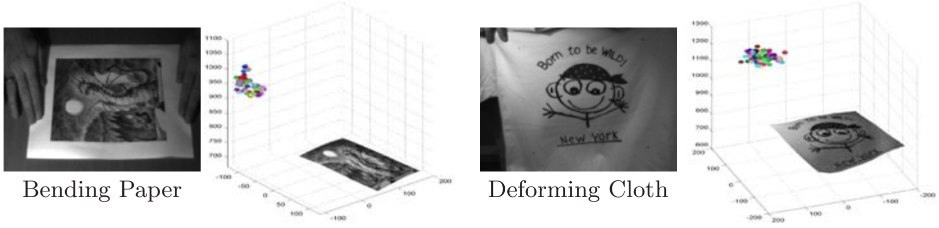
Using Shading to Disambiguate. When using shading, the reconstruction errors depicted in the two right-most boxplots of Fig. 4 exhibit the same patterns as those obtained for the synthetic sequences, which confirms that our method outperforms the other two. As shown in Table 2, we obtain 97% of correct solutions, which represents a 30% increase in performance, using the same definition of "correct" as before. In the bottom of Fig. 5, we plot the estimated light source positions in each frame. Although we did not measure the exact light source locations, the fact that the estimates are tightly clustered is an indication that they are probably correct, given that they all were obtained independently.

Using Motion to Disambiguate. To learn the autoregressive model of Sect. 4.2, we used additional sequences, obtained ground truth data using our stereo rig, and learned the model parameters by probabilistic fitting [19]. In the case of the sheet of paper, as shown in the third column of Fig. 4 and in Table 2, using the motion model yields results similar to those obtained using shading. The performance degrades slightly in the case of cloth because our second order motion model is not accurate enough to perfectly capture the sharp cloth deformations. Nevertheless, our method still outperforms both Salzm08 and Moreno09.

6 Conclusion

For the purpose of single view 3D non-rigid reconstruction, approaches that rely on purely geometric constraints can return incorrect answers because several

	Distant Light		**Nearby Light**	
	Direction Err(deg)	Power Err(%)	Position Err(mm)	Power Err(%)
Random Meshes	6.9 ± 4.3	5.2 ± 2.1	7.4 ± 6.1	6.8 ± 3.3
Wave Sequence	2.1 ± 0.9	2.2 ± 0.8	3.2 ± 0.8	2.8 ± 1.0



Bending Paper

Deforming Cloth

Fig. 5. Estimated lighting parameters. Upper table: Mean error and standard deviation of the lighting parameters–direction, position and power– estimated independently in each frame of the synthetic sequences. Bottom figures: Light source positions estimated independently in all frames of the real sequences. Note how well clustered they are.

different shapes that obey, or nearly obey these constraints, often yield very similar projections. To overcome this problem given that the input data is noisy, we use error propagation techniques to derive an analytical expression of the space of potential candidate shapes and to propose a small but representative number of samples. The best among them can then be chosen based on additional image cues, such as shading or motion, which significantly improves results with respect to state-of-the-art methods.

References

1. Perriollat, M., Hartley, R., Bartoli, A.: Monocular template-based reconstruction of inextensible surfaces. In: BMVC. (2008)
2. Salzmann, M., Fua, P.: Reconstructing sharply folding surfaces: A convex formulation. In: CVPR. (2009)
3. Zhu, J., Hoi, S., Xu, Z., Lyu, M.: An effective approach to 3d deformable surface tracking. In: ECCV. (2008)
4. Salzmann, M., Moreno-Noguer, F., Lepetit, V., Fua, P.: Closed-form solution to non-rigid 3d surface registration. In: ECCV. (2008)
5. Moreno-Noguer, F., M.Salzmann, Lepetit, V., Fua, P.: Capturing 3d stretchable surfaces from single images in closed form. In: CVPR. (2009)
6. Ecker, A., Jepson, A.D., Kutulakos, K.N.: Semidefinite programming heuristics for surface reconstruction ambiguities. In: ECCV. (2008)
7. Cohen, L., Cohen, I.: Finite-element methods for active contour models and balloons for 2-d and 3-d images. PAMI **15** (1993)
8. McInerney, T., Terzopoulos, D.: A finite element model for 3d shape reconstruction and nonrigid motion tracking. In: ICCV. (1993)
9. Metaxas, D., Terzopoulos, D.: Constrained deformable superquadrics and nonrigid motion tracking. PAMI **15** (1993)

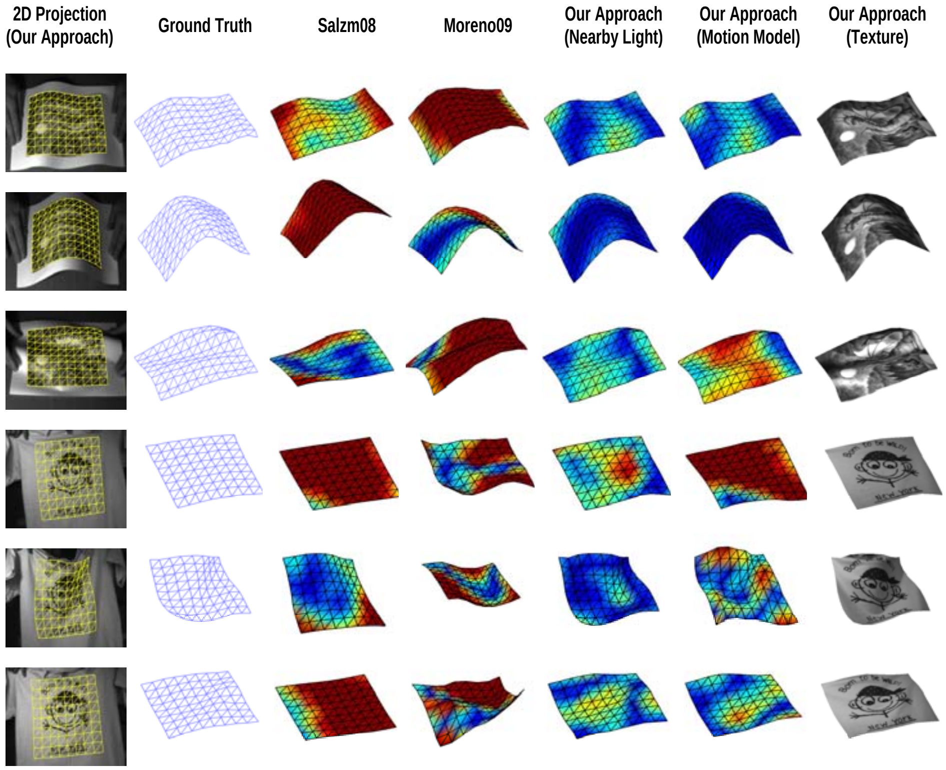

2D Projection (Our Approach)
Ground Truth
Salzm08
Moreno09
Our Approach (Nearby Light)
Our Approach (Motion Model)
Our Approach (Texture)

Fig. 6. Results for the two real sequences. Top three rows: Paper. Bottom three rows: Cloth. The reconstruction errors are again color-coded.

10. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3–d faces. In: SIGGRAPH. (1999)
11. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. In: ECCV. (1998)
12. Gumerov, N., Zandifar, A., Duraiswami, R., Davis, L.: Structure of applicable surfaces from single views. In: ECCV. (2004)
13. Liang, J., DeMenthon, D., Doermann, D.: Flattening curved documents in images. In: CVPR. (2005)
14. Wang, Y., Liu, Z., G., Hua, Wen, Z., Zhang, Z., Samaras, D.: Face re-lighting from a single image under harsh lighting conditions. In: CVPR. (2007)
15. White, R., Forsyth, D.: Combining cues: Shape from shading and texture. In: CVPR. (2006)
16. Ruanaidh, J., Fitzgerald, W.: Numerical Bayesian Methods Applied to Signal Processing. Springer (1996)
17. Hamerly, G., Elkan, C.: Learning the k in k-means. In: NIPS. (2003)
18. Hara, K., Nishino, K., Ikeuchi, K.: Light source position and reflectance estimation from a single view without the distant illumination assumption. PAMI (2005)
19. Blake, A., Isard, M.: Active contours. Springer (1998)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)