Non-Rigid Structure from Motion with Diffusion Maps Prior

Lili Tao Bogdan J. Matuszewski
Applied Digital Signal and Image Processing Research Centre
University of Central Lancashire, UK
{lltao,bmatuszewski1}@uclan.ac.uk

Abstract

In this paper, a novel approach based on a non-linear manifold learning technique is proposed to recover 3D non-rigid structures from 2D image sequences captured by a single camera. Most of the existing approaches assume that 3D shapes can be accurately modelled in a linear subspace. These techniques perform well when the deformations are relatively small or simple, but fail when more complex deformations need to be recovered. The non-linear deformations are often observed in highly flexible objects for which the use of the linear model is impractical.

A specific type of shape variations might be governed by only a small number of parameters, therefore can be well-represented in a low dimensional manifold. We learn a non-linear shape prior using diffusion maps method. The key contribution in this paper is the introduction of the shape prior that constrain the reconstructed shapes to lie in the learned manifold. The proposed methodology has been validated quantitatively and qualitatively on 2D points sequences projected from the 3D motion capture data and real 2D video sequences. The comparisons of the proposed manifold based method against several state-of-the-art techniques are shown on different types of deformable objects.

1. Introduction

The objective of the Structure from Motion (SfM) is to jointly reconstruct 3D shapes and estimate corresponding camera motion trajectories based only on a set of observed image sequences. While the reconstruction of rigid objects has been well-established [19] over the past two decades, deformable shape reconstruction is still challenging, mainly because it is a severely under-constrained problem. This is particularly true for the articulated objects or the object which contains large and complex deformations. Such time-varying shape recovery problem is referred to as Non-Rigid Structure from Motion (NRSfM).

Bregler *et al.* [5] was the first to adopt the factorisation algorithm to deformable 3D structures by introducing a low rank shape model to represent deformable shapes. As a time-varying object usually cannot arbitrarily deform, the idea of this model is to represent a deformable shape as a linear combination of basis shapes. Due to its simplicity, shape basis model has been widely used to tackle the NRSfM [4, 26, 1]. Non-rigid articulated structure representation has also been formulated following the idea of factorisation algorithm [21, 15]. However because of inherently high number of degrees of freedom and motion degeneracy, these methods may fail to provide meaningful reconstruction. To counter this effect, it is common to introduce prior information to define additional constraints for minimisation of the re-projection error. Xiao *et al.*[25] proposed a closed-form solution and showed that orthonormality constraints is insufficient to provide unique solution to estimate basis shapes. Torresani *et al.*[20] employed a form of Probabilistic Principal Components Analysis to provide Gaussian distribution on deformation coefficients as prior knowledge. Del Bue [8] proposed an alternative approach introducing a single shape prior coupled with a bundle adjustment refinement. These approaches may improve the performance for both non-rigid and articulated SfM, obtaining reliable 3D reconstruction but only if an appropriate initial value is provided. Departing from the shape basis model, a trajectory based algorithm was proposed in [2] by Akhter *et al.* who described a duality theorem in 3D structure representation which models independent 3D point trajectories. The main advantage of this representation is that the basis trajectories can be predefined, thus removing a large number of unknowns from the estimation.

The restriction of the most existing NRSfM methods is that they try to explain the complex deformations using a global model. An alternative piecewise model has been recently developed [23, 18, 10]. This model is able to cope well with strongly deforming objects. However, necessity for dividing the surface into a set of overlapping patches (often preformed manually) is generally viewed as the severe drawback of this model.

To move away from the linear combination of basis shapes, Rabaud and Belongie [16] integrated the Lo-

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.201

cally Smooth Manifold Learning algorithm to regularise the NRSfM problem. However, there is no guarantee that the manifold is planar or isometric to a plane. Despite the manifold learning techniques are becoming increasingly popular and have been successfully used in different applications including medical image analysis [24], object classification [14] and segmentation [9], these techniques have not been widely applied in NRSfM problem.

2. Contributions

Although the tremendous progress has been achieved towards solving the Structure from Motion for deformable shapes, one of the still existing limitations of the methods proposed so far is that they mainly address the problem of small deformations. The main reason for their failure when recovering objects with large, complex deformations can be attributed to the reliance on a linear shape model. This paper focuses on modelling non-linear deformable objects with large complex deformations, such as deformable cloth or articulated full-body motion. In this case, the existing methods based on linear space manifold are no longer applicable. We argue that the linear models require more parameters than our method based on non-linear manifold learning approach.

This paper proposes a novel method for reconstruction of 3D deformable structures exhibiting large and complex deformations. The proposed method is based on a recently introduced manifold learning technique called Diffusion Maps [6]. This manifold build as a shape prior, with the reconstructed shapes constrained to lie in the manifold. Our method achieves good results when dealing with objects undergoing significant and complex deformations. In case of articulated deformations, *e.g.*, full-body movement, rather than having an initial segmentation stage to assign different body parts [15] which may lead to unexpected errors, the whole data are considered as a single entity without the need for body part recognition. Learning instead a corresponding low dimensional manifold from the training examples. Such techniques have rarely been applied in the context of non-rigid shape reconstruction. Our approach is to integrate the learned non-linear shape prior manifold into the NRSfM solver. The advantage of our method is that it can be adopted for reconstruction of highly deformable, complex objects.

3. Problem Formulation

Considering a set of 2D images captured by a single camera, tracking *P* feature points in *F* video frames, the $2F \times P$ measurement matrix can be formed as:

$$\mathbf{W}=\begin{bmatrix}\mathbf{x}_{11} & \cdots & \mathbf{x}_{1P}\\ \vdots & \mathbf{x}_{tp} & \vdots\\ \mathbf{x}_{F1} & \cdots & \mathbf{x}_{FP}\end{bmatrix}=\begin{bmatrix}\mathbf{R}_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathbf{R}_F\end{bmatrix}\begin{bmatrix}-\mathbf{S}_1-\\ \vdots\\ -\mathbf{S}_F-\end{bmatrix}=\mathbf{RS} \quad (1)$$

where $\mathbf{W}$ is an observation matrix and contains 2D input points $\mathbf{x}_{tp} = [x_{tp}, y_{tp}]^{\mathrm{T}}$ with indices *t* and *p* referring to the p^{th} point in the t^{th} image. Without loss of generality, we assume that the coordinates of the feature points are given with respect to the centre of gravity calculated for all the points in the corresponding image. We also assume that the orthographic projection accurately models the image acquisition.

The goal is to recover camera orientations matrix $\mathbf{R}$ and the concatenated time-varying shapes matrix $\mathbf{S}$, based only on the 2D measurement in matrix $\mathbf{W}$. It is an under constrained problem since the shape and motion are both changing with time. The number of unknown variables (3*F*+3*FP*) is higher than the number of observed input data (2*FP*) from the observation. To deal with this, two models have proved to be successful.

Low-rank shape model

The points in each observed image can be represented as $\mathbf{x}_t = \mathbf{R}_t\mathbf{S}_t$, where $\mathbf{x}_t$ represents input points, $\mathbf{R}_t$ is a 2×3 projection matrix representing camera orientation and $\mathbf{S}_t \in \mathbb{R}^{3\times P}$ is a 3D shape projected onto the t^{th} frame. Describing the deformation using a shape model in a linear subspace is one way of imposing compactness on $\mathbf{S}$ to reduce the dimensionality of the problem. A deformable 3D shape can be represented as a linear combination of *K* unknown but fixed basis shapes $\mathbf{B}_l$:

$$\mathbf{S}_t = \sum\nolimits_{l=1}^{K} \alpha_{tl}\mathbf{B}_l \quad (2)$$

where $K \ll F, P$. The deformation coefficients α_l are adjustable over time. This low-rank shape model can be obtained by performing Singular Value Decomposition (SVD) or Principal Components Analysis (PCA). The measurement matrix can be decomposed and represented by pose, basis shapes and time varying coefficients matrices, therefore it can be rearranged as:

$$\mathrm{W} = \begin{bmatrix}\alpha_{11}\mathbf{R}_1 & \cdots & \alpha_{1K}\mathbf{R}_1\\ \vdots & \ddots & \vdots\\ \alpha_{F1}\mathbf{R}_F & \cdots & \alpha_{FK}\mathbf{R}_F\end{bmatrix}\begin{bmatrix}-\mathbf{B}_1-\\ \vdots\\ -\mathbf{B}_K-\end{bmatrix} = \mathbf{MB} \quad (3)$$

Since basis shapes $\mathbf{B} \in \mathbb{R}^{3K\times P}$, and $\mathbf{M} \in \mathbb{R}^{2F\times 3K}$ the rank of measurement matrix $\mathbf{W}$ is 3*K* at most in the absence of noise. The factor $\mathbf{M}$ and $\mathbf{B}$ are computed by factorising the measurements $\mathbf{W}$. The solution is not unique and is defined up to a ambiguity matrix $\mathbf{Q} \in \mathbb{R}^{3K\times 3K}$. According to [25], the limitation of the closed-form solution in this approach is that the motion matrix is nonlinear, when an inaccurate set of basis shapes have been chosen, it may not be possible to remove the affine ambiguity.

Smooth trajectories model

According to the duality theorem, as described in [2],

	Shape	Trajectory	Proposed
Camera	$3F$		
Coefficients	FK	$3KP$	$F(K+1)$
Basis	$3KP$	/	/
Total	$3F+FK+3KP$	$3F+3KP$	$3F+F(K+1)$

Table 1: Comparison of number of unknowns in low-rank shape model, trajectory model and our proposed non-linear manifold model

representing a non-rigid shape using the above shape basis model is dual to trajectory basis model, in which each point trajectory is represented as a K dimensional point within an unknown linear trajectory space. The trajectory for each point is approximated by a linear combination of a small number of basis trajectories $\mathbf{A}_l$:

$$\mathbf{T}_p = \sum_{l=1}^{K} \mathbf{A}_l \boldsymbol{\beta}_{pl} \tag{4}$$

where $\boldsymbol{\beta}_{pl}$ are 1×3 coefficient vectors for the basis trajectory The basis trajectory can be predefined in an object independent way using Discrete Cosine Transform (DCT) basis and therefore avoid training process. The model only needs to consider camera parameters and trajectory coefficients, thus requires less parameters than shape basis model (see Table 1).

Non-linear manifold model

Our model departs from the linear shape model. The shape basis $\mathbf{B}$ in the proposed method are selected from the learned shape manifold. Unlike the low rank shape model, where all the reconstructed shapes are represented as a linear combination of unknown but fixed K basis shapes, in the proposed method, the basis shapes may be different at each frame. Although it may seem to increase the number of parameters in the model, it should be recognised that all the basis shapes are selected from the manifold and are not estimated as a part of the optimisation process. The parameters to be estimated in the proposed approach include only the camera motion and shape coefficients, representing the shape in the local linear barycentric coordinates system approximating the manifold at the location corresponding to the current estimate of $\mathbf{S}_t$.

Comparing the three models together, the number of unknowns for each model is given in Table 1. In most cases, $K < 10$, $F, P > 100$, the proposed model requires less parameters than low rank shape model and has a similar order of magnitude as trajectory model. Although the number of parameters depends on number of frames F in our method, it is important to note that they are not depending on the number of feature point P. That makes our approach suitable for the shape which contains large number of feature points.

4. NRSfM with Diffusion Maps

In this section, an overview of the proposed manifold based NRSfM algorithm is given first, followed by a short description of the diffusion maps including description of out-of-sample and pre-image problems.

As known from [25], enforcing only the rotation constraints cannot guarantee the unique solution for the camera motion and the basis shapes. To solved this, the designed shape prior can help to attract a shape towards the manifold and therefore avoid incorrect reconstruction.

A summary of the algorithm for recovery of non-rigid object and estimation of camera motion is given in Algorithm 1. Initial shapes $\mathbf{S}'$ and camera motion $\mathbf{R}'$ are estimated by running a few iteration of the optimisation process in batch NRSfM using linear basis shapes model [17]. For each initial shape, Nyström extension is used for embedding these new samples into the reduced space. Intuitively, if the points in reduced space are relatively close, the corresponding shapes in high-dimensional space should represent similar shapes. Based on this observation, the reconstructed shape at each frame can be represented as weighted sum of K+1 basis shapes from the learned manifold. The coefficients of correspond basis shape are calculated based as barycentric coordinates of K+1 closest points in reduced space. Once the basis shapes and their coefficients have been obtained, an optimisation is applied to minimise the image reprojection error with an additional smoothing term and basic rotation constraint over all frames. However, the quality of optimisation result is depending on the accuracy of initial shapes. Updating basis shapes in each iteration can help to circumvent the problem. The basis shapes are being kept updated as long as 2D measurement error r_t exceeds the defined threshold r_T (10^{-3} in our case) or the error between two adjacent frames is relatively large which implies that the current results are unlikely to explain the shapes well.

4.1. Diffusion Maps

In contrast to linear methods, non-linear approaches are able to handle a wider range of data variability and preserving local structures at the same time. While linear manifold method like PCA is straightforward, the recovered input data lies on a linear subspace of high dimensional space. The problem with this is that the input data may have complex non-linear dependencies and preserving local or indeed global structures in the data may not be possible utilising linear projections.

Diffusion maps is a graph based technique with isometric mapping from original shape space to reduced low-dimensional diffusion space. Assuming $\mathbf{X}$ is a dataset with M samples, the goal of dimensionality reduction problems is to find an embedding from data $\mathbf{X} = \{\mathrm{X}_1 \cdots \mathrm{X}_M\}$ in high N dimensional space to reduced K dimensional space

Algorithm 1 Outline of Diffusion Maps based NRSfM

```
Input: Stream of 2D observations, diffusion map $\Psi$ of training dataset $\mathbf{X}$ (Section 4.1)
Output: 3D deformable shapes $\mathbf{S}$ and camera motion $\mathbf{R}$ for each frame.
1: Initialisation of estimating Initial shapes $\mathbf{S}'$ and camera motion $\mathbf{R}'$.
2: while $(\|r\| > r_T)$ or $\left(\|r_t\| - \|r_{t-1}\| > 10^{-3}\right)$ do
3:     Shape projection onto manifold (shape Embedding) (Section 4.2)
4:     Find $K$+1 closest points $\mathbf{b}_l, l = 1 \cdots K$+1 in low dimensional space, where $K$ is the dimensionality of the reduced space.
5:     Shape update (Section 4.3)
6:     Non-linear optimisation by minimising 2D measurement error and shape smooth term to obtain updated shapes $\mathbf{S}_t$ and camera motion $\mathbf{R}_t$,$t$=$1 \cdots F$ .(Section 4.4)
7: end while
```

x=$\{\mathrm{x}_1 \cdots \mathrm{x}_M\}$. A mapping is defined by: $\Psi : \mathbf{X} {\mapsto} \Psi(\mathbf{X}) = (\Psi_1(\mathbf{X}), \cdots, \Psi_K(\mathbf{X}))$, where $\mathbf{X} \in \mathbb{R}^N$, $K \ll N$.

Given a set of shapes $\mathrm{X}_1 \cdots \mathrm{X}_M \in \mathcal{M}$, where $\mathcal{M}$ is the manifold embedded in $\mathbb{R}^N$, Euclidean distance for each pair of shapes $\|\mathrm{X}_i - \mathrm{X}_j\|^2$ is calculated to build an adjacency graph. The entries of the adjacency matrix W_{ij},$i, j \in 1 \ldots M$ define the weighted similarity graph for all connected vertexes. Using Gaussian kernel $W_{ij} = \exp(-\|\mathrm{X}_i - \mathrm{X}_j\|^2 / 2\delta)$ in this case, where δ is calculated as $\delta = \frac{1}{M}\sum_{i=1}^{M} \min\limits_{j:\mathrm{X}_i \neq \mathrm{X}_j} \|\mathrm{X}_i - \mathrm{X}_j\|^2$. We also apply k-nearest neighbour (kNN) sparsification scheme, retaining k edges for each point and remove other connections to avoid outliers. The work presented in [6], shown that the diffusion distances describe the intrinsic geometric linking of the adjacency matrix, and the diffusion map Ψ can be constructed as:

$$\Psi : \mathrm{X}_i \mapsto [\lambda_1\varphi_1(\mathrm{X}_i), \cdots, \lambda_K\varphi_K(\mathrm{X}_i)]^T \tag{5}$$

The mapping is expressed using eigenvectors φ and eigenvalues λ of diffusion operator P=$D^{-1}\hat{W}$,where each entry $p(\mathrm{X}_i, \mathrm{X}_j)$=$\hat{W}_{ij}/d_{ii}$ with d_{ii}=$\sum_{j=1}^{M} \hat{W}_{ij}$ and d_{ij}=0 for all $i \neq j$,$\forall d{\in}D$. An anisotropic normalized graph Laplacian [7] has been used for renormalizing the adjacency matrix $\hat{W}_{ij} = W_{ij}/q_i q_j$, in which $q_i = \sum_{j=1}^{M} W_{ij}$,$q_j = \sum_{i=1}^{M} W_{ji}$. Fig.1 illustrates the embedding of shapes from *cardboard* data [22] together with representative corresponding shapes extracted from 1000 training samples.

4.2. Out of sample extension

In general, the diffusion map Ψ is only able to provide an embedding for the data which is given in the training set. However, in the NRSfM problem, it is necessary to calculate embedding for shapes which are not presented in the training set. To extent the embedding for new data, the mapping can be approximated with the Nyström extension [3]. Suppose $\mathbf{S}_t \in \mathbb{R}^N$is a new data which has not been presented in the training set. Knowing that for every sample in training dataset:

$$\forall \mathrm{X}_i {\in} \mathbf{X}, \sum_{\mathrm{X}_j \in \mathbf{X}} p(\mathrm{X}_i, \mathrm{X}_j)\varphi_k(\mathrm{X}_j) {=} \lambda_k \varphi_k(\mathrm{X}_i), k{=}1 \ldots M \tag{6}$$

Having a shape $\mathbf{S}_t$ not present in the training set $\mathbf{X}$, an embedding $\mathbf{S}_t \mapsto \left(\hat{\Psi}_1(\mathbf{S}_t), \cdots, \hat{\Psi}_K(\mathbf{S}_t)\right)$ of this new shape is calculated from:

$$\hat{\Psi}_k(\mathbf{S}_t) {=} \sum\nolimits_{\mathrm{X}_j \in \mathbf{X}} p(\mathbf{S}_t, \mathrm{X}_j)\varphi_k(\mathrm{X}_j) \tag{7}$$

where $p(\mathbf{S}_t, \mathrm{X}_j)$ is calculated the same as in Diffusion maps.

4.3. The pre-image problem

The pre-image problem is concerned with finding the inverse mapping of a point $\mathbf{x} \in \mathbb{R}^K$ given in the reduced space back to the manifold $\mathrm{X}_i = \Psi^{-1}(\mathrm{x}_i)$,with $\mathbf{X} \in \mathbb{R}^N$. Assuming we look for a shape $\mathbf{S}_t$ given by its embedding $\mathbf{x}_t$, if this shape $\mathbf{S}_t$ does not exist in the training dataset, the exact pre-image might not be found in that case. To resolve this problem, Arias *et al.* proposed to find an approximate pre-image by optimising a certain optimality criteria [3]. Inspired by this, we assume that the pre-image can be represented as a linear combination of its neighbours on the manifold selected from the training samples. The simplest way to achieve this is to identify the K+1 closest points of $\mathbf{x}_t$ in the reduced space. This can be efficiently calculated by using a Delaunay triangulation. Since diffusion maps provides isometric mapping the data must keep the same structure when embedded into the reduced space and therefore the neighbours on the manifold correspond to the closest neighbours in the reduced space. Each point $\mathbf{x}_t$ can be represented as $\mathbf{x}_t = \sum_{l=1}^{K+1} \theta_{tl}\mathbf{b}_{tl}$, where $\mathbf{b}_{tl}$ is the l^{th} nearest point of $\mathbf{x}_t$. The weights θ_{tl} are computed as the barycentric coordinates of x_t, thus can be obtained by optimising the following function:

$$\arg\min_{\theta_{tl}} \sum_{t=1}^{F} \left\| \mathbf{x}_t {-} \sum_{l=1}^{K+1} \theta_{tl}\mathbf{b}_{tl} \right\|^2 \ \textit{with} \ \sum_{l=1}^{K+1} \theta_{tl} {=} 1, 0 {\leq} \theta_t {\leq} 1 \tag{8}$$

Once the weights θ_{tl} are estimated, The shape $\mathbf{S}_t$ can be approximated as a set of weighted training samples $\mathbf{S}_t$=$\sum_{l=1}^{K+1} \theta_{tl}\mathbf{B}_{tl}$, where the training sample $\mathbf{B}_{tl}$ is the pre-image of $\mathbf{b}_{tl}$.

4.4. Cost function

The cost function to be minimised consists of the reprojection error, shape smoothing terms an rotation constraint.

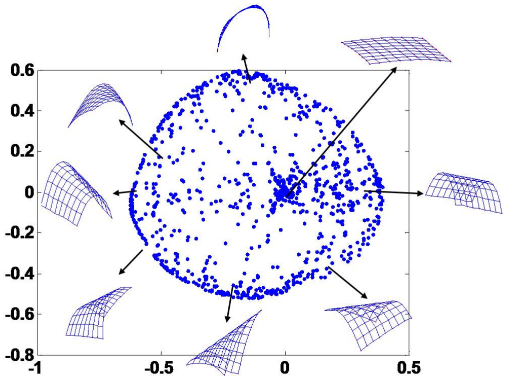

Figure 1: The reduced space of *cardboard* dataset

The cost function is given as:

$$\underset{\mathbf{R}_t,\theta_{tl}}{\arg\min}\sum_{t=1}^{F}\|\mathbf{W}_t-\mathbf{R}_t\mathbf{S}_t\|^2+\varphi_{\mathbf{S}}\sum_{t=2}^{F}\|\mathbf{S}_t-\mathbf{S}_{t-1}\|^2+\varphi_{\mathbf{R}}\sum_{t=1}^{F}\varepsilon_{rot}$$

$$with\sum_{l=1}^{K+1}\theta_{tl}=1, 0\leq\theta_t\leq 1 \qquad (9)$$

where $\varepsilon_{rot}=\left\|\mathbf{R}_t\mathbf{R}_t{}^T-\mathbf{I}\right\|^2$ enforces orthonomality of all $\mathbf{R}_t$. $\varphi_{\mathbf{S}}$ and $\varphi_{\mathbf{R}}$ are regularisation constants.

The cost function above was minimised by using Levenberg-Marquardt algorithm.

The accuracy of the optimised results strongly depends on initialisation since the mapping in the out-of-sample extension is based on initial shapes. To eliminate the effect, basis shapes are updated until the 2D measurement error is smaller than predefined threshold r_T and the error between two adjacent frames is small enough.

5. Experimental Results

The proposed methodology has been validated quantitatively and qualitatively on both motion capture and real data for different types of deformable object. To demonstrate advantages of our method over previously proposed methods, the experiments are mainly focused on reconstructing complex deformations. To demonstrate the performance of the algorithm, extensive experimental evaluation has been provided.

The models and algorithms used for comparison are as follows:

MP: The metric projection method [15].
PTA: The DCT based point trajectory approach [2].
CSF: The column space fitting method [11].
KSFM: The kernel non-rigid structure from motion [12].
IPCA: The incremental principal components analysis based method [17].
DM: The proposed method.

The data which were used for testing include: two articulated face sequences, *surprise* and *talking*, both captured using passive 3-D scanner with 3D tracking of 83 facial landmarks [13]; two surface models, *cardboard* and *cloth* [22]; five different human actions, *walking*, *pick-up*, *yoga*, *drink* and *stretch*, and three dance sequences *dance*, *Indian dance* and *Capoeira*, from CMU motion capture database. Diffusion maps requires training process, so training datasets for two face sequences are taken from the BU-3DFE [27] and for two surface sequences are obtained from [22]. Since no separate training data are available in CMU database, half of each sequence is used for manifold learning and the other half for testing. All the training data has been rigidly co-registered. Same testing data has been applied for other methods which do not require training.

5.1. The influence of embedding dimensionality

For the first set of experiments, we start with tests on motion capture data. The accuracy of 3D shape reconstruction is affected by the dimensionality of the manifold representing prior information. To find the relationship between manifold dimensionality and the reconstruction error, experiments have been carried out with all the test sequences and dimensionality changing between 3 and 10. To simplify visualization of results, all the 12 sequences are separated into 3 groups, those are: small deformation sequences (*surprise*, *talking*, *cardboard*), large deformation sequences (*cloth*, *walking*, *pick-up*, *yoga*, *drink*, *stretch*) and all the dance sequences. For evaluating the results, the normalized means of the 3D error are compared over all frames and all points:

$$e=\frac{1}{\Delta FP}\sum_{t=1}^{F}\sum_{p=1}^{P}e_{tp},\ \Delta=\frac{1}{3F}\sum_{t=1}^{F}(\Delta_{tx}+\Delta_{ty}+\Delta_{tz}) \quad (10)$$

where Δ_{tx},Δ_{ty},Δ_{tz} are the standard deviations of *x*,*y* and *z* coordinates of ground truth shape at t^{th} frame and e_{tp} is the Euclidean distance between corresponding point p at frame t in the reconstructed and ground truth shapes. Fig.2 shows the means of reconstruction error for each group and the overall average results when different manifold dimensions *K* are used. As expected, in general, increasing the number of manifold dimensions decreases the error. This is especially true for the group of dance sequences and the group representing large deformations. Higher dimensional manifolds preserve more information from the original data leading to better results. However for data with small deformations, the 3D error levels off and does not strongly depend on *K*. This does make sense as only a small number of basis shapes is required to describe the data variability containing only relatively small number of degrees of freedom.

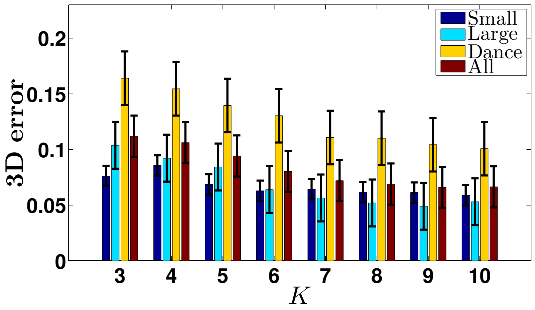

Figure 2: Average normalised mean 3D error and standard deviation of different number of dimensions in reduced space. Bars left to right: Group of small deformation sequences, group of large deformation sequences, group of all dance sequences, all the sequences.

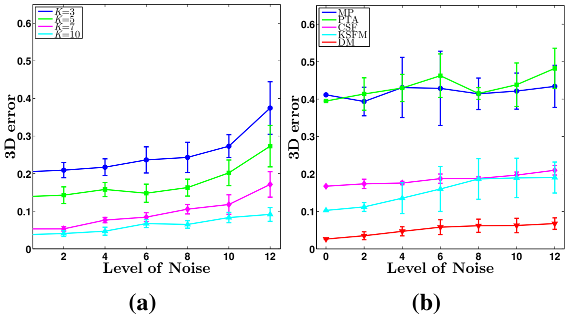

Figure 3: Reconstruction error as function of the measurement noise for the *walking* data. (a) Our method with varying number of manifold dimension K; (b) Our method evaluated against four other methods.

5.2. Sensitivity to noise

In most cases, inaccurate 2D measurement caused by feature tracking/detection error may lead to shape reconstruction failure for most previously proposed approaches, as those are very sensitive to noise. This experiment was designed to test the sensitivity of our method to different level of noise in the observations. Firstly, we performed the experiment on the *walking* data using only our method but with different level of noise and different dimensionality of the learned manifold. The measurement $\mathbf{W}$ was perturbed by Gaussian noise according to the standard deviation of the measurement data with given level of noise. The reconstruction errors are evaluated for 10 trials with each level of noise, which are set to 2%, 4%, 6%, 8%, 10% and 12%. The results are shown in Fig.3a.

Also we compare our method against several state-of-the-art techniques in terms of sensitivity to the noise present in the measurement data. We used MP, PTA, CSF, KSFM and DM for comparison here and the results are given in Fig.3b. The proposed method achieved much smaller errors both in terms of the mean and standard deviations. It is important to note that the results from our method are relatively stable and are not very sensitive to noise, even when the noise level has increased to 12%, the estimated maximum error was 0.0870. Even with the noise presented in the measurements, reasonably accurate shapes are still obtainable, showing that manifold based method can produce results that are better than those obtained by previously proposed methods.

5.3. Comparison with previous methods

For the comparative evaluation, performance of the proposed method is tested against all the 5 other approaches listed above for all 12 sequences. Table 2 summarises the results showing 3D reconstruction errors of each method and each sequence, together with the optimal number of bases for which minimal reconstruction error on the test data is obtained. We followed the same evaluation procedure as reported in [12]; the 3D errors of the PTA, CSF and KSFM methods are chosen with their best parameter K by running the trials with K varying from 2 to 13. The best result for DM method is chosen by changing manifold dimension K from 3 to 10. Considering the ambiguity of estimated camera motion [2], the shapes are aligned using a single global rotation based on Procrustes alignment method.

As shown on the Table 2, trajectory based methods PTA, CSF and KSFM are able to provide comparable results, to the proposed method on objects with small deformations (*e.g.* faces etc.). This is because these objects exhibit mostly a rigid motion, the deformations are only seen around the lips and chin. But those methods provide relatively large error on highly non-rigid human motion sequences (*e.g.* dance etc.). DM is the only method that presents accurate reconstructed results almost at all times, even for full-body motion capture sequences. Note that although the initial shapes of our method may not belong to the manifold $\mathcal{M}$, after optimisation process, the results demonstrate good convergence since the 3D errors are relatively small. An important observation is that, in the trajectory based methods, the optimal number of bases K has to be independently estimated for each sequence. Choosing too big K may lead to an ill-conditioned problem, but the point trajectory cannot be comprehensively represented if K is too small, while the results from our method are more predictable.

Finally more qualitative reconstruction comparison for different methods is shown in Fig. 5 to further highlight the accuracy of our method.

5.4. Real data experiments

We tested our approach on a video sequence showing paper being bended taken from a video camera. In the video, 81 features were tracked along 61 frames showing approximately two periods of bending movement. Fig.4 shows a comparison of our reconstructed shapes with the results obtained from MP, PTA, KSFM methods.

	MP	PTA	CSF	KSFM	IPCA	DM
Surprise	0.2558	0.0386(12)	0.0396(3)	0.0381(4)	0.1289	**0.0352**(10)
Talking	0.0991	0.0862(10)	0.0573(3)	0.0498(4)	0.0986	**0.0350**(10)
Cardboard	0.4185	0.2894(8)	0.3237(3)	0.2753(2)	0.2445	**0.1064**(10)
Cloth	0.3997	0.3526(6)	0.2609(6)	0.1806(2)	0.1909	**0.0287**(7)
Walking	0.4114	0.3948(2)	0.1675(2)	0.1029(5)	0.3264	**0.0265**(9)
IndianDance	0.4576	0.4848(13)	0.3368(7)	0.2343(7)	0.3440	**0.0981**(10)
Capoeira	0.4177	0.5127(6)	0.3648(4)	0.2376(7)	0.4059	**0.0258**(9)
Pick-up	0.4332	0.2433(12)	0.2298(6)	0.2321(6)	0.2900	**0.0634**(10)
Yoga	0.8085	0.1623(11)	0.1467(7)	0.1474(7)	0.2626	**0.0768**(10)
Drink	0.3310	0.0248(13)	0.0224(6)	**0.0186**(12)	0.0843	0.0207(9)
Stretch	0.3988	0.1087(12)	0.0709(8)	0.0736(12)	0.1918	**0.0687**(6)
Dance	0.2210	0.2935(5)	0.2684(2)	0.2369(4)	0.3058	**0.1676**(7)

Table 2: Normalised mean 3D error calculated for different sequences.

6. Comments and Future work

The paper presented a new approach to integrate the idea from non-linear manifold learning techniques into the NRSfM framework for the task of reconstructing complex and highly deformable shapes. The diffusion maps have been introduced in order to build non-linear shape prior manifold. This approach significantly improved the reconstruction quality and is well-adapted to large deformation of complex objects, especially for non-rigid articulated body movement, which cannot be accurately represented in a linear subspace. It should be pointed out that the improved performance, of the proposed method, in terms of 3D shape reconstruction accuracy comes at the cost of required availability of a representative training dataset, and therefore the comparison of the proposed method with respect to the other methods may not be seen as fair. Indeed in this sense it can be also argued that the method does not fit the definition of the SfM problem due to the use of this additional information.

As we only use limited number of shapes in training process, to overcome this, the future work would include collecting and generating data for building a sufficiently dense representation of the manifold to further improve the performance. As manifold learning has shown to be a very powerful approach for analysis of the shapes, we believe the manifold based method is a suitable groundwork for reconstruction of deformable shapes.

Acknowledgements

The motion capture data used in this project was obtained from http://mocap.cs.cmu.edu.

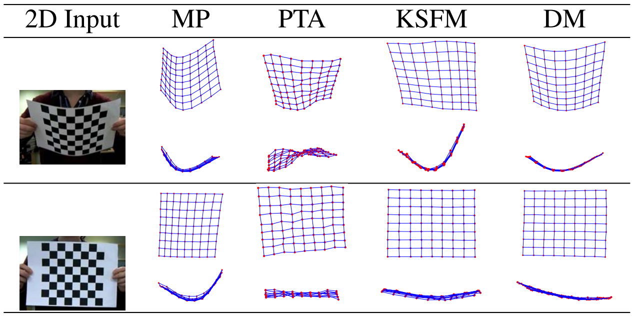

Figure 4. Selected 2D frames from the video sequence of a paper bending. Front and top views of the corresponding 3D reconstructed results using our method (DM), MP, PTA and KSFM.

References

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

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE PAMI*, 2011. 1, 2, 5, 6

[3] P. Arias, G. Randall, and G. Sapiro. Connecting the out-of sample and pre-image problems in kernel methods. In *ICPR*, 2007. 4

[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

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2001. 1

[6] R. Coifman and S. Lafon. Diffusion maps. *Appl. Comp. Harm. Anal.*, 21, 2006. 2, 4

[7] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *National Academy of Sciences*, 102(21), 2005. 4

[8] A. Del Bue. A factorization approach to structure from motion with shape priors. In *CVPR*, 2008. 1

[9] P. Etyngier, F. Segonne, and R. Keriven. Shape priors using manifold learning techniques. In *ICCV*, 2007. 2

[10] J. Fayad, L. Agapito, and A. Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In *ECCV*, 2010. 1

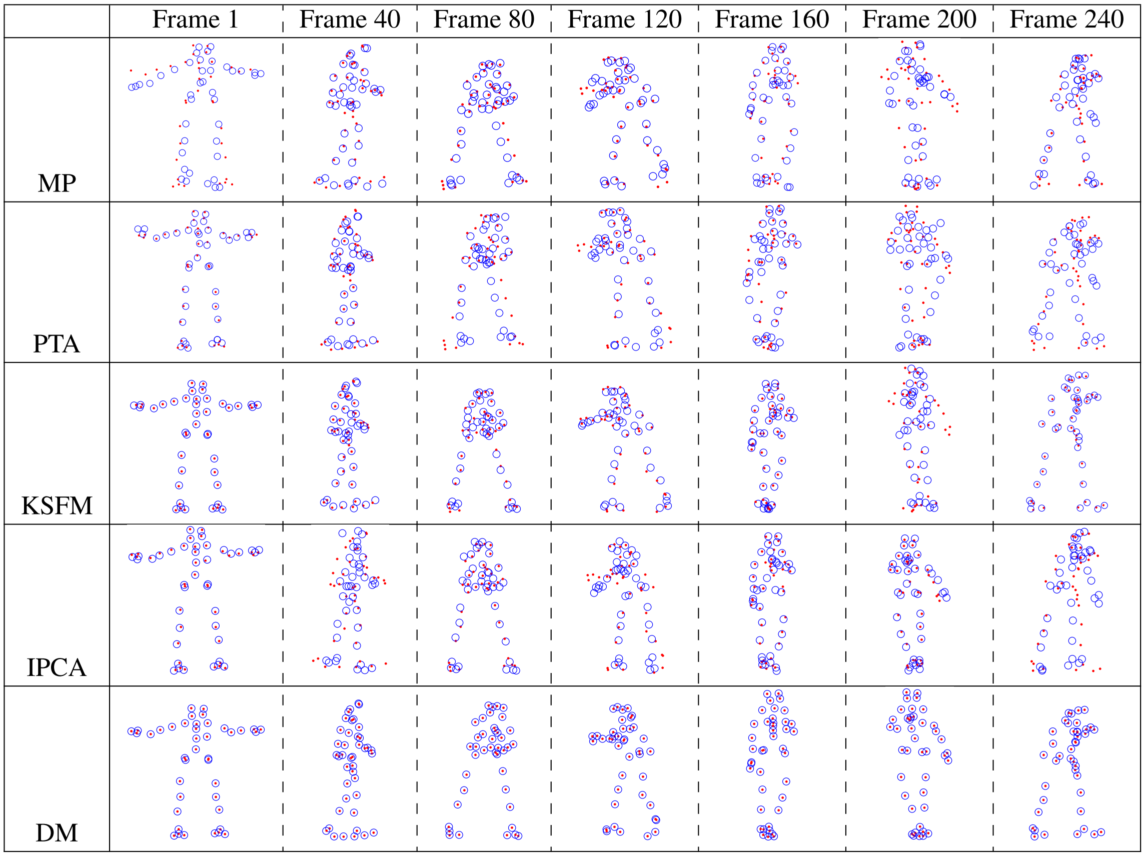

Figure 5. Reconstruction results on the *capoeira* sequence. Reconstructed 3D shapes (blue circles), together with ground truth (red dots) are displayed above.

[11] P. Gotardo and A. Martinez. Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE PAMI*, 33, 2011. 5

[12] P. Gotardo and A. Martinez. Kernel non-rigid structure from motion. In *ICCV*, 2011. 5, 6

[13] B. Matuszewski, W. Quan, L.-K. Shark, A. McLoughlin, C. Lightbody, H. Emsley, and C. Watkins. Hi4d–adsip 3d dynamic facial articulation database. *Image and Vision Computing*, 10:713–727, 2012. 5

[14] L. Mei, J. Liu, A. Hero, and S. Savarese. Robust object pose estimation via statistical manifold modeling. In *ICCV*, 2011. 2

[15] M. Paladini, A. Bue, J. Xavier, M. Stosic, M. Dodig, and L. Agapito. Factorization for non-rigid and articulated structure using metric projections. In *CVPR*, 2009. 1, 2, 5

[16] V. Rabaud and S. Belongie. Re-thinking non-rigid structure from motion. In *CVPR*, 2008. 1

[17] L. Tao, B. Matuszewski, and S. Mein. Non-rigid structure from motion with incremental shape prior. In *ICIP*, 2012. 3, 5

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

[19] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2), 1992. 1

[20] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE PAMI*, 30(5), 2008. 1

[21] P. Tresadern and I. Reid. Articulated structure from motion by factorization. In *CVPR*, 2005. 1

[22] A. Varol, M. Salzmann, P. Fua, and R. Urtasun. A constrained latent variable model. In *CVPR*, 2012. 4, 5

[23] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-free monocular reconstruction of deformable surfaces. In *ICCV*, 2009. 1

[24] R. Wolz, P. Aljabar, J. Hajnal, J. Ltjnen, and D. Rueckert. Manifold learning combining imaging with non-imaging information. In *ISBI*, 2011. 2

[25] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *ECCV*, 2004. 1, 2, 3

[26] J. Yan and M. Pollefeys. A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *IEEE PAMI*, 2008. 1

[27] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d face expression database for facial behavior research. In *AFGR*, 2006. 5