Non-Rigid Metric Shape and Motion Recovery from Uncalibrated Images Using Priors

Alessio Del Bue Xavier Lladó Lourdes Agapito
Department of Computer Science
Queen Mary, University of London
Mile End Road, London, E1 4NS, UK
{alessio,llado,lourdes}@dcs.qmul.ac.uk

Abstract

In this paper we focus on the estimation of the 3D Euclidean shape and motion of a non-rigid object which is moving rigidly while deforming and is observed by a perspective camera. Our method exploits the fact that it is often a reasonable assumption that some of the points are deforming throughout the sequence while others remain rigid. First we use an automatic segmentation algorithm to identify the set of rigid points which in turn is used to estimate the internal camera calibration parameters and the overall rigid motion. Finally we formalise the problem of non-rigid shape estimation as a constrained non-linear minimization adding priors on the degree of deformability of each point. We perform experiments on synthetic and real data which show firstly that even when using a minimal set of rigid points it is possible to obtain reliable metric information and secondly that the shape priors help to disambiguate the contribution to the image motion caused by the deformation and the perspective distortion.

1. Introduction

The extensive work of the past years on *structure from motion* for the case of rigid objects has recently been extended to deal with non-rigid structure. Bregler et al. [4] introduced a representation for non-rigid 3D shape where any configuration can be expressed as a linear combination of basis shapes that define the principal modes of deformation of the object. They proposed a factorization method for weak perspective viewing conditions that exploits the rank constraint on the measurement matrix and enforces orthonormality constraints on camera rotations to recover the motion and the non-rigid 3D shape. Torresani et al. [16] extended the method to a trilinear optimization problem using Alternating Least Squares while Brand [2] proposed an alternative optimization method adding the extra constraint that the deformations should be as small as possible relative to the mean shape.

The main problem with these approaches stems from the fact that the shape and motion are ambiguous. Recently, Xiao et al. [18] proved that the orthogonality constraints were insufficient to disambiguate rigid motion and deformations. They identified a new set of constraints on the shape bases which, when used in addition to the rotation constraints, provide a closed form solution to the problem of non-rigid structure from motion. However, their solution is based on a strong assumption as it requires that there be D image frames (where D is the number of basis shapes) in which the basis shapes are known to be independent. Besides, it was later pointed out by Brand [3] that their method provides an exact solution with noiseless data and when the number of basis shapes D is known but it breaks down with noisy data or when D is not correctly estimated. Non-linear optimization schemes that minimize image reprojection error have also been proposed to refine an initial solution [1, 6]. Torresani et al. [15] proposed an algorithm that learns the time-varying shape of a non-rigid 3D object from uncalibrated 2D tracking data and where prior information on the motion or shape is introduced to avoid ambiguities.

Crucially, all the methods described previously assume the case of images acquired under weak perspective viewing conditions. In this paper we are interested in the case when the images are acquired at closer distances or with a wide field of view and perspective distortions appear. Xiao and Kanade [17] have more recently developed a two step factorization algorithm for reconstruction of 3D deformable shapes under the full perspective camera model. In this paper we present an alternative approach to non-rigid shape and motion recovery under the full perspective camera model. Our method adopts similar assumptions to the work by Del Bue et al. [5] where the observation is made that frequently a scene might contain a mixture of rigid and non-rigid points. Similarly, our approach first performs

rigid and non-rigid motion segmentation on the image data to separate both types of motion where the main contribution in this paper is to deal with the projective camera case. To obtain the metric upgrade information we perform self-calibration on the rigid set of points which provides estimates for the camera intrinsic parameters, the overall rigid motion and the mean shape. We then formalise the problem of non-rigid shape estimation as a constrained non-linear minimization using the estimates given by the self-calibration algorithm as the starting point for the minimization and providing priors on the degree of rigidity of each of the points. Finally we show results on synthetic and real data.

2. Non-rigid factorization: Background

2.1. Weak perspective camera model

The model introduced by Bregler et al. [4] to express non-rigid deformations is point-wise and the 3D shape $\mathtt{S}_i$ at frame i (in non-homogeneous coordinates) is approximated by a linear combination of a set of D basis shapes $\mathtt{B}_d$ which represent the principal modes of deformation of the object:

$$\mathtt{S}_i = \sum_{d=1}^{D} l_{id}\mathtt{B}_d \qquad \mathtt{S}_i, \mathtt{B}_d \in \Re^{3\times N} \quad l_{id} \in \Re \qquad (1)$$

where each basis shape $\mathtt{B}_d$ is a $3 \times N$ matrix which contains the 3D locations of N object points for that particular mode of deformation. Assuming an orthographic camera model the shape is then projected onto an image giving N image points:

$$\mathtt{W}_i = \begin{bmatrix} u_{i1} & \dots & u_{iN} \\ v_{i1} & \dots & v_{iN} \end{bmatrix} = \mathtt{R}_i \left(\sum_{d=1}^{D} l_{id}\mathtt{B}_d \right) \qquad (2)$$

where $[u_{ij} v_{ij}]^T$ are the horizontal and vertical image coordinates of point j – referred to the centroid of the object – and $\mathtt{R}_i$ encodes the first two rows of the rotation matrix. If all N points are tracked in F image frames we may construct the measurement matrix $\mathtt{W}$ which can be expressed as:

$$\mathtt{W} = \begin{bmatrix} \mathtt{W}_1 \\ \vdots \\ \mathtt{W}_F \end{bmatrix} = \begin{bmatrix} l_{11}\mathtt{R}_1 & \dots & l_{1D}\mathtt{R}_1 \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F & \dots & l_{FD}\mathtt{R}_F \end{bmatrix} \begin{bmatrix} \mathtt{B}_1 \\ \vdots \\ \mathtt{B}_D \end{bmatrix} = \mathtt{M}\,\mathtt{B} \qquad (3)$$

Clearly, the rank of the measurement matrix is constrained to be at most $3D$. This rank constraint can be exploited to factorize the measurement matrix into a motion matrix $\tilde{\mathtt{M}}$ and a shape matrix $\tilde{\mathtt{B}}$ by truncating the SVD of $\mathtt{W}$ to rank $3D$. However, this factorization is not unique since any invertible $3D \times 3D$ matrix $\mathtt{Q}$ can be inserted in the decomposition leading to: $\mathtt{W} = (\tilde{\mathtt{M}}\mathtt{Q})(\mathtt{Q}^{-1}\tilde{\mathtt{B}})$. The problem is to find a transformation matrix $\mathtt{Q}$ that renders the appropriate replicated block structure of the motion matrix shown in Equation (3) and that removes the affine ambiguity, upgrading the reconstruction to a metric one.

2.2. Perspective camera model

If we now assume a perspective projection model for the camera a 3D point $\mathbf{X}_j$ will be projected onto image frame i according to the equation:

$$\mathbf{x}_{ij} = \frac{1}{\lambda_{ij}}\,\mathtt{P}_i\mathbf{X}_j \qquad (4)$$

where $\mathbf{x}_{ij}$ and $\mathbf{X}_{ij}$ are both expressed in homogeneous coordinates (i.e. $\mathbf{x}_{ij} = [u_{ij} v_{ij} 1]^T$ and $\mathbf{X}_j = [X_j Y_j Z_j 1]^T$), $\mathtt{P}_i$ is the projection matrix and λ_{ij} is the projective depth for that point. Scaling the image coordinates of all the points in all the views by their corresponding projective depth gives a $3F \times N$ measurement matrix:

$$\mathtt{W} = \begin{bmatrix} \lambda_{11}\mathbf{x}_{11} & \dots & \lambda_{1N}\mathbf{x}_{1N} \\ \vdots & & \vdots \\ \lambda_{F1}\mathbf{x}_{F1} & \dots & \lambda_{FN}\mathbf{x}_{FN} \end{bmatrix} = \begin{bmatrix} \mathtt{P}_1 \\ \vdots \\ \mathtt{P}_F \end{bmatrix} \mathtt{X} \qquad (5)$$

where $\mathtt{X} = [\mathbf{X}_1 \dots \mathbf{X}_N]$ is a $4 \times N$ shape matrix which contains the homogeneous coordinates of the N 3D points. In the case of rigid structure, when the measurement matrix is rescaled with the correct projective depths, the rank of $\mathtt{W}$ is constrained to be at most 4. Various algorithms [9, 11] have been proposed to use the rank 4 constraint to estimate the projective depths first and then obtain a projective reconstruction of the scene based on the factorization of the scaled measurement matrix.

However, for non-rigid shape the 3D structure changes from frame to frame where $\mathtt{X}_i = [\mathbf{X}_{i1} \dots \mathbf{X}_{iN}]$ is the shape at frame i. As we mentioned above, the deformation of a shape can often be explained as a linear combination of a set of shape bases. In the projective case [17] the 3D vectors are expressed in homogeneous coordinates and so the shape may be written as:

$$\mathtt{X}_i = \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathtt{B}_d \\ \mathbf{1} \end{bmatrix} \qquad \mathtt{X}_i \in \Re^{4\times N} \quad \mathtt{B}_d \in \Re^{3\times N} \qquad (6)$$

where $\mathtt{B}_d$ are the $(3 \times N)$ shape bases, l_{id} are the corresponding deformation coefficients and $\mathbf{1}$ is an N-vector of ones. The projection of the shape at any frame i onto the image is then governed by the projection equation $\mathtt{W}_i = \mathtt{P}_i\mathtt{X}_i$. In matrix form this can be re-written for all frames as [17]:

$$\mathtt{W} = \begin{bmatrix} l_{11}\mathtt{P}_1^{(1:3)} & \dots & l_{1D}\mathtt{P}_1^{(1:3)} & \mathbf{P}_1^{(4)} \\ \vdots & & \vdots & \vdots \\ l_{F1}\mathtt{P}_F^{(1:3)} & \dots & l_{FD}\mathtt{P}_F^{(1:3)} & \mathbf{P}_F^{(4)} \end{bmatrix} \begin{bmatrix} \mathtt{B}_1 \\ \vdots \\ \mathtt{B}_D \\ \mathbf{1} \end{bmatrix} \qquad (7)$$

where $\mathtt{P}_i^{(1:3)}$ are the first three columns of the projection matrix, $\mathbf{P}_i^{(4)}$ is the fourth and $\mathbf{1}$ is an N-vector of ones. Clearly, the rank of the measurement matrix is at most $3D + 1$ for the projective case [17]. Once more if the projective depths λ_{ij} were known the measurement matrix could be rescaled and decomposed into motion and projective shape matrices using factorization. In their most recent work Xiao and Kanade [17] proposed a new method to estimate the projective depths using the $3D + 1$ subspace constraint and then upgrade the projective reconstruction to a metric one using an extension of their affine closed form solution to the perspective camera case. However, their method still relies on the assumption that there be D frames in which the basis shapes are known to be independent.

In this paper we propose an alternative method to obtain a metric reconstruction of a non-rigid object observed by a perspective camera using the assumption that some of the points on the object might be undergoing rigid motion. The point trajectories are segmented and then the rigid set is used to obtain the information that will allow to upgrade the structure from projective to metric space. Finally we formalise the problem of non-rigid shape estimation as a constrained non-linear minimization. In the next section we describe the automatic segmentation algorithm we propose to separate the rigid and non-rigid motions.

3. Segmentation of rigid and non-rigid motion under perspective viewing

In the case of a weak perspective camera, the rank of a measurement matrix containing a set of rigid points is constrained to be at most 3. However (see equation 5), when the camera is described by the perspective model, the rank of the measurement matrix is 4, provided that the measurement matrix has been rescaled with the correct estimates of the projective depths λ_{ij}. When the points in the measurement matrix are non-rigid the rank increases to $3D+1$ in the projective camera case where D is the number of basis shapes. Unfortunately, the rank constraint cannot be used directly to segment rigid and non-rigid points, since the rigid points could always be explained as non-rigid points with zero configuration weights for the non-rigid shape bases.

Instead, our approach is based on the fact that rigid points will satisfy the epipolar geometry while the non-rigid points will give a high residual in the estimation of the fundamental matrix between pairs of views. We use a RANSAC algorithm [7] to estimate the fundamental matrices and to segment the scene into rigid and non-rigid points (which we consider to be outliers).

However, a well known drawback of random sampling and consensus techniques is the computational cost required to obtain a valid set of points when the percentage of outliers is high due to the large number of samples needed to be drawn from the data. Unfortunately, this is the most likely scenario in non-rigid structure from motion where we normally deal with a small proportion of completely rigid points. In this paper we exploit a measure of the degree of deformability of a point to infer a prior distribution of the probability of a trajectory being rigid or non-rigid given that measure. These distributions are then used as priors to perform guided sampling over the set of trajectories in a similar approach to the one proposed by Tordoff and Murray [14] for the stereo matching problem.

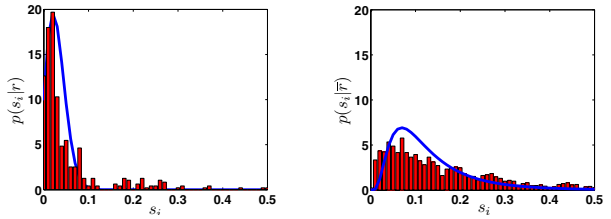
3.1. Building the rigidity priors

3.1.1 Degree of non-rigidity

Kim and Hong [10] introduced the notion of Degree of Non-rigidity (*DoN*) of a point viewed by an orthographic camera as an effective measure of the deviation of the point from the average shape. If the average 3D shape of a time varying shape $\mathtt{S}_i = [\mathbf{S}_{i1} \dots \mathbf{S}_{iN}]$ (in non-homogeneous coordinates) is given by $\bar{\mathtt{S}} = [\bar{\mathbf{S}}_1 \dots \bar{\mathbf{S}}_N]$ the Degree of Non-rigidity for point j can be expressed as $DoN_j = \sum_{i=1}^{F}(\mathbf{S}_{ij} - \bar{\mathbf{S}}_j)(\mathbf{S}_{ij} - \bar{\mathbf{S}}_j)^T$. The projection of the DoN will be thus be given by $s_j = \sum_{i=1}^{F} \mathtt{R}_i(\mathbf{S}_{ij} - \bar{\mathbf{S}}_j)(\mathbf{S}_{ij} - \bar{\mathbf{S}}_j)^T \mathtt{R_i}^T = \sum_{i=1}^{F}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^T$ where $\mathbf{x}_{ij}$ are the image coordinates of point j in frame i and $\bar{\mathbf{x}}_j$ are the coordinates of its projected mean shape. An estimate of the average projected 2D shape $\bar{\mathbf{x}}_j$ could simply be given by the rank-3 approximation of the measurement matrix $\mathtt{W}$ computed using singular value decomposition and given by $SVD_3(\mathtt{W}) = \bar{\mathtt{M}}\bar{\mathtt{B}}$. The projected deviation from the mean for all the points would then be defined by $\{\mathbf{x}_{ij} - \bar{\mathbf{x}}_j\} = \tilde{\mathtt{W}} = \mathtt{W} - \bar{\mathtt{M}}\bar{\mathtt{B}}$. Kim and Hong computed a more sophisticated estimate of the average shape, but for simplicity in this paper we have used the above description which has shown to give a good measure of the degree of deformability.

Notice that the previous definitions all assume affine viewing conditions. However, in this paper we are dealing with points in projective space so we need to re-define the measure of non-rigidity. First, the original measurement matrix $\mathtt{W}$ must be re-scaled by the estimated projective weights λ_{ij}. We calculate the projective depths λ_{ij} using subspace constraints ([9]) and express the rescaled measurement matrix as $\mathtt{W}_{rescaled} = \{\lambda_{ij}\mathbf{w}_{ij}\}$. Then we estimate the mean shape as the rank-4 approximation of the rescaled measurement matrix computed using singular value decomposition and given by $SVD_4(\mathtt{W}_{rescaled}) = \bar{\mathtt{W}} = \bar{\mathtt{M}}\bar{\mathtt{B}}$. The projected deviation from the mean would then be defined as before by $\{\mathbf{x}_{ij} - \bar{\mathbf{x}}_j\} = \tilde{\mathtt{W}} = \mathtt{W} - \bar{\mathtt{M}}\bar{\mathtt{B}}$ and the projection of the DoN can finally be computed as:

$$s_j = \sum_{i=1}^{F}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^T. \tag{8}$$



(a) (b)

Figure 1. Conditional densities for the score given: (a) that a point is rigid $p(s|r)$ or (b) non-rigid $p(s|\bar{r})$

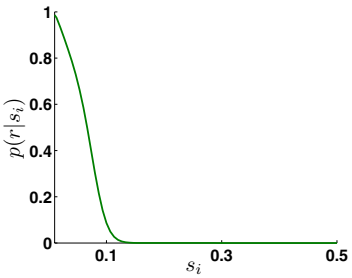


Figure 2. Estimated prior given by the estimated densities $p(s|r)$ and $p(s|\bar{r})$.

3.1.2 Computation of the prior

Tordoff and Murray [14] showed that guided sampling based on knowledge extracted from the images can greatly improve the performance of a random sampling method, especially in the presence of noise or of a high number of outliers. In these cases standard RANSAC becomes computationally prohibitive given the large number of random samples that must be drawn from the data. In this paper we use the 2D projection of the DoN defined in the previous section to provide a score s_j for each point trajectory which will be used to build a prior distribution of the conditional probability of each point in the object being rigid or non-rigid given this score.

We have inferred the conditional probability density functions for the score s given that a point is rigid $p(s|r)$ (see Figure 1(a)) or non-rigid $p(s|\bar{r})$ (see Figure 1(b)) by computing the normalised frequency histograms over many experimental trials with synthetic (and real) sequences with different perspective distortions, degrees of deformation and ratios of rigid/non-rigid points. We have then approximated the histograms by fitting appropriate analytical functions. To derive the prior conditional density function of a point being rigid given the non-rigidity score $p(r|s)$ we use Bayes theorem:

$$p(r|s) = \frac{p(s|r)p(r)}{p(s)} \propto \frac{p(s|r)}{p(s|r) + p(s|\bar{r})} \qquad (9)$$

Figure 2 shows an example of a prior obtained from our experiments. We are studying alternative measures to the DoN which might give better prior distributions of the probability of a point being rigid. Note that although the computation of the score is specific to each method the derivation of the prior given the distribution of the score is general.

3.2. Guided RANSAC

We use guided RANSAC to estimate the fundamental matrices between pairs of views. This process will be used to provide a segmentation of the image trajectories into rigid and non-rigid ones since the non-rigid trajectories will not satisfy the epipolar geometry and will therefore give a high residual in the computation of the pairwise fundamental matrices. To speed up the process we use the prior derived in the previous section to draw the point samples: points with the highest conditional probability of being rigid will be chosen more frequently. The method employed to estimate the fundamental matrix is the standard 8-point algorithm [8]. Notice that we do not consider outliers in the point matching from frame to frame.

We show results of the guided sampling RANSAC algorithm applied to the segmentation of rigid and non-rigid points in the experimental section.

4. Non-rigid shape and motion estimation

Once the scene has been segmented into the rigid and non-rigid point sets we compute metric non-rigid shape in two steps. First we use the rigid point set to estimate the camera intrinsic parameters – which provide the necessary information to upgrade the structure to metric – and the overall rotations and translations. Secondly we formulate the estimation of metric non-rigid shape as a global non-linear minimization.

4.1. Computing the metric upgrade

Given a set of rigid points observed by a perspective camera a variety of projective reconstruction algorithms [13, 9] can be applied to the rigid trajectories to obtain the 3D shape in projective space. It is then possible to upgrade this reconstruction to a metric one by using a self-calibration method which will provide estimates of the internal camera calibration parameters and of the motion and shape parameters in Euclidean space. We have used the well-known self-calibration method proposed by Pollefeys et al. [12]. The main advantage of this method is that it allows to impose different constraints on each of the camera intrinsic parameters (focal length, principal point and aspect ratio) since the camera calibration matrix K is parameterized explicitly in terms of them. Each of the parameters may be considered to be known, unknown but constant between

views or unknown and varying. As we will show in the experimental section the algorithm provides good results even when a very small number of rigid points is used.

4.2. Non-linear optimization

Our approach to solve for the non-rigid shape and motion given the 2D image tracks is to minimize image reprojection error. The cost function being minimised is the geometric distance between the measured image points and the estimated reprojected points $\chi = \sum_{i,j} \parallel \mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij} \parallel^2 = \sum_{i,j} \parallel \mathbf{x}_{ij} - \mathtt{P}_i \mathbf{X}_{ij} \parallel^2$ where $\mathtt{P}_i$ is the projection matrix for frame i and $\mathbf{X}_{ij}$ are the metric 3D coordinates of point j in frame i. We parameterize the projection matrices in terms of the calibration matrices $\mathtt{K}_i$, the rigid rotation matrices $\mathtt{R}_i$ and the translation vectors $\mathbf{T}_i$. The coordinates of the non-rigid points $\mathbf{X}_{ij}$ can be parameterized in terms of the basis shapes $\mathtt{B}_{jd}$ and the deformation coefficients l_{id}. We may now write the non-linear minimization scheme as:

$$\min_{\mathtt{K}_i \mathtt{R}_i \mathbf{T}_i \mathbf{B}_{dj} l_{id}} \sum_{i,j} \parallel \mathbf{x}_{ij} - \mathtt{K}_i [\mathtt{R}_i | \mathbf{T}_i] \begin{bmatrix} \sum_{d=1}^{D} l_{id} \mathbf{B}_{dj} \\ 1 \end{bmatrix} \parallel^2 \quad (10)$$

This problem is known as bundle-adjustment and it can be solved using a sparse implementation of a non-linear minimization algorithm such as Levenberg-Marquardt.

4.3. Initialization

The initial estimates for the calibration matrices $\mathtt{K}_i$, the overall rotations $\mathtt{R}_i$ and translations $\mathbf{T}_i$ have been obtained directly from the self-calibration algorithm applied on the rigid point set. An initial estimate for the first basis shape of the non-rigid points (which encodes the mean shape) can be easily computed given the rescaled measurement matrix of the non-rigid points $\mathtt{W}^{nonrigid}$, and the projection matrices $\mathtt{P}_i$ (provided by the self-calibration algorithm) using the expression:

$$\begin{bmatrix} \mathbf{B}_1^{nonrigid} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathtt{P}_1 \\ \vdots \\ \mathtt{P}_F \end{bmatrix}^{+} \mathtt{W}^{nonrigid} \quad (11)$$

where $\mathbf{B}_1^{nonrigid}$ encodes the first basis of the non-rigid points. The coordinates of the rest of the basis shapes which encode the $D-1$ non-rigid components $\mathtt{B}_d$ with $d = 2, \ldots, D$ are initialised to small random values. Finally, the configuration weights associated with the mean shape l_{i1} are initialised to 1 while the rest of the weights l_{id} are initialised to small values. This randomised initialisation of basis shapes and coefficients has previously been used by different authors [16, 6, 5] who report good convergence of the algorithm.

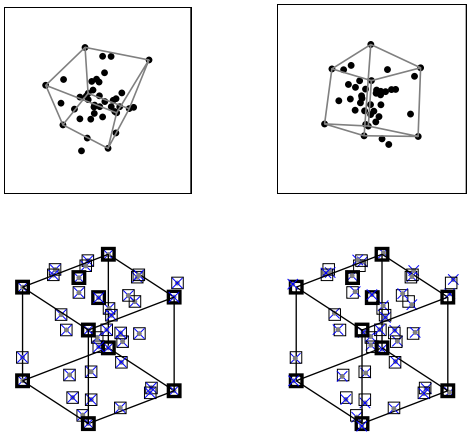


Figure 3. First row: 2 views of the synthetic sequence with camera setup 2. Second row: Ground truth (squares) vs reconstruction (crosses) for (left) no noise (right) noise=2 pix.

4.4. Rigidity priors

If the motion of a point j is completely rigid for the entire sequence, the structure referring to that point can be expressed entirely by the first basis ($d = 1$) called the rigid basis. From this it follows that for a rigid point $\mathbf{B}_{dj} = \mathbf{0} \quad \forall \; d > 1$ where $\mathbf{B}_j = [\mathbf{B}_{1j}^T \ldots \mathbf{B}_{Dj}^T 1]^T$. Note that $\mathbf{B}_j$ is a $3D+1$ vector which encodes the D basis shapes for point j and $\mathbf{B}_{dj}$ is a 3-vector which contains 3D coordinates of basis shape d for point j. Notice that this forces $3(D-1)$ zeros in each column of the shape matrix corresponding to a rigid point. We write these expectations as priors on the coordinates of the basis vectors $\mathbf{B}_{dj}$ and solve the problem as a Maximum A Posteriori (MAP) estimation.

5. Results

5.1. Experiments with synthetic data

The synthetic 3D data consisted of a set of random points sampled inside a cube of size $50 \times 50 \times 50$ units. Different sequences were generated using 10 points which remained rigid throughout the sequence (which included the vertices of the cube) and 10, 20, 30, 40 and 80 which deformed. The deformations for the non-rigid points were generated using random basis shapes as well as random deformation weights. The first basis shape had the largest configuration weight equal to 1.

The data was projected onto 50 images using a perspective camera model and applying rotations and translations over all the axes. We used 2 different camera setups with different levels of perspective distortion. We varied the distance of the object to the camera and the focal length (setup 1: d=150, f=800; setup 2: d=100, f=500). Figure 3 shows

Noise level	0	0.5	1	1.5	2
10/10	0	0.2	0.3	0.5	0.6
10/20	0	0.3	0.4	0.5	0.7
10/40	1.6	1.9	1.7	2.2	2.3
10/80	1.8	2.0	2.7	2.6	2.4

Table 1. Mean misclassification error (expressed in number of points) for different levels of noise and different ratios of rigid/non-rigid points.

two frames of a sequence obtained with camera setup 2. For all the experiments we assumed that the focal length, aspect ratio and the principal point were constant over the sequence while the skew was set to be 0. Gaussian noise of different levels was added to the image coordinates.

5.1.1 Motion segmentation

The experimental setup described above is first used to obtain an indication of the validity of our segmentation approach presented in Section 3. Guided RANSAC was performed using a conditional density prior given the non-rigidity score that was generated empirically from the data. The camera setup with strongest perspective distortions (d=100, f=500) was used and the ratio of rigid/non-rigid points was varied to study the performance of the algorithm given increasing numbers of outliers.

Various trials were performed with the number of random samples fixed to 2500. The distance threshold t which decides whether a point is an inlier or an outlier (rigid or non-rigid in this case) was set empirically to be $t = 4.12$. It was fixed by taking into account the sum of the residuals given by the estimation of F-1 fundamental matrices using normalised coordinates. Table 1 shows the degree of misclassification (measured as number of misclassified points) for varying ratios of rigid/non-rigid points and for increasing levels of noise. Note that a very good segmentation is achieved for ratios of rigid/non-rigid points above $10/40$. At $10/40$ the mean misclassification error was approximately of 2 points and higher values of misclassification appear for a rigid/non-rigid point ratio of $10/80$. We have noticed a better algorithmic behaviour in the case of stronger perspective distortions compared to weaker ones since the effects of perspective distortions and deformations are less ambiguous in such cases. Obviously, a misclassification error in the segmentation will affect the recovery of the 3D structure and camera parameters. We will analyse this situation in the next section and we will see that for the error levels we see here, the effect of misclassification is not very significant on the final 3D reconstruction.

5.1.2 3D metric reconstruction

The aim of this section is to show the performance of our reconstruction approach under different situations assuming that a correct segmentation has been achieved. In particular, we performed an exhaustive evaluation over three different sets of experiments:

1. Varying the number of basis shapes ($D = 3$, 4 and 5). We used 10 rigid and 30 non-rigid points with the first camera setup to project the 3D data.
2. Varying the ratio of rigid/non-rigid points. We used 10 rigid points while varying the number of non-rigid to 10, 20, 40 and 80. We fixed the number of basis shapes to 3 and used the first camera setup.
3. Varying the levels of perspective distortion (camera setup 1 and 2). We fixed the number of basis shapes to 5 and used 10 rigid and 30 non-rigid points.

We also used the last experimental setup to test the behaviour of our algorithm when some non-rigid points were misclassified as rigid points after the automatic segmentation. Following the results obtained in the previous section we added 2 non-rigid points to our rigid set.

The results obtained for the three sets of experiments are summarised in Figure 4 where we show the r.m.s. 2D image reprojection error expressed in pixels, 3D reconstruction error expressed in percentage relative to the scene size (which we defined as the maximum of the x, y and z coordinates) and absolute rotation error expressed in degrees. The plots of this figure show the mean values corresponding to 5 random trials on each level of noise. Our proposed method appears to perform well in the presence of image noise. Note that the 3D reconstruction error is well below 1.5% even for large perspective distortions and a large proportion of non-rigid points. The 2D error is also small and it appears to be of the same order as the image noise. The bottom row of Figure 4 also includes results obtained for both camera setups when 2 non-rigid points were incorrectly classified as rigid points showing that the impact of the presence of some outliers in the motion segmentation on the 3D reconstruction is not severe.

The non-linear optimization algorithm usually converges within around 30 iterations. Results in Figure 4 show that the algorithm always converges in the absence of noise. In this sense, the added priors are fundamental to avoid local minima given by ambiguous configurations of motion, perspective distortion and deformation parameters. Besides, the convergence in the absence of noise validates the randomised initialization used for the non-rigid basis shapes and coefficients.

Table 2 summarises the results of the focal length estimation. Observe that reliable estimates are obtained even in the presence of noise.

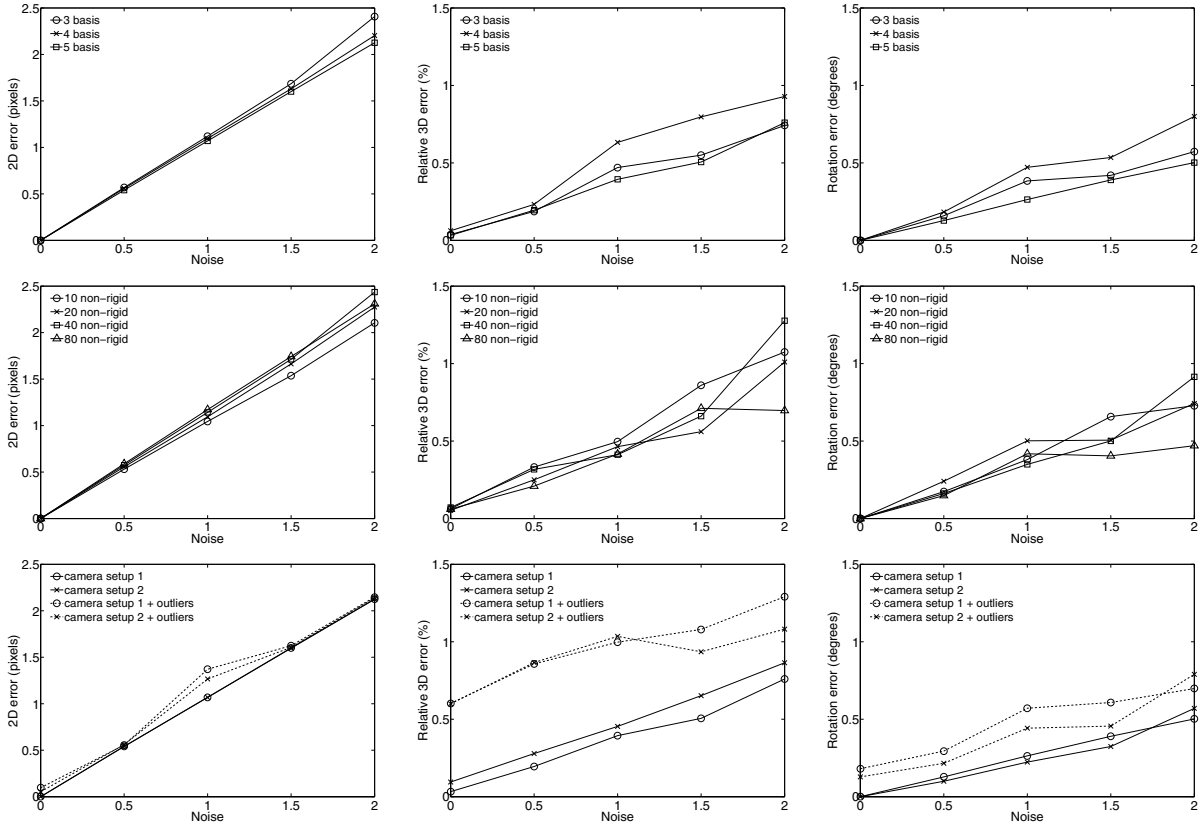


Figure 4. 2D, 3D and rotation error curves for each set of experiments: different number of basis shapes (top row), rigid/non-rigid points (middle row) and camera setups (bottom row). The bottom row also includes results obtained when 2 non-rigid points were incorrectly classified as rigid showing that the impact of errors in the motion segmentation on the 3D reconstruction is not severe.

Noise level	0	0.5	1	1.5	2
mean	0	0.18	0.50	0.88	1.17
std. dev	0	0.15	0.48	0.83	0.93
maximum	0	0.70	2.04	3.51	5.15

Table 2. Mean, standard deviation and maximum relative error (%) of the focal length for the different levels of noise. Results obtained over all the experimental tests.

5.2. Real experiment

In this experiment we use real 3D data of a human face undergoing rigid motion while performing different facial expressions. The 3D data was captured using a VICON motion capture system by tracking a subject wearing 37 markers on the face. The 3D points were then projected synthetically onto an image sequence 74 frames long using a perspective camera model. The size of the face model was $169 \times 193 \times 102$ units and the camera setup was such that the subject was at a distance of 300 units from the camera and the focal length was 600 pixels so the perspective effects were significant.

In this case the segmentation of points was done manually selecting 14 rigid points situated on the nose, temples and the side of the face. Figure 5 shows the ground truth and reconstructed shapes from front, side and top views. The selected set of rigid points is highlighted in the frontal view of the first frame. The obtained 2D reprojection error was 0.67 pixels, the absolute 3D error was 2.24 units while the estimated focal length was 595.12. The results are satisfactory considering that the selected rigid points were not perfectly rigid during all the sequence. Note the deformations are very well captured by the model even for the frames in which the facial expressions are more exaggerated.

6. Conclusions

In this paper we have presented a new approach to the computation of metric non-rigid shape from a sequence of uncalibrated images. The method first performs a segmentation of scene points into rigid and non-rigid and then obtains the metric upgrade information and estimates for the overall motion and mean shape from the rigid point set. The estimation of non-rigid shape is then formulated as a non-

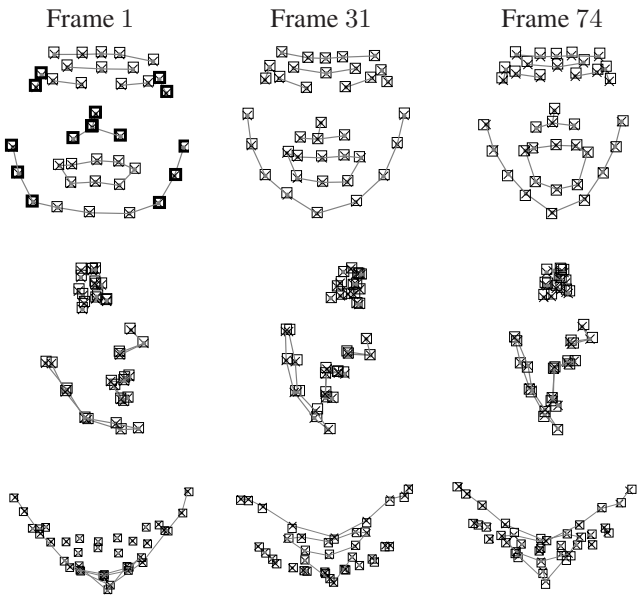


Figure 5. Front, side and top views of the reconstructed face for frames 1, 31 and 74. Crosses indicate estimated reconstructed points while squares refer to the ground truth. Highlighted marks on the frontal view of frame 1 indicate rigid points.

linear optimization problem. Our experiments show that even when using a minimal set of rigid points it is possible to obtain reliable metric information. We have also shown that the reconstruction algorithm appears to have some resilience to some of non-rigid points being misclassified as being rigid.

7. Acknowledgements

This work was supported by EPSRC grant GR/S61539/01. Alessio Del Bue holds a Queen Mary, University of London PhD studentship.

References

[1] H. Aanæs and F. Kahl. Estimation of deformable structure and motion. In *Workshop on Vision and Modelling of Dynamic Scenes, ECCV'02, Copenhagen, Denmark*, 2002.

[2] M. Brand. Morphable models from video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, December 2001.

[3] M. Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *CVPR (2)*, pages 122–128. IEEE Computer Society, 2005.

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 690–696, June 2000.

[5] A. Del Bue, X. Llado, and L. Agapito. Non-rigid face modelling using shape priors. In S. G. . X. T. W. Zhao, editor, *IEEE International Workshop on Analysis and Modelling of Faces and Gestures, held in conjunction with ICCV-05*, volume 3723 of *Lecture Notes in Computer Science*, pages 96–107. Springer-Verlag, 2005.

[6] A. Del Bue, F. Smeraldi, and L. Agapito. Non-rigid structure from motion using nonparametric tracking and non-linear optimization. In *IEEE Workshop in Articulated and Non-rigid Motion ANM04, held in conjunction with CVPR2004*, pages 8–8, Washington, June 2004.

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pages 726–740. Kaufmann, Los Altos, CA., 1987.

[8] R. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, June 1997.

[9] A. Heyden, R. Berthilsson, and G. Sparr. An iterative factorization method for projective structure and motion from image sequences. *Image and Vision Computing*, 17(13):981–991, November 1999.

[10] T. Kim and K.-S. Hong. Estimating approximate shape and motion of deformable objects with a monocular view. In *Proc. Asian Conference on Computer Vision*, Jeju Island, Korea, January 2004.

[11] S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, volume 2, pages 430–437, June 2000.

[12] M. Pollefeys, R. Koch, and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

[13] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. 4th European Conference on Computer Vision, Cambridge*, pages 709–720, April 1996.

[14] B. Tordoff and D. Murray. Guided-MLESAC: Faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1523–1535, October 2005.

[15] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[16] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2001.

[17] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *Proc. 10th International Conference on Computer Vision, Beijing, China*, October 2005.

[18] J. Xiao, J. Chai and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Proc. 8th European Conference on Computer Vision, Prague, Czech Republic*, May 2004.