Dense Variational Reconstruction of Non-Rigid Surfaces from Monocular Video*

Ravi Garg Anastasios Roussos Lourdes Agapito
School of EECS, Queen Mary University of London
[rgarg,troussos,lourdes]@eecs.qmul.ac.uk

Abstract

*This paper offers the first variational approach to the problem of dense 3D reconstruction of non-rigid surfaces from a monocular video sequence. We formulate non-rigid structure from motion (NRS*f*M) as a global variational energy minimization problem to estimate dense low-rank smooth 3D shapes for every frame along with the camera motion matrices, given dense 2D correspondences.*

*Unlike traditional factorization based approaches to NRS*f*M, which model the low-rank non-rigid shape using a fixed number of basis shapes and corresponding coefficients, we minimize the rank of the matrix of time-varying shapes directly via trace norm minimization. In conjunction with this low-rank constraint, we use an edge preserving total-variation regularization term to obtain spatially smooth shapes for every frame. Thanks to proximal splitting techniques the optimization problem can be decomposed into many point-wise sub-problems and simple linear systems which can be easily solved on GPU hardware. We show results on real sequences of different objects (face, torso, beating heart) where, despite challenges in tracking, illumination changes and occlusions, our method reconstructs highly deforming smooth surfaces densely and accurately directly from video, without the need for any prior models or shape templates.*

1. Introduction

Recovering completely dense 3D models of a scene observed by a moving camera, where an estimate of its 3D location is obtained for every pixel in the image, is a key problem in computer vision. Rigid structure from motion (S*f*M) algorithms have made significant progress towards this goal, with dense approaches to *multi-view stereo* (MVS) [14, 28] able to acquire highly accurate models from a collection of fully calibrated images. Recent variational approaches to (S*f*M) have even allowed to perform *real-time* dense reconstruction of rigid scenes [19] while estimating the unknown camera motion from live video acquired with a handheld camera; or to deal with sceces containing multiple independently moving rigid objects [24].

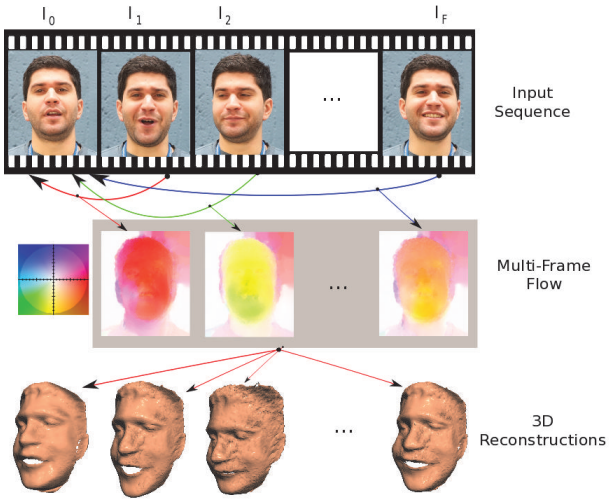


Figure 1. Our proposed pipeline for dense NRS*f*M. The first row shows the input image stream. Dense long-term 2D trajectories are first computed for every pixel in the reference frame using [15] and used as input to our dense NRS*f*M algorithm.

These dense S*f*M approaches produce impressive and detailed models of 3D objects purely from video sequences. However, their common drawback is that they can only handle scenes with rigid objects. In contrast, the field of non-rigid structure from motion (NRS*f*M) focuses on the reconstruction of deformable objects from video. Results from this field have significantly advanced in recent years in terms of their ability to reconstruct strong realistic non-rigid motions [10, 26, 30] and to recover from its inherent ambiguities with the use of additional priors on the deformations or the camera motion [4, 32]. In particular, the popular low-rank shape constraint, first proposed by Bregler *et al.* [7], has recently been shown to provide sufficient prior information, together with camera orthonormality constraints, to constrain the problem [1] and avoid ambiguous solutions and practical algorithms have followed [12, 20].

*This research was funded by the European Research Council under ERC Starting Grant agreement 204871-HUMANIS. The authors wish to thank C. Russell and S. Vicente for fruitful discussions.

However, in contrast to their rigid counterparts, NRS*f*M methods are typically sparse, *i.e*. they can only reconstruct a small set of salient points. This results in very low resolution 3D models that cannot capture fine detail. The leap to dense non-rigid shape estimation has been slowed down by the requirement for dense long-term 2D correspondences which are particularly challenging to obtain in the presence of non-rigid motion and large displacements. In practice, it is only recently that robust methods have emerged that can provide dense 2D trajectories for image points throughout the full sequence when the scene contains non-rigid motion [15, 22, 23].

In this paper we take advantage of these recent advances in variational optimization approaches to both: *(i)* dense estimation of rigid 3D shape(s) from video [19, 24] and *(ii)* dense estimation of long-term 2D trajectories in videos of non-rigid motion [15, 23]. Using a multi-frame motion flow field as input, we adopt a variational setting to provide an energy optimization approach to NRS*f*M that can provide 3D estimates for all the pixels in a reference frame of a video sequence. The novelty of our method resides in combining the low-rank shape prior with a powerful edge preserving spatial regularization prior that estimates smooth but detailed non-rigid shapes. Our results show that spatial smoothness can act as an important additional cue to help resolve the ambiguities inherent to the NRS*f*M problem and acquire accurate non-rigid shapes.

2. Related work

In their seminal work, Bregler *et al*. [7] pioneered the first solution to non-rigid structure from motion (NRS*f*M) by extending Tomasi and Kanade's [31] rigid factorization approach. Their insight was to incorporate a statistical shape prior on the time evolving non-rigid shape into the factorization formulation. This prior was expressed as a low-rank shape constraint: the 3D shape at any frame in the sequence can be expressed as a linear combination of an *unknown low-rank shape basis* governed by time-varying coefficients. Although this prior has proved to be a powerful constraint and led to a wealth of solutions, in isolation it is not sufficient to recover unambiguous deformable shape and camera motion from video. The problem remains ill-posed and the focus of NRS*f*M methods has been to resolve the inherent ambiguities, in particular to solve the core problem of upgrading the solutions to metric space, using additional constraints [4, 6, 13, 20, 32].

Most approaches have required the addition of extra priors on the shape or camera matrices to resolve the ambiguities such as: temporal smoothness [4, 32], near rigidity (rigid component explains most of the motion) [4, 32], smooth time trajectories [2, 16, 21], basis priors [4] or use of a pre-defined trajectory basis (such as DCT) [2, 21]. However, recently it was shown that orthonormality constraints on the camera matrix were sufficient in addition to the low-rank shape prior [1], and practical methods have emerged [12, 20] .

Most NRS*f*M methods impose the low-rank constraint explicitly by parameterizing the non-rigid shapes using a pre-defined number of basis shapes and time-varying coefficients. However, Dai *et al*. [12] recently noted that it is this explicit representation of non-rigid shape in terms of basis and coefficients that leads to additional basis ambiguities. Instead, they imposed the low-rank shape constraint directly on the matrix of time-varying shapes via trace norm minimization as the tightest possible relaxation of rank minimization. Trace norm has also been successfully used in compressed sensing and matrix completion [8] and more recently for factorisation based rigid structure from motion [3].

Similarly to Dai *et al*. [12], in this paper we adopt a trace norm minimization approach to estimate low-rank non-rigid shapes. However, our method departs substantially from Dai *et al*.'s by: *(i)* making the problem scalable to the use of a dense multi-frame flow field as input to the NRS*f*M problem; *(ii)* embedding the low-rank shape constraint within a global energy minimization framework which allows to incorporate powerful spatial regularization and recover smooth 3D shapes; and *(iii)* eliminating the requirement that the exact number of basis shapes be known in advance – in contrast, an important drawback of Dai *et al*.'s approach is that it still requires this information for the metric upgrade step. The result is the first dense template-free formulation of NRS*f*M. Our approach provides robust dense 3D estimates for every pixel in the reference image of a time-varying shape without the use of any prior models, using only the original footage.

Previous attempts to dense NRS*f*M have come from: piecewise approaches that reconstruct local patches using simple local models [10, 27] but require a post-processing step to stitch all the local reconstructions into a single smooth surface, and template based approaches [5] that require a 3D template to be provided.

Our system Given a video sequence acquired with a single camera as input, our approach provides a complete pipeline for dense NRS*f*M integrating 2D image matching and 3D reconstruction in two steps (Figure 1 illustrates our approach):

Dense 2D correspondences: First dense 2D correspondences are established in the image sequence. Here we take advantage of recent advances in robust and dense variational multi-frame motion estimation. While most works on 2D motion estimation for video sequences focus on estimating frame-to-frame optical flow fields, recently new Lagrangian approaches have been proposed that can handle the estimation of long-term trajectories that associate

each world point with its entire 2D image trajectory over an image sequence [15, 23]. In other words, these methods allow the computation of dense 2D correspondences from a reference frame to each of the subsequent images in the sequence which is the essential information needed for 3D reconstruction. These methods impose subspace constraints – the 2D trajectories are assumed to lie on a low-dimensional space – that implicitly act as a trajectory regularization term. This in turn leads to temporally consistent motion fields and allows to cope with occlusions and large displacements caused by non-rigid motion. More specifically, we adopt the formulation of [15] for colour images with soft subspace constraints.

Dense 3D reconstruction: Given these dense correspondences as input, our new variational energy optimization approach alternates between solving for the camera matrices and the non-rigid shape for every frame in the sequence. Our energy combines: *(i)* a geometric data term that minimizes image reprojection error, *(ii)* a trace norm term that minimizes the rank of the time-evolving shape matrix and *(iii)* an edge-preserving spatial regularization term that provides smooth 3D shapes.

3. Problem formulation

Consider an image sequence $I_1, \dots, I_F$ of F frames with N pixels each where I_{ref} is chosen to be the reference frame (this will often be the first frame). The input to our algorithm is a set of *dense* 2D tracks that have been estimated in a pre-processing step. For every pixel in the reference image I_{ref}, each track encodes its image location in the subsequent F frames. Let $p = 1, \dots, N$ be an index for the pixels and (x_{fp}, y_{fp}) the location of the p-th point in the f-th frame, $f = 1, \dots, F$. Note that, in the reference frame, this location coincides with the location of the p-th pixel on the image grid.

We adopt an orthographic camera model, where the 2×3 camera matrix $\boldsymbol{R}_f$ projects 3D points (X_{fp}, Y_{fp}, Z_{fp}) onto image frame f following the projection equation:

$$\underbrace{\begin{bmatrix} x_{f1} & \cdots & x_{fN} \\ y_{f1} & \cdots & y_{fN} \end{bmatrix}}_{\boldsymbol{W_f}} = \boldsymbol{R}_f \underbrace{\begin{bmatrix} X_{f1} & \cdots & X_{fN} \\ Y_{f1} & \cdots & Y_{fN} \\ Z_{f1} & \cdots & Z_{fN} \end{bmatrix}}_{\boldsymbol{S_f}} \quad (1)$$

where $\boldsymbol{W_f}$ stores the 2D locations of all N points in frame f and the $3 \times N$ matrix $\boldsymbol{S_f}$ represents the 3D shape observed in the frame f. Since the objects we are observing are non-rigid, the shape matrix $\boldsymbol{S_f}$ will be different for each frame. Note that we have eliminated the translation component from (1) by registering the image coordinates to the centroid in each frame f. Stacking equation (1) vertically for every frame $f \in \{1. \cdots, F\}$, we can now formulate the projection of the time varying shapes in all the frames as:

$$\boldsymbol{W} = \boldsymbol{R}\boldsymbol{S} \quad (2)$$

where, $\boldsymbol{W}$ is the input measurement matrix that contains the full 2D tracks, $\boldsymbol{S}$ is the non-rigid shape matrix and $\boldsymbol{R}$ is the motion matrix:

$$\underbrace{\boldsymbol{W}}_{2F \times N} = \begin{bmatrix} \boldsymbol{W_1} \\ \vdots \\ \boldsymbol{W_F} \end{bmatrix}, \underbrace{\boldsymbol{R}}_{2F \times 3F} = \begin{bmatrix} \boldsymbol{R}_1 & & \bigcirc \\ & \ddots & \\ \bigcirc & & \boldsymbol{R}_F \end{bmatrix}, \underbrace{\boldsymbol{S}}_{3F \times N} = \begin{bmatrix} \boldsymbol{S_1} \\ \vdots \\ \boldsymbol{S_F} \end{bmatrix} \quad (3)$$

We now define the problem of NRS*f*M as the joint estimation of: *(i)* the set of orthographic camera matrices [1] $\boldsymbol{R}$, and *(ii)* the set of 3D shapes $\boldsymbol{S}$ or equivalently the 3D coordinates (X_{fp}, Y_{fp}, Z_{fp}) of every point in every frame. The matrix $\boldsymbol{S}$ can be also be interpreted as the trajectory matrix, since its columns correspond to the 3D trajectories of each point. It is also useful in our formulation to represent $\boldsymbol{S}$ using the "permutation" operator $P(\boldsymbol{S})$ that re-arranges the entries of $\boldsymbol{S}$ into a $F \times 3N$ matrix such that the f-th row of $P(\boldsymbol{S})$ contains the X, Y and Z coordinates of all points of the shape at frame f (*i.e.* all values of $\boldsymbol{S}_f$).

4. Dense reconstruction with trace norm and spatial smoothness prior

To solve the dense NRS*f*M problem as defined in the previous section, we propose to minimize an energy of the following form, jointly with respect to the motion matrix $\boldsymbol{R}$ and the shape matrix $\boldsymbol{S}$:

$$E(\boldsymbol{R}, \boldsymbol{S}) = \lambda E_{data}(\boldsymbol{R}, \boldsymbol{S}) + E_{reg}(\boldsymbol{S}) + \tau E_{trace}(\boldsymbol{S}) \quad (4)$$

where E_{data} is a data attachment term, E_{trace} favours a low-rank shape matrix, and E_{reg} is a term for the spatial regularization of the trajectories in $\boldsymbol{S}$. The positive constants λ and τ are weights that control the balance between these terms. We now describe each of these terms in detail.

The **first term** (E_{data}) is a quadratic penalty of the image reprojection error

$$E_{data} = \frac{1}{2} \|\boldsymbol{W} - \boldsymbol{R}\boldsymbol{S}\|_{\mathcal{F}}^2 \quad (5)$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm of a matrix. This term penalizes deviations of the image measurements from the orthographic projection equation (2).

The **second term** (E_{reg}) enforces edge-preserving spatial regularization of the dense 3D trajectories that constitute the columns of $\boldsymbol{S}$. To formulate this term, let i be an index (i=1,2,3) that selects the X, Y or Z coordinate of a 3D point. S_f^i will then be the i-th row of the 3D shape $\boldsymbol{S_f}$. Since the 3D points that we reconstruct are associated with projected pixel locations on the reference image I_{ref}, each element of S_f^i is associated with a specific pixel of I_{ref}. By

[1] Each $\boldsymbol{R}_f$ must satisfy the orthonormality constraint $\boldsymbol{R}_f \boldsymbol{R}_f^T = \boldsymbol{I}_{2\times 2}$

arranging these elements in the image grid of I_{ref}, we consider S_f^i as a discrete 2D image of the same size as I_{ref}. We now denote the 2D gradient of this image at pixel p by $\nabla S_f^i(p)$. Following [9], we define this discrete gradient using forward differences in both horizontal and vertical directions. We now define E_{reg} as the summation of discretized Total Variation regularizers $TV\{\cdot\}$ [25]:

$$E_{reg} = \sum_{f=1}^{F}\sum_{i=1}^{3} TV\{S_f^i\} = \sum_{f=1}^{F}\sum_{i=1}^{3}\sum_{p=1}^{N} \|\nabla S_f^i(p)\| \quad (6)$$

Total Variation based regularization smooths while preserving discontinuities and has been successfully applied to various related optic flow estimation [15, 34], and 3D reconstruction methods [19].

The **third term** (E_{trace}) penalises the number of independent shapes needed to represent the deformable scene. This is based on the realistic assumption that the shapes that a deforming object undergoes over time lie on a low-dimensional linear subspace [7]. Most NRS*f*M methods [7, 20, 32] assume that the dimension of the shape subspace is known beforehand. However, instead of using some a priori dimension for this subspace to enforce a hard rank constraint, similarly to [12], we penalize the rank of the $F \times 3N$ matrix $P(\boldsymbol{S})$. This is implemented using the *trace norm* $\|\cdot\|_*$ (a.k.a. *nuclear norm*), which is the tightest convex relaxation of the rank of a matrix and is given by the sum of its singular values Λ_j:

$$E_{trace} = \|P(\boldsymbol{S})\|_* = \sum_{j=1}^{min(F,3N)} \Lambda_j \quad (7)$$

5. Optimisation of the proposed energy

In this section, we solve the minimization of the proposed energy (4), that can be written as follows:

$$\min_{\boldsymbol{S},\boldsymbol{R}} \frac{\lambda}{2}\|\boldsymbol{W} - \boldsymbol{R}\boldsymbol{S}\|_{\mathcal{F}}^2 + \sum_{f,i,p} \|\nabla S_f^i(p)\| + \tau\|P(\boldsymbol{S})\|_* \quad (8)$$

Note that this energy is biconvex (not convex), due to the bilinear term in E_{data}. To minimize it we alternate between the estimation of the motion matrix $\boldsymbol{R}$ and the shape matrix $\boldsymbol{S}$ leaving the other fixed as described in Algorithm 1. The different components of this algorithm are presented in the rest of this section.

5.1. Motion matrix estimation

The first alternation step of Algorithm 1 involves the refinement of the motion matrix $\boldsymbol{R}$ by minimising (8) w.r.t. $\boldsymbol{R}$, assuming that the shape matrix $\boldsymbol{S}$ is known. We parameterize the camera matrices $\boldsymbol{R}_f$ using quaternions, which guarantee orthonormality. Only the term E_{data} depends on $\boldsymbol{R}$ and we minimise it using Levenberg-Marquardt.

Algorithm 1: Variational non-rigid reconstruction

Initialise $\boldsymbol{R}$, $\boldsymbol{S}$;
for *alternation* = $1, \ldots, k$ **do**
  Fix $\boldsymbol{S}$ and minimise (8) w.r.t. $\boldsymbol{R}$ using Levenberg-Marquardt algorithm;
  Fix $\boldsymbol{R}$ and minimise (8) w.r.t. $\boldsymbol{S}$ by alternating between Algorithms 2 and 3 until convergence;

5.2. Shape estimation

The second alternation step of Algorithm 1 assumes that the motion matrix $\boldsymbol{R}$ is known and minimises (8) w.r.t. to the shape matrix $\boldsymbol{S}$. Although the energy (8) is convex w.r.t. $\boldsymbol{S}$, it is non-trivial to minimise it using standard gradient descent methods. To facilitate such minimisation we use proximal splitting techniques [11] to decouple the trace norm and TV regularisation parts of the energy. We introduce an auxiliary variable $\bar{\boldsymbol{S}}$ and minimize (8) by alternating between the following two minimizations:

$$\min_{\boldsymbol{S}} \frac{1}{2\theta}\|\boldsymbol{S} - \bar{\boldsymbol{S}}\|_{\mathcal{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{W} - \boldsymbol{R}\boldsymbol{S}\|_{\mathcal{F}}^2 + \sum_{f,i,p} \|\nabla S_f^i(p)\| \quad (9)$$

$$\min_{\bar{\boldsymbol{S}}} \frac{1}{2\theta}\|\boldsymbol{S} - \bar{\boldsymbol{S}}\|_{\mathcal{F}}^2 + \tau\|P(\bar{\boldsymbol{S}})\|_* \quad (10)$$

where θ is a quadratic relaxation parameter that is relatively small so that the optimal $\boldsymbol{S}$ and $\bar{\boldsymbol{S}}$ are close. We can now efficiently solve the sub-problems (9) and (10) using convex optimization techniques.

Solving problem (9). The energy in (9) is convex but due to the TV regularisation term it is non-differentiable. However using the Legendre-Fenchel transform [17], one can dualise the regularization term in (9) and rewrite the corresponding minimisation in its primal-dual form as:

$$\min_{\boldsymbol{S}} \max_{\boldsymbol{q}} \Big\{ \frac{1}{2\theta}\|\boldsymbol{S} - \bar{\boldsymbol{S}}\|_{\mathcal{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{W} - \boldsymbol{R}\boldsymbol{S}\|_{\mathcal{F}}^2 + \sum_{f,i,p} \big\{ S_{fi}(p)\, \nabla^* \boldsymbol{q}_f^i(p) - \delta(\boldsymbol{q}_f^i(p)) \big\} \Big\} \quad (11)$$

where, $\boldsymbol{q}$ is the dual variable that contains the 2-vectors $\boldsymbol{q}_f^i(p)$, for each frame f, coordinate i and pixel p. Also, ∇^* is the adjoint of the discrete gradient operator ∇ and can be expressed as $\nabla^* = -\mathrm{div}(\cdot)$, where div is the divergence operator, after discretisation using backward differences [9]. Finally, $\delta(\cdot)$ is the indicator function of the unit ball:

$$\delta(\boldsymbol{s}) = \begin{cases} 0 & if \ \ \|\boldsymbol{s}\| \leq 1 \\ \infty & if \ \ \|\boldsymbol{s}\| > 1 \end{cases} \quad (12)$$

Following duality principles [9, 17], we solve the saddle point problem (11) by deriving a primal-dual algorithm, described in Algorithm 2. This algorithm allows a high degree of parallelization and can be solved efficiently on a GPU.

Algorithm 2: Primal dual algorithm for problem (9)

Input: Measurement matrix $\boldsymbol{W}$, current motion matrix estimates $\boldsymbol{R}$ and low rank shapes $\bar{\boldsymbol{S}}$.
Output: Spatially smooth shapes $\boldsymbol{S}$.
Parameters: λ, θ and step size σ of dual update.
Initialise the dual variable $\boldsymbol{q}$ using the estimates from the previous run of this algorithm (If this is the first run, initialize $\boldsymbol{q}$ with $\mathbf{0}$).

while *not converge* **do**

$$\underbrace{\boldsymbol{D_q}}_{3F\times N} = \begin{bmatrix} \nabla^* \boldsymbol{q}_1^1(1) & \dots & \nabla^* \boldsymbol{q}_1^1(N) \\ \vdots & \ddots & \vdots \\ \nabla^* \boldsymbol{q}_F^3(1) & \dots & \nabla^* \boldsymbol{q}_F^3(N) \end{bmatrix};$$

$$\boldsymbol{S} = \left(\lambda \boldsymbol{R}^T\boldsymbol{R} + \tfrac{1}{\theta}\boldsymbol{I}_{3F\times 3F}\right)^{-1}\left(\lambda \boldsymbol{R}^T\boldsymbol{W} + \tfrac{\bar{\boldsymbol{S}}}{\theta} - \boldsymbol{D_q}\right);$$

for $f = 1$ **to** F, $i = 1$ **to** 3, $p = 1$ **to** N **do**

$$\boldsymbol{q}_f^i(p) = \frac{\boldsymbol{q}_f^i(p) + \sigma \nabla S_f^i(p)}{\max(1, \|\boldsymbol{q}_f^i(p) + \sigma \nabla S_f^i(p)\|)};$$

Solving problem (10). Notice that the quadratic term in (10) can also be written as $\|P(\boldsymbol{S}) - P(\bar{\boldsymbol{S}})\|_{\mathcal{F}}^2$. Thus this is a convex minimisation problem that can be solved using the *soft impute* algorithm proposed in [18]. The steps that we follow are summarized in Algorithm 3. The solution $\bar{\boldsymbol{S}}$ is actually a low rank approximation of the spatially smooth shape matrix $\boldsymbol{S}$.

Algorithm 3: Soft impute algorithm for problem (10)

Input: Current estimate of spatially smooth shapes $\boldsymbol{S}$.
Output: Low rank approximation $\bar{\boldsymbol{S}}$ of the shape matrix.
Parameters: τ, θ.

$[\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V}]$ = Singular Value Decomposition of $P(\boldsymbol{S})$;
$\bar{\boldsymbol{D}} = \min(\boldsymbol{D} - \theta\tau \boldsymbol{I}_{3F\times 3F}, \mathbf{0})$;
// (where $\min(\cdot,\cdot)$ is an element-wise operator)
$\bar{\boldsymbol{S}} = P^{-1}(\boldsymbol{U}\bar{\boldsymbol{D}}\boldsymbol{V}^T)$;

5.3. Initialization of R and S

We adopt the following procedure to initialize $\boldsymbol{R}$ and $\boldsymbol{S}$. Assuming a dominant rigid component is present in the scene, we use the rigid factorization algorithm of [31] to estimate the initial camera matrices $\{\boldsymbol{R}_1 \cdots \boldsymbol{R}_F\}$ and a mean shape. This mean shape is used as a rigid initialization of the shape matrix $\boldsymbol{S}$: the mean 3D shape is replicated for every $\boldsymbol{S}_f$ in every frame f.

6. Experimental evaluation

As a preprocessing step, we normalize the measurement matrix $\boldsymbol{W}$ so that all its entries are within $[-1, 1]$. In addition, the different terms of the proposed energy (4) are normalized by applying the factor $\frac{1}{FN}$ to E_{data} and E_{reg} and $\frac{1}{\sqrt{FN}}$ to E_{trace}. In practice, we combine their effect by applying a normalized weight $\hat{\tau}$ to the trace norm term, defined as $\tau = \hat{\tau}\sqrt{FN}$. Next, experiments on synthetic and real sequences are described [2].

6.1. Synthetic face sequences

In this section we evaluate the performance of our method quantitatively on sequences generated using dense ground truth 3D data of a deforming face. We use 10 meshes of dense 3D data of different facial expressions captured using structured light [33]. We generate four different sequences that differ in the number of frames and the range and smoothness of the camera rotations and deformations, see Figure 2(a). By projecting the 3D data onto an image using an orthographic camera, we derived dense 2D tracks, which we feeded as input to the NRS*f*M estimations.

We compare the results of our algorithm against two state of the art NRS*f*M methods, Metric Projections (MP) [20] and Trajectory Basis (TB) [2], since publicly available code exists for both and the algorithms were scalable. For MP and TB we report the result for the number of basis shapes (≥ 2) that gave the lowest 3D error.

Table 6.1 shows the results for each method. We define the normalised per frame RMS error for the reconstructed 3D shape $\boldsymbol{S}_f$ with respect to the corresponding ground truth shape $\boldsymbol{S}_f^{GT}$ as: $e_{3D} = \frac{||\boldsymbol{S}_f - \boldsymbol{S}_f^{GT}||_{\mathcal{F}}}{||\boldsymbol{S}_f^{GT}||_{\mathcal{F}}}$. We report the mean RMS error over all frames, after per-frame rigid alignment with the corresponding ground truth shape using Procrustes analysis. We also provide the rank of the reconstructed result for each case[3]. Next we give details about each synthetic sequence and discuss the corresponding results.

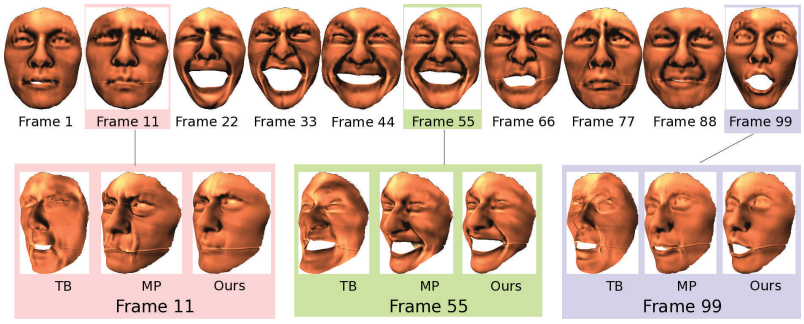
Sequence 1. In this 10 frame long sequence, each frame corresponds to a different facial expression. The face is rotated about the vertical axis from $+30°$ to $-30°$ with respect to the frontal view. This is a challenging setup since the rank of the 3D shape is very high (close to 10). The results shown on the first row of Table 6.1 reveal that both MP and TB fail to reconstruct this sequence while our approach performs well.

Sequence 2. This setup is equivalent to the previous one except that the rotations now ranged from $+90°$ to $-90°$. This simple fact allows MP and TB to reduce their errors substantially, which indicates that large rotations help in NRS*f*M. The main drawback however, is that establishing 2D correspondences under this amount of rotation would be unrealistic in a real world scenario due to occlusions.

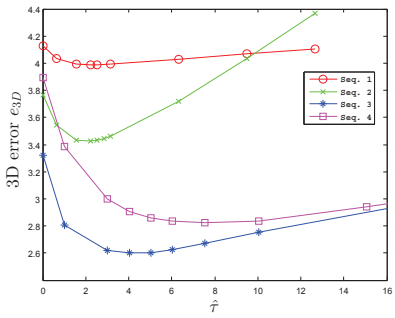
Sequence 3. 99 frame long sequence generated by linearly interpolating between pairs of views to obtain smooth 3D deformations. Realistic rotations that contain some high frequencies are simulated by upscaling (by a factor of 4)

[2] For more details and videos, visit: http://www.eecs.qmul.ac.uk/~rgarg/Variational_NRSfM

[3] Since the trace norm approximation of the rank results in some singular values being small but not exactly 0, for our method (case $\tau \neq 0$) we report the rank of the optimal 3D reconstructions retaining the singular values of $P(\boldsymbol{S})$ that explain 99% of the shape variance.



(a) Ground truth 3D shapes (top row) and dense 3D reconstructions for selected frames (bottom row) in Sequence 4 using TB [2], MP [20] and our approach.See supplementary material for videos.



(b) Normalized RMS 3D error with varying trace norm strength for synthetic experiments.

Figure 2. Results on synthetic sequences.

	TB [2]	MP [20]	Ours	Ours ($\tau = 0$)
Seq. 1	18.38%(2)	19.44%(3)	**4.01**%(9)	4.13%(10)
Seq. 2	7.47%(2)	4.87%(3)	**3.45**%(9)	3.76%(10)
Seq. 3	4.50%(4)	5.13%(6)	**2.60**%(9)	3.32%(99)
Seq. 4	6.61%(4)	5.81%(4)	**2.81**%(9)	3.89%(99)

Table 1. Quantitative evaluation on 4 synthetic face sequences. We show average RMS 3D reconstruction errors for TB [2], MP [20] and our approach. In all cases, the rank of the reconstructed result is shown in brackets.

those estimated by our algorithm on the real face sequence (Figure 3). Since this sequence assumes smooth deformations and high frequency rotations, the conditions are ideal for TB [2] which outperforms MP. However, our method outperforms both baseline methods.

Sequence 4. This setup is equivalent to the previous one with the exception that the rotations are projected onto a low frequency subspace to simulate the case of both smooth rotations and deformations. This scenario shows the failure of TB [2] to cope with rotation and deformation spaces that share frequencies [21]. Once more, our approach achieves the lowest 3D reconstruction errors.

For our method, we provide an additional column showing the results obtained in the case when the trace norm term was switched off, with $\tau = 0$. As expected, the rank of the reconstructions was much higher as were the 3D errors. Figure 2(b) shows the effect on 3D errors of varying the normalized trace norm weight parameter $\hat{\tau}$. We observe that the optimal reconstruction in all the synthetic sequences is achieved with a similar value for $\hat{\tau}$.

In conclusion, these experiments reveal some of the strengths of our algorithm: *(i)* our approach can reconstruct even in the case of small out-of-plane rotations where other methods break down, *(ii)* it can cope both with smooth or high frequency rotations and deformations.

6.2. Experiments on real sequences

In this section we present a qualitative evaluation of our variational approach on three monocular video sequences captured in natural environments under changes in lighting, occlusions and large displacements.

Face sequence. Human faces undergoing different facial expressions have been reconstructed in the past by NRS*f*M methods; however, generally only of a few, often manually tracked, feature points (fewer than 100). This 120 frame long sequence of a subject performing natural expressions was acquired under natural lighting conditions and displays occlusions due to out-of-plane rotations.

To overcome the challenges in establishing dense 2D correspondences due to the lack of texture on the skin, in this sequence we used the gradient of all color image channels (concatenated in a sequence of 6D vector-valued images) as input to the multi-frame optical flow algorithm [15]. Figure 3 shows some of the frames of the sequence, and our fully dense 3D reconstructions rotated using the recovered rotation matrices.

Back Sequence [26]. This is a 150 frame long sequence of the back of a person deforming sideways and stretching. The textured pattern worn by the subject was used to facilitate sparse feature matching in [26] but is not necessary here. Figure 4 shows images and resulting 3D dense shapes $\boldsymbol{S_f}$, rotated according to the estimated matrices $\boldsymbol{R}_f$.

Heart Sequence. In-vivo reconstruction from laparoscopic sequences is an area where NRS*f*M can be extremely useful as stereo capture inside the body is often impossible or can only be done with a very small baseline [29]. We chose a challenging monocular sequence of a beating heart taken during bypass surgery[4]. Figure 5 shows some frames and the recovered dense shapes. Not only is our approach robust to the moving specularities on the video but it can recover the rhythmic deformations of the heart well, despite the very small rotational motion component.

7. Conclusion

This paper presents the first variational approach for dense 3D reconstruction of non-rigid scenes from a monocular sequence without prior scene knowledge. We have used

[4]Video available from http://hamlyn.doc.ic.ac.uk/vision

Figure 3. 3D reconstruction results of the real face sequence. (a) Input images. (b) Corresponding 3D shapes from original viewpoint of the camera while the face rotates and deforms. (c-d) Shape deformation as observed from three-quarter and profile views respectively (after taking out the rotational component of the face). (e) Rendered surfaces from a different viewpoint, using computed deformations and rotations with augmented texture placed on the reference image. See supplementary material for videos.

the trace norm prior for low rank shapes along with TV regularization to formulate the dense NRS*f*M problem as a global energy minimisation scheme. Experimental results on challenging real sequences show that our approach can successfully generate dense 3D reconstructions even in the presence of small rotations and low image texture. A future extension of this work will be to incorporate photometric image matching and 3D reconstruction into a single optimization framework.

References

[1] I. Akhter, Y. Sheikh, and S. Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *CVPR*, 2009. 1, 2

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *PAMI*, 2011. 2, 5, 6

[3] R. Angst, C. Zach, and M. Pollefeys. The generalized trace-norm and its application to structure-from-motion problems. In *ICCV*, 2011. 2

[4] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *CVPR*, 2008. 1, 2

[5] A. Bartoli, Y. Gerard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *CVPR*, 2012. 2

[6] M. Brand. A direct method for 3D factorization of nonrigid motion observed in 2D. In *CVPR*, 2005. 2

[7] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000. 1, 2, 4

[8] E. J. Candès. The power of convex relaxation: The surprising stories of matrix completion and compressed sensing. In *SODA*, 2010. 2

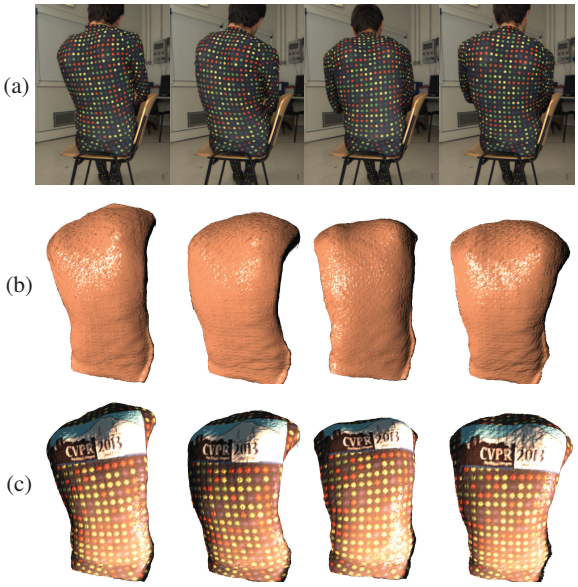


Figure 4. 3D reconstruction results for the back sequence. (a) Input images. (b) 3D reconstruction of the deformed surface. (c) Textured rendering of the result (with an additional light source).

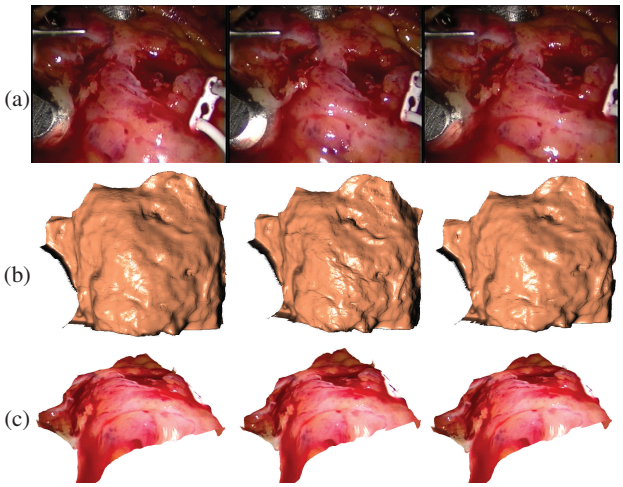


Figure 5. 3D reconstruction results for the heart sequence. (a) Input images. (b) Front view of the estimation of the deforming and rotating surface. (c) Textured rendering of the reconstruction from a side view, using in all frames the texture of the reference image.

[9] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *JMIV*, 2011. 4

[10] T. Collins and A. Bartoli. Locally affine and planar deformable surface reconstruction from video. *VMV*, 2010. 1, 2

[11] P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, 2011. 4

[12] Y. Dai, H. Li, and M. He. A simple prior-free method for non rigid structure from motion factorization. In *CVPR*, 2012. 1, 2, 4

[13] A. Del Bue. A factorization approach to structure from motion with shape priors. In *CVPR*, 2008. 2

[14] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2010. 1

[15] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *IJCV*, 2013. 1, 2, 3, 4, 6

[16] P. Gotardo and A. Martinez. Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *PAMI*, 2011. 2

[17] A. Handa, R. Newcombe, A. Angeli, and A. Davison. Applications of Legendre-Fenchel transformation to computer vision problems. Technical Report DTR11-7, Imperial College, 2011. 4

[18] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *JMLR*, 2010. 5

[19] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense tracking and mapping in real-time. In *ICCV*, 2011. 1, 2, 4

[20] M. Paladini, A. Del Bue, J. Xavier, L. Agapito, M. Stosic, and M. Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *IJCV*, 2012. 1, 2, 4, 5, 6

[21] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3D reconstruction of a moving point from a series of 2D projections. In *ECCV*, 2010. 2, 6

[22] D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. In *3DPVT*, 2010. 2

[23] S. Ricco and C. Tomasi. Dense lagrangian motion estimation with occlusions. In *CVPR*, 2012. 2, 3

[24] A. Roussos, C. Russell, R. Garg, and L. Agapito. Dense multibody motion estimation and reconstruction from a handheld camera. In *ISMAR*, 2012. 1, 2

[25] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 1992. 4

[26] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *CVPR*, 2011. 1, 6

[27] C. Russell, J. Fayad, and L. Agapito. Dense non-rigid structure from motion. In *3DIMPVT*, 2012. 2

[28] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006. 1

[29] D. Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *MICCAI*, 2012. 6

[30] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *CVPR*, 2010. 1

[31] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 1992. 2, 5

[32] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 2008. 1, 2, 4

[33] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. In *SIGGRAPH*, 2005. 5

[34] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *DAGM*, 2007. 4