

# Sequential Non-Rigid Structure-from-Motion with the 3D-Implicit Low-Rank Shape Model<sup>\*</sup>

Marco Paladini<sup>1</sup>, Adrien Bartoli<sup>2</sup>, and Lourdes Agapito<sup>1</sup>

<sup>1</sup> Queen Mary University of London, Mile End Road, E1 4NS London, UK

<sup>2</sup> Clermont Université, France

**Abstract.** So far the Non-Rigid Structure-from-Motion problem has been tackled using a batch approach. All the frames are processed at once after the video acquisition takes place. In this paper we propose an incremental approach to the estimation of deformable models. Image frames are processed online in a sequential fashion. The shape is initialised to a rigid model from the first few frames. Subsequently, the problem is formulated as a model based camera tracking problem, where the pose of the camera and the mixing coefficients are updated every frame. New modes are added incrementally when the current model cannot model the current frame well enough. We define a criterion based on image reprojection error to decide whether or not the model must be updated after the arrival of a new frame. The new mode is estimated performing bundle adjustment on a window of frames. To represent the shape, we depart from the traditional explicit low-rank shape model and propose a variant that we call the 3D-implicit low-rank shape model. This alternative model results in a simpler formulation of the motion matrix and provides the ability to represent degenerate deformation modes. We illustrate our approach with experiments on motion capture sequences with ground truth 3D data and with real video sequences.

## 1 Introduction

The reconstruction of 3D scenes from monocular video sequences is one of the fundamental problems in computer vision. Following the success on rigid structure recovery in recent years there has been a wealth of research on modelling deformable structures. Most Non-Rigid Structure-from-Motion (NR SfM) algorithms to date rely on the foundational model proposed by Bregler *et al.* [4] which describes the time varying structure of a deforming object as a linear combination of basis shapes. The pose, the basis and the time varying coefficients are then estimated using a batch approach – all the frames in the sequence are processed at once after the acquisition.

While batch and real-time sequential rigid SfM are mature fields that have now consolidated into commercial applications, NR SfM is still at its infancy.

---

<sup>\*</sup> This work was partially funded by the European Research Council under ERC Starting Grant agreement 204871-HUMANIS and the British Council/Alliance Research Programme. A. Bartoli was funded by ANR through the HFIBMR Project.

Some batch algorithms exist [3, 13, 10] but there is still a need to define deformable shape models and estimation algorithms that will allow to push NR SfM forward to a scenario where it might emulate the successes of its rigid counterpart, in terms of robust performance and application to real world cases. In this paper we advance the state of the art in NR SfM in two main directions, both proposing a new sequential estimation paradigm and an alternative low-rank shape model.

Our first contribution is the definition of a new estimation paradigm that extends NR SfM to the sequential domain. We propose a rank-growing engine which will determine when the rank of the model should be increased and if necessary will estimate the new mode.

We divide the sequential non-rigid shape estimation into two processes: model-based tracking of the camera pose and shape coefficients and model update. The first process assumes that a current up-to-date model, of a certain rank, of the 3D shape observed so far exists and performs *model based camera tracking*: when a new frame arrives this module estimates the current camera pose and the shape parameters using as input the 2D coordinates of image features matched in the last  $W$  frames, where  $W$  is the width of a sliding window. The second process is a *model update* module which decides, based on the image reprojection error given by the camera tracking module, whether or not the current model is able to explain the deformations viewed in the new frame. If the current model does not have enough descriptive power to capture the deformations observed in the new frame, the model update module will add a new mode and estimate its parameters using bundle adjustment on a sliding window. The entire system is bootstrapped from a rigid reconstruction obtained from a small number of initial frames.

Our second contribution is an alternative low-rank shape model that provides the ability to represent modes of deformation of dimensionality lower than 3 (for instance deformations on a plane or along a line).

We call it the *3D implicit low-rank shape model* since it does not use an explicitly defined 3D shape basis. This has two main advantages. First, the motion matrix in our model has a simpler structure than in the classical model, which allows for a linear estimation of camera pose and shape coefficients from a single frame, and can be used to initialise the bundle adjustment in the sequential framework. Second, our model handles deformations whose rank is not a multiple of 3 and thus avoids one to explicitly compute the rank of a particular shape basis. When the deformations are processed one frame at a time, having the flexibility to update the model with 1-dimensional modes fits the sequential estimation paradigm more naturally, since there is a much higher chance of observing lower dimensional deformations.

It is important to note that in this paper we do not try to solve the matching problem. Instead, we rely on point correspondences between frames being available. The integration of the feature tracking problem with the camera tracking and model update processes (which are the focus of this paper) is beyond the scope of this work although we certainly intend to address it in our future work.

## 2 Related Work

The ability to reconstruct a deformable 3D surface from a monocular sequence when the only input information is a set of point correspondences between images is an ill posed problem unless more constraints than just the reprojection error are used. The seminal work of Bregler *et al.* [4] was the first to propose a solution to the NR SfM problem for the orthographic camera case. This model not only provided an elegant extension of the rigid factorisation framework [12] but has also opened up new computational and theoretical challenges in the field.

Current solutions to NR SfM focus on the definition of optimization criteria to guarantee the convergence to a well behaved solution. This is often only achieved through the addition of temporal and spatial smoothness priors. Bundle adjustment has become a popular optimization tool to refine an initial rigid solution while incorporating temporal and spatial smoothness priors on the motion and the deformations.

Aanaes *et al.* [1] were the first to formulate the problem using bundle adjustment using smoothness priors. Later, Del Bue *et al.* [5] incorporated the constraint that some of the points on the object were rigid while Bartoli *et al.* [3] used a coarse to fine shape model where new deformation modes are added iteratively to capture as much of the variance left unexplained by previous modes as possible. Torresani *et al.* [13] formulate the problem using Probabilistic Principal Components Analysis introducing priors as a Gaussian distribution on the deformation weights. More recently, Paladini *et al.*'s [10] work focuses on ensuring that the solution lies on the correct motion manifold where the metric constraints are exactly satisfied. All these approaches are initialised from a rigid solution and they use temporal and spatial smoothness priors on the motion and shape parameters. Olsen *et al.* [9] proposed the surface shape prior and an implicit model that simplifies the estimation process but leads to a non-Euclidean 3D reconstruction.

The linear subspace model has also allowed closed-form solutions to be proposed for the cases of both affine [14] and perspective [16, 6] cameras. Recently, a set of new approaches has departed from the low-rank linear shape model. Rabaud and Belongie [11] adopt a manifold learning framework assuming that only small neighbourhoods of shapes are well modelled with a linear subspace.

Akhter *et al.* [2] described the structure of a non-rigid body in trajectory space as a linear combination of DCT basis trajectories with the obvious advantage that the basis is object independent.

The common attribute to all NR SfM algorithms proposed so far is that they are batch methods. Our new sequential approach is motivated by recent developments in the area of sequential real-time SfM methods for rigid scenes [7, 8]. In particular, our approach is inspired by the work of Klein and Murray [7] in which they develop a real time system based on two parallel threads – the camera tracking thread which performs real time model based pose estimation and the mapping thread which runs in a constant loop performing bundle adjustment on a small set of key-frames. To the best of our knowledge our work is the first in NR SfM to depart from the batch formulation and reformulate the shape

estimation sequentially. First we introduce a new variant to the low-rank linear basis shape model that we believe is better suited to a sequential formulation.

### 3 New Deformation Model

#### 3.1 Classical Explicit Low-Rank Shape Model

In the case of deformable objects the observed 3D points change as a function of time. In the low-rank shape model defined by Bregler *et al.* [4] the 3D points deform as a linear combination of a fixed set of  $K$  rigid shape bases according to time varying coefficients. In this way,  $\mathbf{S}_f = \sum_{k=1}^K l_{fk} \mathbf{B}_k$  where the matrix  $\mathbf{S}_f = [\mathbf{X}_{f1}, \dots, \mathbf{X}_{fP}]$  contains the 3D coordinates of the  $P$  points at frame  $f$ , the  $3 \times P$  matrices  $\mathbf{B}_k$  are the shape bases and  $l_{fk}$  are the coefficient weights. If the 3D shape is known, this model can be obtained from the PCA decomposition of the  $\mathbf{S}^*$  that contains the 3D shape in all the frames.

$$\mathbf{S}_{F \times 3P}^* = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_F^* \end{bmatrix} = \begin{bmatrix} X_{11} & Y_{11} & Z_{11} & \cdots & X_{1P} & Y_{1P} & Z_{1P} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ X_{F1} & Y_{F1} & Z_{F1} & \cdots & X_{FP} & Y_{FP} & Z_{FP} \end{bmatrix} \quad (1)$$

A PCA decomposition of rank  $K$  of  $\mathbf{S}^*$  would give  $\mathbf{L}\mathbf{B}^*$ , where  $\mathbf{L}$  is the  $F \times K$  matrix of deformation weights  $l_{ik}$ , and the  $K \times 3P$  matrix  $\mathbf{B}^*$  can be rearranged to give the basis shapes  $\mathbf{B}_k$ . If we assume an orthographic projection model the coordinates of the 2D image points observed at each frame  $i$  are then given by:

$$\mathbf{W}_i = \mathbf{R}_i \left( \sum_{k=1}^K l_{ik} \mathbf{B}_k \right) + \mathbf{T}_i \quad (2)$$

where  $\mathbf{R}_i$  is a  $2 \times 3$  *Stiefel matrix* and  $\mathbf{T}_i$  aligns the image coordinates to the image centroid. The aligning matrix  $\mathbf{T}_i$  is such that  $\mathbf{T}_i = \mathbf{t}_i \mathbf{1}_P^T$  where the 2-vector  $\mathbf{t}_i$  is the 2D image centroid and  $\mathbf{1}_P$  a vector of ones.

When the image coordinates are registered to the centroid of the object and we consider all the frames in the sequence, we may write the measurement matrix as:

$$\mathbf{W} = \begin{bmatrix} l_{11} \mathbf{R}_1 & \dots & l_{1K} \mathbf{R}_1 \\ \vdots & \ddots & \vdots \\ l_{F1} \mathbf{R}_F & \dots & l_{FK} \mathbf{R}_F \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_K \end{bmatrix} = \mathbf{M}\mathbf{S} \quad (3)$$

Since  $\mathbf{M}$  is a  $2F \times 3K$  matrix and  $\mathbf{S}$  is a  $3K \times P$  matrix in the case of deformable structure the rank of  $\mathbf{W}$  is constrained to be at most  $3K$ . The motion matrices now have a complicated repetitive structure  $\mathbf{M}_i = [\mathbf{M}_{i1} \dots \mathbf{M}_{iK}] = [l_{i1} \mathbf{R}_i \dots l_{iK} \mathbf{R}_i]$  that makes the model estimation difficult.

Olsen *et al.* [9] proposed to consider an implicit model where the repetitive structure of the motion matrix is not used. While this simplifies the estimation problem, the recovered model does not directly provide usable motion and

shape parameters, unless a mixing matrix is computed [4, 14]. The mixing matrix computation problem has not received a simple solution so far.

### 3.2 Proposed 3D-Implicit Low-Rank Shape Model

In this paper we propose to depart from the traditional basis shapes model, and embrace a different formulation that will fit the problem of sequential structure recovery more naturally since it allows for the rank of the shape model to grow one by one with the arrival of a new frame, instead of multiples of three.

The data in the shape matrix may be re-arranged in a different form, stacking the shape matrices vertically for all frames  $F$ . Each matrix  $\mathbf{S}_f \in \mathbb{R}^{3 \times P}$  contains the 3D coordinates of  $P$  points in frame  $f$ .

$$\mathbf{S}_{3F \times P} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_F \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1P} \\ Y_{11} & Y_{12} & \cdots & Y_{1P} \\ Z_{11} & Z_{12} & \cdots & Z_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ X_{F1} & X_{F2} & \cdots & X_{FP} \\ Y_{F1} & Y_{F2} & \cdots & Y_{FP} \\ Z_{F1} & Z_{F2} & \cdots & Z_{FP} \end{bmatrix} \quad (4)$$

If we assume that the shape matrix  $\mathbf{S}$  is low-rank we can perform Principal Components Analysis to obtain a PCA basis as  $\mathbf{S} = \mathbf{U}_d \mathbf{V}_d$ , where  $d$  is the rank of the decomposition,  $\mathbf{U}_d \in \mathbb{R}^{3F \times d}$  and  $\mathbf{V}_d \in \mathbb{R}^{d \times P}$ . We can also explicitly include an average rigid (mean) shape in the model, therefore the shape at frame  $f$  would be given by:

$$\mathbf{S}_f = \bar{\mathbf{S}} + [\mathbf{U}_{f1} \cdots \mathbf{U}_{fr}] \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_r \end{bmatrix} \quad (5)$$

where  $\bar{\mathbf{S}}$  is the mean shape,  $d = 3+r$ ,  $\mathbf{U}_{fr}$  is the 3-vector  $[U(x)_{fr} U(y)_{fr} U(z)_{fr}]^T$  and  $\mathbf{V}_r$  are the rows of matrix  $\mathbf{V}$ .

Therefore we can consider  $\mathbf{V}$  to be a PCA basis of the shape (row) space of  $\mathbf{S}$  and  $\mathbf{U}$  to contain the time varying coefficients. Note that in this case the shape matrix  $\mathbf{V}$  has dimensions  $r \times P$  where  $r$  is the rank of the decomposition and  $P$  is the number of points in the shape. For each frame  $3r$  coefficients are needed to express the configuration of the shape.

We assume that the shape at instant  $f$  is then projected onto an image following an orthographic camera model. The 2D coordinates of the points can then be expressed as:

$$\mathbf{W}_f = \begin{bmatrix} u_{f1} \cdots u_{fP} \\ v_{f1} \cdots v_{fP} \end{bmatrix} = \mathbf{R}_f \mathbf{S}_f + \mathbf{T}_f = \mathbf{R}_f (\bar{\mathbf{S}} + \mathbf{U}_f \mathbf{V}) + \mathbf{T}_f \quad (6)$$

where  $\mathbf{R}_f$  is a  $[2 \times 3]$  orthographic camera projection matrix, it encodes the first two rows of the camera rotation matrix and  $\mathbf{T}_f$  the translation for frame  $f$ . If we

now register all the measurements to their centroid in each frame the projection of the shape in all frames can be written as:

$$\mathbf{W} = \begin{bmatrix} \mathbf{R}_1 & & & \\ & \mathbf{R}_2 & & \\ & & \ddots & \\ & & & \mathbf{R}_F \end{bmatrix} \left( \begin{bmatrix} \bar{\mathbf{S}} \\ \bar{\mathbf{S}} \\ \vdots \\ \bar{\mathbf{S}} \end{bmatrix} + \begin{bmatrix} \mathbf{U}_{11} & \cdots & \mathbf{U}_{1r} \\ \mathbf{U}_{21} & \cdots & \mathbf{U}_{2r} \\ \vdots & & \vdots \\ \mathbf{U}_{F1} & \cdots & \mathbf{U}_{Fr} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_r \end{bmatrix} \right) \quad (7)$$

In our model, the basis shapes are not explicitly used as in the classical model, while the camera projection is explicitly modeled. We thus call our model the *3D-implicit low-rank shape model*. Our model combines Bregler *et al.* [4]’s explicit model and Olsen *et al.* [9]’s implicit model. It has the following two main advantages:

1. **Simplicity.** The motion matrix is block diagonal and only contains the rotation matrices instead of a mixture of the coefficients and the rotations. The fact that the 3D basis is not explicitly available in our model is not a problem since one is generally more interested in recovering the 3D shape of the observed scene than the basis shapes – the basis shapes can be estimated a posteriori by forming and factorizing the matrix  $\mathbf{S}^*$  in equation (1). As we explain below, it also is an advantage not to have explicit 3D basis shapes.
2. **Any-rank deformations.** Our formulation allows us to define shape models where the rank is not a multiple of 3. In other words, in the explicit model, a basis shape always has to be of rank 3, whereas in the real world not all deformations are of rank 3. Xiao and Kanade [15] propose to explicitly find the rank of a particular deformation mode (which can be one of 1, 2 or 3). Our model circumvents this difficult problem.

## 4 A Sequential Approach to NR SfM

In this paper we depart from the batch formulation of NR SfM and we propose a sequential approach based on the alternative low-rank shape model outlined in the previous section. Our approach can be seen as a two process formulation. The system holds a current up-to-date model, of a certain rank, encapsulated in matrix  $\mathbf{V}$ . The first process is a model based camera tracking module. Given the current estimate of  $\mathbf{V}$ , when a new frame arrives, the camera tracking module estimates the new pose  $\mathbf{R}_f$  and the new deformation coefficients  $\mathbf{U}_f$  for the current frame. If the current model explains well the measurements the image reprojection error will be low. However, if the error goes above some defined threshold the rank of the model must be increased and the model updated. In that case, a model update module will update the current model adding a new row to matrix  $\mathbf{V}$ . As the sequence is processed the model will become more complicated, until all the possible object deformations have been observed. Our sequential approach to NR SfM is summarised in Algorithm 1. We now describe in detail the two main modules of our sequential system: the camera tracking module and the model update module.

---

**Algorithm 1** Sequential Non-Rigid Structure-from-Motion (NR SfM)
 

---

**Input:** 2D point correspondences

**Output:** 3D coordinates of the deforming surface for each frame.

- 1: Initialise model to mean rigid shape  $\bar{\mathbf{S}}$  estimated via rigid factorization on the first few frames.
  - 2: **loop**
  - 3:   new frame  $f$  arrives
  - 4:   run *camera tracking process*: estimate camera pose  $\mathbf{R}_i$  and coefficients  $\mathbf{U}_i$
  - 5:   **while** (image reprojection error is above threshold) **do**
  - 6:     run *model update process*:
  - 7:       increase rank  $r \leftarrow r + 1$
  - 8:       estimate new row of  $\mathbf{V}$  and new column of  $\mathbf{U}_f$
  - 9:   **end while**
  - 10:   go to process next frame;  $f \leftarrow f + 1$
  - 11: **end loop**
- 

## 5 Camera Tracking Given a Known Model $\mathbf{V}$

If the matrix  $\mathbf{V}$  is known in advance, the NR SfM problem is reduced to the estimation of the camera pose  $\mathbf{R}_f$  and the mixing coefficients  $\mathbf{U}_f$  for each frame. In that case, the pose of the camera and the coefficients can be updated sequentially for each frame using a model based approach.

We adopt a sliding window approach where we perform bundle adjustment on the last  $W$  frames where  $W$  is the width of a pre-defined window. The cost to be minimised is the image reprojection error over all frames in the window:

$$\min_{\mathbf{R}_i, \mathbf{U}_i} \sum_{i=f-W}^f \|\mathbf{W}_i - \mathbf{R}_i(\bar{\mathbf{S}} + \mathbf{U}_i\mathbf{V})\|_F^2 \quad (8)$$

To this cost function we add a temporal smoothness prior to penalise strong variations in the camera matrices of the form  $\|\mathbf{R}_i - \mathbf{R}_{i-1}\|_F^2$ , and a shape smoothness prior (similar to the one used in [3]) that ensures that points that lie close to each other in space should stay close. The shape smoothness is defined as  $\sum_{i=f-W}^f D^{i,i-1}$ , where  $D^{i,i-1}$  is the change in the euclidean distance between 3D points over two frames:  $D^{i,i-1} = \sum_{a,b=1}^P \phi_{a,b} |d^2(\mathbf{X}_{i,a}, \mathbf{X}_{i,b}) - d^2(\mathbf{X}_{i-1,a}, \mathbf{X}_{i-1,b})|$ . The weight  $\phi_{a,b}$  is a measure of the closeness of points  $a$  and  $b$ , defined as a  $P \times P$  affinity matrix  $\phi_{a,b} = \rho(d^2(\mathbf{X}_a, \mathbf{X}_b))$  where  $\rho$  is a truncated Gaussian kernel. The final cost function can now be written as:

$$\min_{\mathbf{R}_i, \mathbf{U}_i} \sum_{i=f-W}^f \|\mathbf{W}_i - \mathbf{R}_i(\bar{\mathbf{S}} + \mathbf{U}_i\mathbf{V})\|_F^2 + \lambda \sum_{i=f-W}^f \|\mathbf{R}_i - \mathbf{R}_{i-1}\|_F^2 + \psi \sum_{i=f-W}^f D^{i,i-1} \quad (9)$$

The mean shape  $\bar{\mathbf{S}}$  and the shape model  $\mathbf{V}$  are assumed to be known. This nonlinear minimization requires an initial estimate for the camera pose  $\mathbf{R}_f$  and the shape coefficients  $\mathbf{U}_f$  in the current frame  $f$ . Algorithms to obtain linear estimates for  $\mathbf{R}_f$  and  $\mathbf{U}_f$  are described in Section 5.1.

The steps of the complete algorithm to track the current pose of the camera and the shape coefficients given the shape model can be summarised as follows. Each time a new frame  $f$  of feature tracks is available:

- Obtain initial estimates for the current pose  $\mathbf{R}_f$  and mixing coefficients  $\mathbf{U}_f$  using the linear estimation plus prior described in Section 5.1.
- Minimize the cost function (9) with smoothness priors using bundle adjustment to obtain optimized values for the rotations  $\mathbf{R}_i$  and shape coefficients  $\mathbf{U}_i$  in all the frames in the sliding window.
- If the reprojection error of the window becomes higher than a threshold, signal the modelling process to increase the rank of the  $\mathbf{V}$  matrix.

### 5.1 Initialization: Linear Estimation of $\mathbf{U}_f$ and $\mathbf{R}_f$

Consider new image measurements become available for a new frame. These can be arranged in a  $2 \times P$  matrix for that single frame called  $\mathbf{W}_f$ . The projection model gives us the relation  $\mathbf{W}_f = \mathbf{R}_f(\mathbf{S} + \mathbf{U}_f\mathbf{V}) + \mathbf{T}_f$ .

**Linear estimation of  $\mathbf{R}_f$ .** For every new frame the camera pose  $\mathbf{R}_f$  must be initialised before Bundle Adjustment. For this purpose, we approximate the shape with the rigid mode to obtain an initial estimate of the camera rotation. This means we need to find the camera pose  $\mathbf{R}_f$  that satisfies  $\mathbf{W}_f = \mathbf{R}_f\mathbf{S}$ , while respecting the smoothness prior  $\lambda\mathbf{I} \text{vec}(\mathbf{R}_f) = \lambda \text{vec}(\mathbf{R}_{f-1})$ . Using the relation  $\text{vec}(\mathbf{AXB}) = [\mathbf{B}^T \otimes \mathbf{A}] \text{vec}(\mathbf{X})$ , where  $\otimes$  is the Kronecker product and  $\text{vec}(\cdot)$  is the column-major vectorisation of a matrix, and using  $\mathbf{W}_f = \mathbf{I}_2\mathbf{R}_f\mathbf{S}$  we can write:

$$\text{vec}(\mathbf{W}_f) = [\mathbf{S}^T \otimes \mathbf{I}_2] \text{vec}(\mathbf{R}_f) \quad (10)$$

$$\begin{bmatrix} [\mathbf{S}^T \otimes \mathbf{I}_2] \\ \lambda\mathbf{I} \end{bmatrix} \text{vec}(\mathbf{R}_f) = \begin{bmatrix} \text{vec}(\mathbf{W}_f) \\ \lambda \text{vec}(\mathbf{R}_{f-1}) \end{bmatrix} \quad (11)$$

The resulting  $\mathbf{R}_f$  will not be orthonormal (i.e. not a truncated rotation matrix), so we find the closest orthonormal rigid projection using SVD.

**Linear estimation of  $\mathbf{U}_f$ .** First we take away the contribution to the image measurements given by the known translation and mean shape component to give  $\tilde{\mathbf{W}}_f = \mathbf{W}_f - \mathbf{T}_f - \mathbf{R}_f\mathbf{S} = \mathbf{R}_f\mathbf{U}_f\mathbf{V}$ , which can be rewritten as  $\text{vec}(\tilde{\mathbf{W}}_f) = [\mathbf{V}^T \otimes \mathbf{R}_f] \text{vec}(\mathbf{U}_f)$ . This provides a linear equation on the unknown vector  $\mathbf{U}_f$ . However, this is not sufficient to produce an acceptable solution, because  $\mathbf{U}_f$  is a  $3 \times r$  matrix where each column  $\mathbf{U}_{f_r}$  is a 3-vector  $[U(x)_{f_r} U(y)_{f_r} U(z)_{f_r}]^T$  that contains the PCA coefficients of all 3D coordinates, while  $\tilde{\mathbf{W}}_f$  contains 2D projections. However, this problem can be overcome by including a temporal smoothness prior term that penalises solutions that are far from the value for the previous frame  $\mathbf{U}_{f-1}$ . Thus the prior term is of the form  $\lambda\mathbf{I} \text{vec}(\mathbf{U}_f) = \lambda \text{vec}(\mathbf{U}_{f-1})$ . We can join both linear equations and solve the linear system:

$$\begin{bmatrix} [\mathbf{V}^T \otimes \mathbf{R}_f] \\ \lambda\mathbf{I} \end{bmatrix} \text{vec}(\mathbf{U}_f) = \begin{bmatrix} \text{vec}(\tilde{\mathbf{W}}_f) \\ \lambda \text{vec}(\mathbf{U}_{f-1}) \end{bmatrix} \quad (12)$$

## 6 Sequential Update of the Shape Model

In NR SfM the 3D object the camera observes varies over time. The current model will encode the modes of deformation that the object has exhibited so far in the sequence. However, if the object deforms in different ways that are not encoded in the model the camera tracking will fail. Therefore, a mechanism is needed to update the model when new modes of deformation appear. In that case, the rank of the model should grow and the parameters of the model should be fit to the new data.

The difficulty of updating the model in an sequential way is doublefold. Firstly, when each new frame arrives, we need a mechanism to decide whether or not the current model continues to fit the data well enough. While the shape model can still describe the data, we can continue to do model based camera tracking. We decide this based on the image reprojection error. Secondly, if the model can no longer explain the data, the rank of the model needs to grow to incorporate the new mode of deformation and the parameters of the new row of  $V$  and the new column of  $U$  must be estimated.

### 6.1 Rank Increase Criterion

The rank selection criterion will decide to increase the rank only if the current data does not fit the model well enough, i.e. if the existing modes do not model the current frame well. Therefore we use the image reprojection error as the criterion – if the error increases above a certain threshold we increase the rank of the shape model. This results in a new row being added to the PCA basis  $V$  and a new column to the PCA components  $U$ . However, the new mode is recovered from the current frame only, so it has no influence over past frames. Therefore for all past frames we can set the  $3(f - 1)$  components of the new column of  $U$  to 0.

### 6.2 Model Update: Estimating New Row of $V$ and New Column of $U$

When the camera tracking module processes a new frame that it cannot model well enough (the reprojection error is above the defined threshold), the model is updated by increasing the rank. Ideally once all the different modes of deformation that an object can exercise are incorporated in the PCA basis, the rank will remain stable and the camera tracking process will be able to reconstruct the incoming frames.

Given new image correspondences for frame  $f$ , the rank of  $U, V$  must be increased. From the current estimate of  $U_{f,1:r-1}$  and  $V_{1:r-1}$  we can rewrite the model for the new frame as

$$\tilde{W}_f = R_f(\tilde{S} + U_{f,1:r-1}V_{1:r-1} + U_{f,r}V_r). \quad (13)$$

Both the residual of the current model  $A = \tilde{W}_f - R_f(\tilde{S} + U_{f,1:r-1}V_{1:r-1})$  and the current camera rotation  $R_f$  are known. We need to estimate  $Z = U_{f,r}V_r$ , the

contribution of the new rank, subject to the following constraints:

$$\mathbf{A} = \mathbf{R}_f \mathbf{Z} \quad \text{rank}(\mathbf{Z}) = 1 \quad (14)$$

This problem is difficult to solve in closed form, therefore we approximate it using a linear solution as follows. We define  $\mathbf{C}$  as the closest rank-1 approximation of  $\mathbf{A}$  obtained using SVD, then compute  $\mathbf{Z}$  as  $\mathbf{Z} = \mathbf{R}_f^\dagger \mathbf{C}$ . Finally, we can decompose  $\mathbf{Z}$  using a rank-1 SVD decomposition to obtain a new row for  $\mathbf{V}$ .

**Non-linear refinement** Once initial estimates are available for the new row of  $\mathbf{V}$  and the new column of  $\mathbf{U}$ , they can be refined minimising image reprojection error over a sliding window of  $W$  frames

$$\min_{\mathbf{v}_r, \mathbf{u}_{ir}} \sum_{i=f-W}^f \|\mathbf{W}_i - \mathbf{R}_i(\bar{\mathbf{S}} + \mathbf{U}_i \mathbf{V})\|_F^2 \quad (15)$$

incorporating the smoothness priors described in section 5. Once the model is updated, the camera tracking module can resume *model based tracking* with the new model  $\mathbf{V}$  with rank  $r + 1$ .

### 6.3 Bootstrapping

One of the known challenges in sequential approaches to rigid SfM is the initialization [7]. It is common to run the system in batch mode for a few frames to obtain a first model of the scene before starting the sequential operation. In the current experiments we run a rigid factorization algorithm on a few initial frames to obtain the rigid mean shape  $\bar{\mathbf{S}}$ . Once this is available the camera tracking and model update loop can start. An alternative approach that does not require manual intervention is the following. Start performing rigid factorization in batch. When a new frame arrives, if the reprojection error of rigid factorization over the frames observed so far is below the threshold then we keep performing rigid factorization. However, if the error becomes higher than our threshold, the mean shape of the non-rigid model is set to the rigid model obtained so far and we start our sequential NR SfM algorithm.

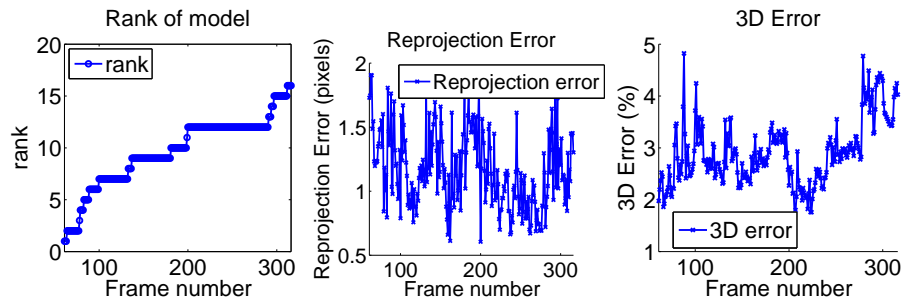
## 7 Experiments

### 7.1 Motion capture sequence *CMU-face*

First we tested our sequential method based on the 3D-implicit low-rank shape model on a motion capture sequence with ground truth data<sup>3</sup>. This sequence from the CMU Motion Capture Database<sup>4</sup> contains 316 frames of motion capture data of the face of a subject wearing 40 markers performing deformations

<sup>3</sup> Videos of the experimental results can be found on the project website <http://www.eecs.qmul.ac.uk/~lourdes/SequentialNRSFM>

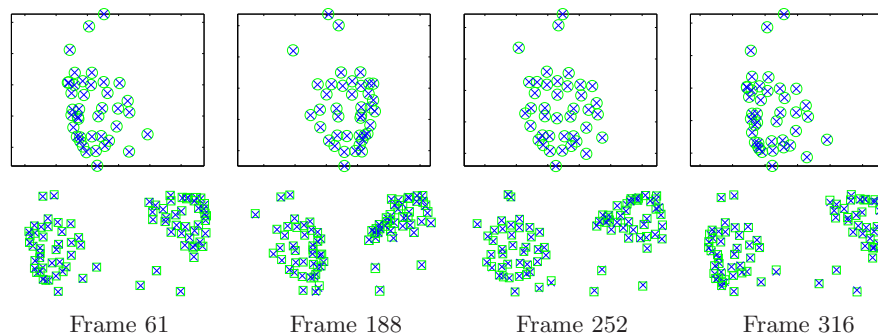
<sup>4</sup> Available from <http://mocap.cs.cmu.edu>



**Fig. 1.** Results of sequential NR SfM on the CMU-face sequence. Left: Value of the rank of the model for each frame, increasing as more frames are processed. Middle: 2D Reprojection error given by the camera tracking process. Right: 3D error of the reconstruction for each frame.

while rotating. This sequence was also used by Torresani *et al.* [13] to perform quantitative tests with ground truth data. We projected the 3D data synthetically using an orthographic camera model.

Prior to the start of our sequential algorithm and with the purpose of bootstrapping the camera tracking module, we ran a batch rigid SfM algorithm [12] on the first 60 frames of the sequence to estimate the mean shape  $\bar{\mathbf{S}}$ . The PCA basis matrix  $\mathbf{V}$  was initialised to 0. We then ran our new sequential algorithm based on the camera tracking and the model update modules, together with the rank detection engine. The average 3D error is 2.9%, with a 0.7 pixels 2D reprojection error on the  $600 \times 600$  pixels images. The reprojection threshold was fixed to 1.2 pixels.



**Fig. 2.** 3D Reconstruction results obtained on the *CMU-face* sequence using camera tracking and model updating. First row: 2D image points (green circles) and reprojections (blue crosses). Second row: Views of the 3D reconstruction (crosses) compared with ground truth MOCAP data (squares)

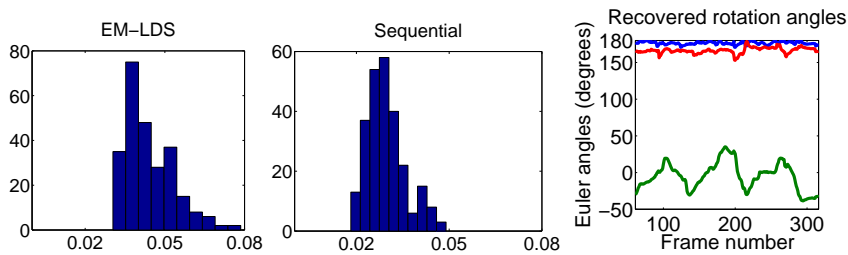
In Figure 1 we show results of the rank estimation, the 2D image reprojection error and the 3D error for each frame in the sequence using our sequential estimation formulation. The average image reprojection error over the whole sequence is less than a pixel. In Figure 3 (left) we compare results of the 3D error obtained with our method (Sequential), with Torresani *et al.*'s state of the art batch NR SfM algorithm (EM-LDS) [13]. We show the histogram of 3D error values taking into account all the frames in the sequence. The results show that our new sequential algorithm provides results comparable to Torresani *et al.*'s [13] batch state of the art algorithm. We show smooth estimates of the rotation angles for all the frames in the sequence in Figure 3 (right). In Figure 2 we show the 2D image reprojection error and the 3D reconstructions (blue crosses) we obtained for some frames in the sequence comparing them with ground truth values (green squares).

## 7.2 Real Data

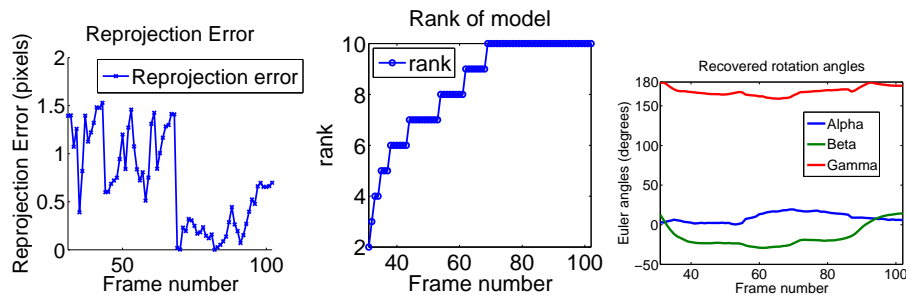
We used the *actress* sequence, also used by Bartoli *et al.* [3], which consists of 102 frames of a video showing an actress talking and moving her head. In Figure 5 we show results of the 3D reconstructions obtained for some of the frames in the sequence. The camera tracking was bootstrapped with a rigid model obtained using Tomasi and Kanade's rigid factorization algorithm [12] on the first 30 frames. The threshold for increasing the rank was a reprojection error of 0.9 pixels. From figure 4 we can see that the rank is increased, and the estimation of new frame parameters keeps the reprojection error low.

## 8 Conclusions

We have undergone a re-thinking of the NR SfM problem for monocular sequences providing a sequential solution. Our new sequential algorithm is able to automatically detect and increase the complexity of the model. Current state



**Fig. 3.** (Left) Histogram of 3D error values built from all the frames, comparing results of our method (Sequential) with Torresani *et al.*'s state of the art batch (EM-LDS) [13]. The 3D errors obtained with our Sequential approach are comparable to the results from the batch method EM-LDS. (Right) Rotation angles estimated with the camera tracking module.

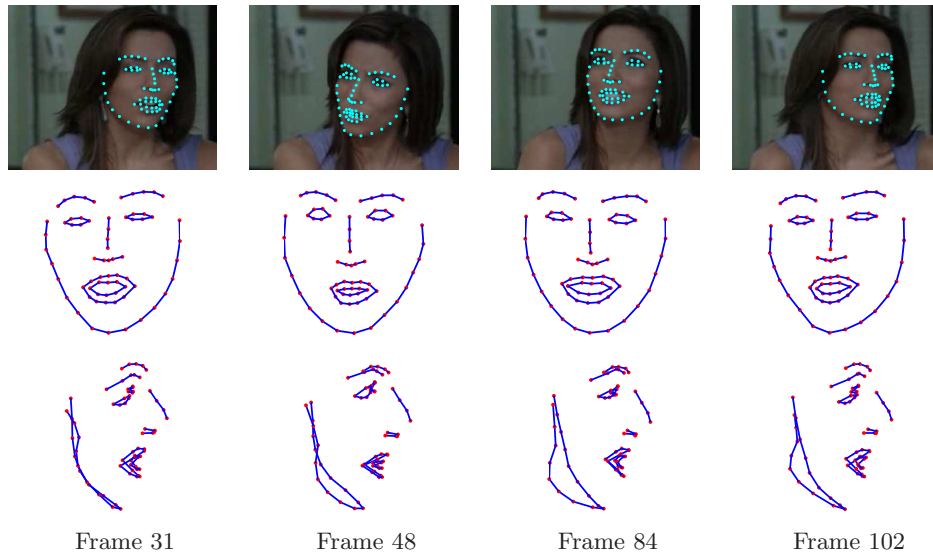


**Fig. 4.** Results on the *actress* sequence. Left: Reprojection error of the frame-by-frame reconstruction obtained with our method. Middle: The value of the rank, increased as more frames are processed. Right: Rotation angles estimated with the camera tracking module.

of the art methods for NR SfM are batch and rely on prior knowledge of the model complexity (usually the number of basis shapes,  $K$ ). Our 3D-implicit low-rank shape model simplifies the projection model and allows the rank to grow one-by-one making it well suited to frame-by-frame operation. We have shown quantitative results on a motion capture sequence and shown our system in operation on a real sequence. Future work will pursue the goal of merging the feature tracking and modelling of image data into a single process. Concerning real time capability, our current MATLAB implementation is not real time. However, the sliding window approach ensures that the computation time per frame is bounded i.e. it does not grow with the number of frames. Therefore we foresee that with appropriate code optimisation we would be able to achieve real-time performance.

## References

1. Aanæs, H., Kahl, F.: Estimation of deformable structure and motion. In: Workshop on Vision and Modelling of Dynamic Scenes, Copenhagen, Denmark (2002)
2. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid Structure from Motion in Trajectory Space. In: Neural Information Processing Systems (2008)
3. Bartoli, A., Gay-Bellile, V., Castellani, U., Peyras, J., Olsen, S., Sayd, P.: Coarse-to-Fine Low-Rank Structure-from-Motion. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska (2008)
4. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head, South Carolina (2000)
5. Del Bue, A., Lladó, X., Agapito, L.: Non-rigid metric shape and motion recovery from uncalibrated images using priors. In: IEEE Conf. on Computer Vision and Pattern Recognition, New York, NY (2006)
6. Hartley, R., Vidal, R.: Perspective nonrigid shape and motion recovery. In: 10th European Conf. on Computer Vision, Marseille, France (2008)



**Fig. 5.** Qualitative results on the *actress* sequence using camera tracking and model update. First row: The input images with superimposed feature tracking data. Second and Third rows: Front and side views of the 3D reconstruction of 4 frames of the sequence.

7. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07). Nara, Japan (November 2007)
8. Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P.: Generic and real-time structure from motion using local bundle adjustment. *Image Vision Comput.* 27(8), 1178–1193 (2009)
9. Olsen, S.I., Bartoli, A.: Implicit non-rigid structure-from-motion with priors. *J. Math. Imaging Vis.* 31(2-3), 233–244 (2008)
10. Paladini, M., Del Bue, A., Stosic, M., Dodig, M., Xavier, J., Agapito, L.: Factorization for non-rigid and articulated structure using metric projections. In: IEEE Conf. on Computer Vision and Pattern Recognition, Miami, Florida (2009)
11. Rabaud, V., Belongie, S.: Re-thinking non-rigid structure from motion. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska (2008)
12. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision* 9(2) (1992)
13. Torresani, L., Hertzmann, A., Bregler, C.: Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(5) (2008)
14. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision* 67(2) (2006)
15. Xiao, J., Kanade, T.: Non-rigid shape and motion recovery: Degenerate deformations. In: IEEE Conf. on Computer Vision and Pattern Recognition, Washington D.C. (2004)
16. Xiao, J., Kanade, T.: Uncalibrated perspective reconstruction of deformable structures. In: 10th Int. Conf. on Computer Vision, Beijing, China (2005)