Machine Vision and Applications (2016) 27:801–819
DOI 10.1007/s00138-016-0769-3

ORIGINAL PAPER

Robust deformable shape reconstruction from monocular video with manifold forests

Lili Tao[1,2] · **Bogdan J. Matuszewski**[2]

Received: 28 October 2014 / Revised: 16 March 2016 / Accepted: 20 April 2016 / Published online: 19 May 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Existing approaches to recover structure of 3D deformable objects and camera motion parameters from an uncalibrated images assume the object's shape could be modelled well by a linear subspace. These methods have been proven effective and well suited when the deformations are relatively small, but fail to reconstruct the objects with relatively large deformations. This paper describes a novel approach for 3D non-rigid shape reconstruction, based on manifold decision forest technique. The use of this technique can be justified by noting that a specific type of shape variations might be governed by only a small number of parameters, and therefore can be well represented in a low-dimensional manifold. The key contributions of this work are the use of random decision forests for the shape manifold learning and robust metric for calculation of the re-projection error. The learned manifold defines constraints imposed on the reconstructed shapes. Due to a nonlinear structure of the learned manifold, this approach is more suitable to deal with large and complex object deformations when compared to the linear constraints. The robust metric is applied to reduce the effect of measurement outliers on the quality of the reconstruction. In many practical applications outliers cannot be completely removed and therefore the use of robust techniques is of particular practical interest. The proposed method is validated on 2D points sequences projected from the 3D motion capture data for ground truth comparison and also on real 2D video sequences. Experiments show that the newly proposed method provides better performance compared to previously proposed ones, including the robustness with respect to measurement noise, missing measurements and outliers present in the data.

Keywords Deformable shape reconstruction · Nonlinear manifold learning · Manifold forests · Missing data and outliers

✉ Lili Tao
lili.tao@bristol.ac.uk

[1] Visual Information Laboratory, Department of Computer Science, University of Bristol, Bristol, UK

[2] Robotics and Computer Vision Laboratory, Computer Vision and Machine Learning Research Group, College of Science and Technology, University of Central Lancashire, Preston, UK

1 Introduction

Simultaneous recovery of three-dimensional (3D) sparse feature points representing evolving non-rigid 3D object (simply referred to as 3D structure or shape in the rest of the paper) and a relative camera motion over time from images obtained from a single uncalibrated camera is a challenging, under-constrained problem. The complexity of this problem can be made apparent by realising that reconstruction of the landmarks' 3D positions cannot be uniquely derived based on the knowledge of the locations of their corresponding projections in a single image alone. This can be seen from a simple observation that any 3D point along the projection line, linking the optical centre of the camera and the selected image point, can be equally considered as a valid 3D landmark if no additional information is available. In the stereovision the additional information comes from another image of a static scene taken from a different position, the knowledge of the correspondence between feature points in both images and the known camera motion. In that case the 3D landmarks' reconstruction is obtained by triangulation. For the problem described in this paper not only the camera motion is not known but also, in general, the 3D shape (represented by 3D landmarks) is changing between successive images. This is a hard problem

because, as explained in Sect. 3, the number of unknowns defined by the 3D landmark's coordinates and the camera motion parameters is increasing faster than the size of the measurement data, consisting of the corresponding 2D image points. Therefore, when more measurements are available the harder the problem is as the difference between number of unknown and known is increasing. This is an example of the so-called ill-posed problem. To solve this kind of problems additional information need to be embedded into the problem or/and the problem needs to be reformulated. This process is called regularisation. The fundamental objective of the regularisation is to limit the number of feasible solutions by introducing constrains reflecting our prior knowledge about the problem, e.g. by forcing the solution to have a specific form or belong to a specific subspace or a manifold. Methods addressing deformable shape reconstruction from a monocular video differ essentially by the way such regularisation is introduced.

The methods proposed for dealing with this problem can be categorised, by the type of the regularisation technique used, into three major classes: The low-rank shape models [22], shape trajectory approaches [2,14,15], and template-based methods [17,27,39]. In all these methods the regularisation is achieved by reduction in the problem dimensionality. For example, in the low-rank models, the rank is defined by the number of elementary "shape's building blocks" or basis shapes from which the reconstructed shape is constructed. In principle, higher rank increases the flexibility of the model, leading to possibly more accurate reconstructions, but at the cost of the method increased susceptibility to the observation noise. Low-rank shape was firstly introduced in [6], where the factorisation algorithm is adopted to solve deformable shape reconstruction problem. 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 problem of Non-Rigid Structure from Motion (NRSfM) [1,4,42]. However, the shape bases are different in each sequence, thus need to be estimated for every sequence. Besides, for relatively complex deformable shapes, a large number of bases are required to fit the model. Considering those drawbacks, a trajectory-based algorithm was proposed in [2] exploiting 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 problem. Template-based reconstruction is an alternative method which usually relies on a known reference frame and works well especially for inextensible surfaces. Since this is still an ill-conditioned problem [21], the most commonly used constraints in the reconstruction involve preserving Geodesic distances as the surface deforms and thus regularise the problem by solving either convex optimisation problem [7,29] or in closed-form sets of quadratic equations [19,30]. The existing 3D reconstruction technologies have been successfully used in many different areas, ranging from medical imaging and biometrics to computer gaming and film production. A variant of that methodology, mostly dealing with static scenes, called simultaneous localisation and mapping (SLAM) is used for robot navigation where reconstruction is used to build a 3D representation of the environment and the camera pose estimation is equivalent to robot positioning in that environment [12,20].

However most of the existing approaches are restricted by the fact that they try to explain the complex deformations using a linear model. To move away from the linear representations of deformable shapes, recent methods have integrated the manifold learning algorithm [28] to regularise the shape reconstruction problem by constraining the shapes as to be well represented by the learned manifold. In simple terms, a manifold can be thought of as a smooth surface/curve embedded in a relevant multi-dimensional space. The advantage of using a manifold constraint, if such constraint adequately represents properties of the reconstructed objects, is a compact representation leading to robust regularisation. If for example, for a hypothetical problem, it is known that a valid shape could be accurately represented by points on a curve, the manifold method would effectively have the dimensionality of one, whereas the linear method would still require, possibly high-dimensional subspaces, to accurately model all turns and twists of the curve. In this case, there is also no guarantee that the reconstructed object would belong to that curve. Rabaud and Belongie firstly claimed in their work [24] that the possible 3D shapes of an object may not lie in a linear low-dimensional manifold. Based on a low-rank shape model, the work assumed that shapes lie on a d-dimensional manifold, and every neighbourhood of shape approximately lies on a d-dimensional linear subspace. In order to minimise the cost function which consists of the reprojection error and smoothing terms. The initial values are calculated by Rigid Shape Chain also introduced in [24], with sequences clustered into several rigid shapes. After initialisation, the optimisation on the shapes is performed using two criteria: the cost function and the shape manifold dimensionality constraint for which the locally smooth manifold learning technique has been used. Later they proposed a method focusing on a globally linear manifold and use shape embedding as initialisation [25]. Similarly, the work in [45] attempts to learn the 3D reconstruction of human motion with an assumption that human poses lie in a union of subspaces.

Other manifold-based methods were inspired by the basis trajectory model. Gotardo and Martinez demonstrated the "kernel trick" which used for nonlinear dimensionality reduction [31] can also be applied to standard NRSfM problem [15]. Recently Hamsici et al. [16] modelled the shape

coefficients in a manifold feature space. This method has ability to recover shapes from a newly observed image. The mapping is learned from the corresponding 2D measurement data of upcoming reconstructed shapes, rather than a fixed set of trajectory bases. They introduced rotation-invariant kernels (RIK) to provide similarity measure for two 3D shapes based on their 2D projections. The problem still remains though that the 2D observations can be completely different when the images are taken from different view angles. Similarly because of different depths, similar 2D images may not represent similar 3D shapes. In comparison, [15] defines a nonlinear model while [16] models 3D shapes in a linear space; [15] uses point trajectory bases as input data for building a kernel function while [16] uses shapes directly from 2D images.

The problem becomes more difficult when the observations are incomplete. Most algorithms assume that all feature points are detected in all images. This is unlikely to happen in practice as some of the feature points will not be detected in all images. This could be because of the feature point detection problems or because some parts of the 3D object may not be visible from all the camera positions, which means some of the entries in the measurement matrix may be unknown. The methods addressing this problem can be divided into three categories: imputation, alternation and nonlinear optimisation. The problem was firstly addressed by filling the missing entries using complete subset of the data in rigid [36] and non-rigid reconstruction problem [44]. These imputation algorithms are simple but cannot handle well real data, which often tend to be very noisy. To overcome this, the alternation algorithms solve the problem based on a closed-form solution using a rank constraint imposed on the measurement matrix without estimating the missing values in advance [18,23]. The idea is to iteratively update the motion and shape in terms of observed measurements. Another commonly used method for addressing the missing data problem is to employ nonlinear minimisation of suitably designed cost function. The measurements can be gradually refined to produce jointly optimal 3D structures and camera motion. This is known as bundle adjustment which has been studied for many years [38]. The major benefit of this method is that the additional constraints can be effectively included in the cost function, though the inherently high number of degrees of freedom may lead to failure of obtaining reliable reconstruction results.

Despite many years of research, one significant problem still remains, the most existing approaches cannot cope well with outliers. Earlier work suggested that outliers have to be removed before doing any further processing [8]. Vidal et al. presented a geometrical algorithm for 3D motion segmentation which dealt with the data by using RANSAC to detect the outliers [40]. The outliers are usually caused by matching errors between two frames, and this may severely affect the trajectory methods because the trajectories passing through the feature points do not belong to any of the trajectory spaces. Additionally, most methods have been using least-square estimator, which is well known not to be robust to outliers. Simply removing the outliers in advance may not be feasible in practice, especially when real-time processing is required. Developing robust estimations is necessary for obtaining reliable solutions. The work in [13] presented a rank-constrained factorisation algorithm that effectively calculates a low-rank approximation of a measurement matrix in the presence of the data outliers. The problem is solved by replacing squared residual error function by L_1 norm which is often used to reduce sensitivity of the model.

Figure 1 provides an overview of the proposed reconstruction system. In this paper, it is assumed that the feature points have been detected in the images and the 2D point correspondences are provided as input to the reconstruction algorithms. Although the paper is focused on the solution of the reconstruction problem, when both points and their correspondences are given, the imperfection in both point position and correspondence estimation are indirectly addressed by introduction of a robust metric. The reminder of the paper is organised as follows: Sect. 2 highlights the novel contributions of the paper. Section 3 describes the problem of deformable shape reconstruction and presents the notation used throughout the paper. Section 4 introduces the generic concept of the manifold forest. Subsequently Sect. 5 provides detailed description of the proposed manifold forest implementation aiding the shape reconstruction. In both these sections, an effort has been made to explain advantages of using manifold forest in shape reconstruction. Section 6 describes a novel robust approach dealing with missing data and outliers. Finally, Sects. 7 and 8 present experimental results and conclusions.

2 Novelty

Although the problem of deformable shape and motion recovery has been studied for many years, one of the severe limitations of the most existing approaches 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 nonlinear deformable objects with large complex deformations, such as deformable cloth. In this case, most existing methods based on linear space are no longer suitable.

Contrary to other techniques using manifold in the shape reconstruction, our manifold is learned based on the 3D shapes rather than on 2D observations. The proposed implementation is based on the manifold forest method described in [11]. The main advantage of using manifold forest as com-

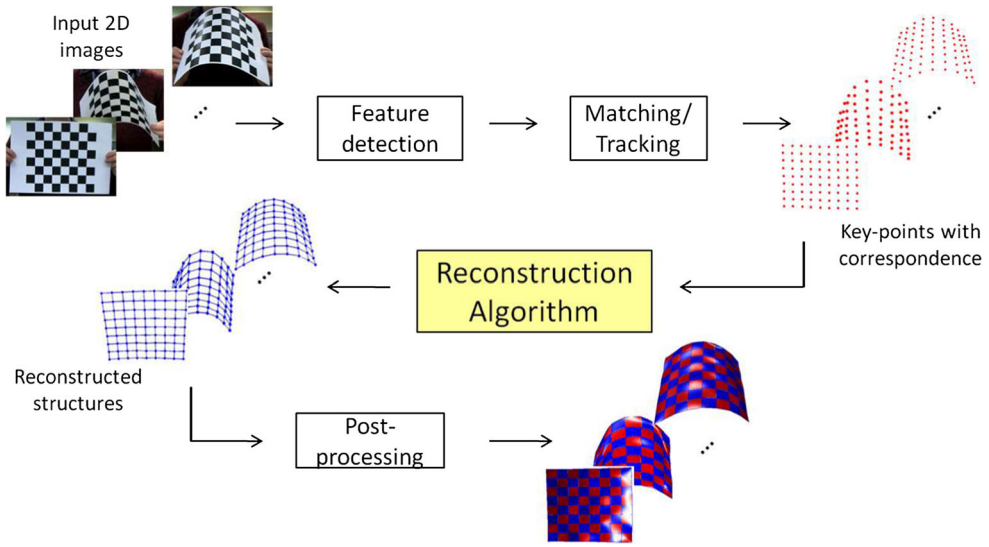

Fig. 1 The pipeline of a complete 3D objects reconstruction system

pared for example to standard diffusion maps [9] is the fact that in the manifold forest the neighbourhood topology is learned from the forests data clustering rather than being defined by the Euclidean distance. To the best of authors' knowledge, random forests technique has never been applied in the context of non-rigid shape reconstruction using. The idea of integrating nonlinear manifold-based approaches into 3D deformable reconstruction was firstly introduced by the authors in [34], where the shape prior is introduced in the form of the diffusion maps. In that work, the structure of data is estimated using Euclidean distances between pairs of data items, whereas the method proposed in this paper learns the structure from the data, based on random forests techniques.

This paper updates and extends the work in [33] with the following four main differences: (a) The method presented in this paper has an additional step in the algorithm, solving the problem when some elements of the measurement matrix are missing; (b) Considering the majority of algorithms are based on minimising squared residual of an error function which makes them sensitive to outliers, another improvement is to reduce the effect of outliers by replacing the L_2 estimator by robust M-estimator [26]; (c) A modification of the method is described when only a relatively small number of training shapes is available. This was firstly introduced by the authors in [32] but without random forests manifold learning technique; (d) More comprehensive set of experiments is presented in the experimental section.

3 Basic formulation

Throughout this paper, vectors and matrices are denoted as lower- and upper-case bold letters, whereas sets are represented by calligraphic letters. We assume that 2D points (features) are obtained from F frames under an orthographic camera projection model. The problem consists of the recovery of shapes $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots \mathbf{S}_F\}$ and camera rotations $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_F\}$ from 2D observations $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_F\}$, thus can be formulated as the 2D reprojection error minimisation problem:

$$\arg\min_{\mathcal{R},\mathcal{S}} \sum_{t=1}^{F} \|\mathbf{Y}_t - \mathbf{P} \cdot \mathbf{R}_t \cdot \mathbf{S}_t\|^2 \tag{1}$$

where, $\mathbf{P}$ represents orthographic camera projection matrix, $\mathbf{Y}_t$ is a $2 \times P$ matrix of detected 2D feature points' coordinates in the t^{th} image, and $\mathbf{S}_t \in \mathbb{R}^{3\times P}$ contains coordinates of P 3D points describing shape represented in the t^{th} frame, $\|\cdot\|$ indicates Frobenius norm. The camera translation has been eliminated by expressing 2D observations $\mathcal{Y}$ with respect to the data points centroid calculated in each observed image.

The goal is to recover camera orientations $\mathcal{R}$ and the concatenated time-varying shapes $\mathcal{S}$, based only on the 2D measurement $\mathcal{Y}$. It is an under-constrained problem since the shape and motion are both changing with time. The number of unknown variables $(3F + 3FP)$ is higher than the size of observed data $(2FP)$. To deal with this, a low-rank shape model has proved to be successful as shown in [37], where the shape $\mathbf{S}_t$ is represented as a linear combination of K unknown but fixed basis shapes $\mathcal{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots \mathbf{B}_K\}$:

$$\mathbf{S}_t = \sum_{l=1}^{K} \theta_{tl} \mathbf{B}_l \tag{2}$$

where $K \ll F, P$. The deformation coefficients θ_{tl} are adjustable over time t. This low-rank shape model can be

obtained by performing singular value decomposition (SVD) of the measurement matrix $\mathbf{Y}$, for which the measurement matrix can be decomposed and represented by pose $\mathcal{R}$, basis shapes $\mathcal{B}$ and time-varying coefficients θ_{tl}, and it can be rearranged as:

$$\mathrm{Y} = \begin{bmatrix} \theta_{11}\mathbf{R}_1 \cdots \theta_{1K}\mathbf{R}_1 \\ \vdots \quad \ddots \quad \vdots \\ \theta_{F1}\mathbf{R}_F \cdot \theta_{FK}\mathbf{R}_F \end{bmatrix} \begin{bmatrix} -\mathbf{B}_1 - \\ \vdots \\ -\mathbf{B}_K - \end{bmatrix} = \mathbf{MB} \tag{3}$$

Since basis shapes $\mathbf{B} \in \mathbb{R}^{3K \times P}$ and $\mathbf{M} \in \mathbb{R}^{2F \times 3K}$, the rank of the measurement matrix $\mathbf{Y}$ is $3K$ at most, in the absence of noise. The matrices $\mathbf{M}$ and $\mathbf{B}$ are computed by factorising the measurement matrix $\mathbf{Y}$. The solution is not unique and is defined up to an ambiguity matrix $\mathbf{Q} \in \mathbb{R}^{3K \times 3K}$. According to [41], 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.

Our model departs from the linear shape model. The shape basis in the proposed method are selected from the learned shape manifold. The shape $\mathbf{S}_t$ is represented as a linear combination of $n+1$ (where n is the dimension of the manifold introduced in Sect. 4.1) basis shapes $\mathbf{B}_{tl}$, adaptively selected from the learned manifold: $\mathbf{S}_t = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{B}_{tl}$. 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 for each frame. Such approach adds an extra flexibility to the reconstruction process allowing a better adaptation of the method to the temporal shape changes. Although it may seem that this increases the number of parameters in the model, it should be clarified 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 $\mathbf{S}_t$.

4 Manifold forests

In this paper, the manifold forests are constructed upon diffusion maps with the neighbourhood topology learned through random forest data clustering. It generates efficient representations of complex geometric structures even when the observed samples are non-uniformly distributed. This section gives an introduction to diffusion maps and randomised decision forests first, and then describes the application of random forests in learning diffusion map manifolds.

4.1 Diffusion maps

In many problems, data are difficult to represent or analysed due to their high-dimensional structure. However, in same cases, complex data might be governed by a small number of parameters. The goal of the manifold learning is to find the embedding function, mapping the data set $\mathcal{X}$ form a high, $N = 3P$-dimensional space to a reduced, n-dimensional space. The diffusion map is a graph-based nonlinear technique with quasi-isometric mapping from original shape space onto a lower-dimensional diffusion space. Unlike linear methods, nonlinear approaches are able to handle a wider range of data variability, preserving local structures at the same time. The problem with linear manifold methods is that the input data may have complex nonlinear dependencies and preserving global or indeed local structures in the data may not be possible utilising linear projections.

Assuming $\mathcal{X}$ is a dataset with M samples, the goal of dimensionality reduction problems is to find a mapping of the data $\mathcal{X} = \{\mathbf{X}_1 \ldots \mathbf{X}_M\}$ given in high N-dimensional space to data $\{\mathbf{x}_1 \ldots \mathbf{x}_M\}$ given in a reduced n-dimensional space. A mapping is defined by: $\Psi : \mathbf{X} \mapsto \Psi(\mathbf{X}) = (\Psi_1(\mathbf{X}), \ldots, \Psi_n(\mathbf{X})) = \mathbf{x}$, where $\mathbf{X} \in \mathbb{R}^N$, $n \ll N$.

Given a set of shapes $\mathbf{X}_1 \ldots \mathbf{X}_M \in \mathcal{M}$, where $\mathcal{M}$ is the manifold embedded in $\mathbb{R}^N$, Euclidean distance for each pair of shapes $\left\|\mathbf{X}_i - \mathbf{X}_j\right\|^2$ is calculated to build an adjacency graph. The entries of the affinity matrix $\mathbf{W} = [W_{ij}]$, $i, j \in 1 \ldots M$ define the weighted similarity graph for all connected vertexes. W_{ij} could be calculated in number of different ways, often Gaussian kernel is used. In that case: $W_{ij} = \exp(-\left\|\mathbf{X}_i - \mathbf{X}_j\right\|^2/2\delta)$, where δ is a kernel scale. k-nearest neighbour (kNN) sparsification scheme can also be applied, retaining k edges for each graph vertex and removing other connections to avoid outliers.

Coifman et al. presented a justification behind using normalised graph Laplacian [9] by connecting them to diffusion distance. Each entry of the diffusion operator $\mathbf{G}$ is constructed as $G(\mathrm{X}_i, \mathrm{X}_j) = W'_{ij}/\Upsilon_{ii}$ with $\Upsilon_{ii} = \sum_j W'_{ij}$. $\mathbf{W}'$ is a renormalised version of the affinity matrix $\mathbf{W}$ using an anisotropic normalised graph Laplacian, such that $W'_{ij} = W_{ij}/q_i q_j$ with $q_i = \sum_j W_{ij}$, $q_j = \sum_i W_{ji}$. The convergence of optimal embedding Ψ for diffusion maps is proven in [9] and is found via eigenvectors φ and their corresponding n biggest eigenvalues λ of the operator $\mathbf{G}$, such that $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_n$,

$$\Psi : \mathrm{X}_i \mapsto [\lambda_1\varphi_1(\mathrm{X}_i), \ldots, \lambda_n\varphi_n(\mathrm{X}_i)]^{\mathrm{T}} \tag{4}$$

where $\varphi_j(\mathrm{X}_i)$ represents ith element (corresponding to the ith training sample X_i) of the jth eigenvector of $\mathbf{G}$.

4.2 Randomised decision forest

Random forest [10] has become a popular method given its capability to handle high-dimensional data, avoid overfitting, and enabling simple parallel implementation. The decision trees in our method are built by making decisions in each node of the tree based on randomly selected features. A random decision forest is an ensemble of such decision trees. The trees are different and independent from each other. Although other choices are possible, this paper is focused only on the binary decision forest.

Given a set of training data $\mathcal{X}$ with M samples: $\mathbf{X}_i \in \mathcal{X}, i = 1 \ldots M$, where each sample contains $3P$ features. The trees are randomised, by randomly selecting a single feature at each internal node. The decision function at the internal node is used to decide whether the data $\mathbf{X}_i$ reaching that node should be assigned to its left or right child node. The threshold α_m of the decision function at node m is selected as result of the maximisation of the information gain:

$$\alpha_m^* = \arg\max_{\alpha_m} I_m \tag{5}$$

with the generic information gain I_m defined as:

$$I_m = H\left(\mathcal{X}_m\right) - \sum_{i\in\{L,R\}} \frac{\left|\mathcal{X}_m^i\right|}{\left|\mathcal{X}_m\right|} H\left(\mathcal{X}_m^i\right) \tag{6}$$

where $|\cdot|$ indicates a cardinality for the dataset. $\mathcal{X}_m$ denotes the training data $\mathcal{X}$ reaching node m. $\mathcal{X}_m^L, \mathcal{X}_m^R$ are the subsets assigned to the left and right child nodes of node m. In this paper it is assumed that data is adequately represented by the Gaussian distribution [11]. In that case the differential entropy $H\left(\mathcal{X}_m\right)$ can be calculated analytically as:

$$H\left(\mathcal{X}_m\right) = \frac{1}{2}\ln\left((2\pi e)^N \left|\Lambda(\mathcal{X}_m)\right|\right) \tag{7}$$

where $|\Lambda(\mathcal{X}_m)|$ is the determinant of the covariance matrix estimated from the $\mathcal{X}_m$ training data. By substituting (7) into (6), the information gain can be rewritten as:

$$I_m \propto \frac{1}{2}\ln\left(\left|\Lambda(\mathcal{X}_m)\right|\right) - \frac{1}{2}\sum_{i\in\{L,R\}} \frac{\left|\mathcal{X}_m^i\right|}{\left|\mathcal{X}_m\right|} \ln\left(\left|\Lambda\left(\mathcal{X}_m^i\right)\right|\right) \tag{8}$$

The trees are trained until the number of samples in a leaf is less than the pre-specified limit or the depth of the tree has exceeded the pre-defined depth.

Once the random forest has been trained, the new sample can be simply put through each tree. Depending on the result of the decision function at each internal node, the new data is sent to the left or right child node until it arrives at a leaf. The samples ending up in the same leaf are likely to be statistically similar and are expected to represent the same neighbourhood of the manifold. As such similarity measure is statistical in nature, thus the results is averaged over many decision trees. If the samples end up in the same leaf for the majority of the trees they are considered to be drawn from the similar location on the manifold.

4.3 Forest model for manifold learning

In the proposed method, the affinity model in manifold learning is built by applying random forest clustering. The data partition is defined based on the leaf node $l(\cdot)$ the input data $\mathbf{X}_i$ would reach. The entries of the affinity matrix $\mathbf{W}^t$ for tree t are calculated as,

$$W_{ij}^t = e^{-L^t(\mathbf{X}_i,\mathbf{X}_j)}, i, j \in 1 \ldots M \tag{9}$$

where the distance L can be obtained using different models. For instance, Gaussian or binary affinity for the data ending up in the same leaf node are defined as follows [11]:
Gaussian affinity model

$$L^t\left(\mathbf{X}_i, \mathbf{X}_j\right) = \begin{cases} \frac{\|\mathbf{X}_i - \mathbf{X}_j\|^2}{2\delta} & l(\mathbf{X}_i) = l(\mathbf{X}_j) \\ \infty & \text{otherwise} \\ c & \end{cases} \tag{10}$$

Binary affinity model

$$L^t\left(\mathbf{X}_i, \mathbf{X}_j\right) = \begin{cases} 0 & l(\mathbf{X}_i) = l(\mathbf{X}_j) \\ \infty & \text{otherwise} \end{cases} \tag{11}$$

In this paper binary affinity model is used as it is simple and efficient, and can be considered to be parameter free. In a forest, each randomly trained clustering tree produces a disjoint partition of the data, and they are independent and different with respect to each other. A graphic representation of building affinity matrix using binary model is shown in Fig. 2. However, as affinity matrix computed for a single randomly trained tree is not representative of the correct similarities of the data, the ensemble of T trees is used to calculate a much smoother affinity matrix $\mathbf{W}$. The affinity matrix representing a forest of size T is calculated by averaging over all affinity matrices from each single tree:

$$\mathbf{W} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{W}^t \tag{12}$$

The comparison results of diffusion maps and manifold forests are shown in Fig. 3. Figure 3a illustrates synthetically generated data of a 3D parabola surface given by the equation $f(x, y) = \frac{x^2+y^2}{2}$. Its corresponding embeddings in 2D reduced space using diffusion maps and random forest manifold with the Gaussian affinity model are shown in

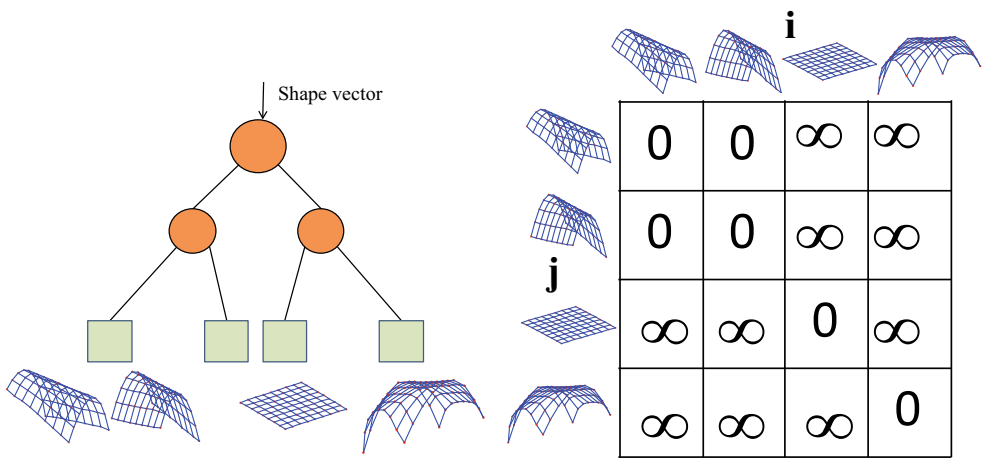

Fig. 2 A graphic representation of building affinity matrix using binary model. *Left* Decision tree model. Shapes are stored in different leaves. *Right* Distance matrix L corresponding to that tree

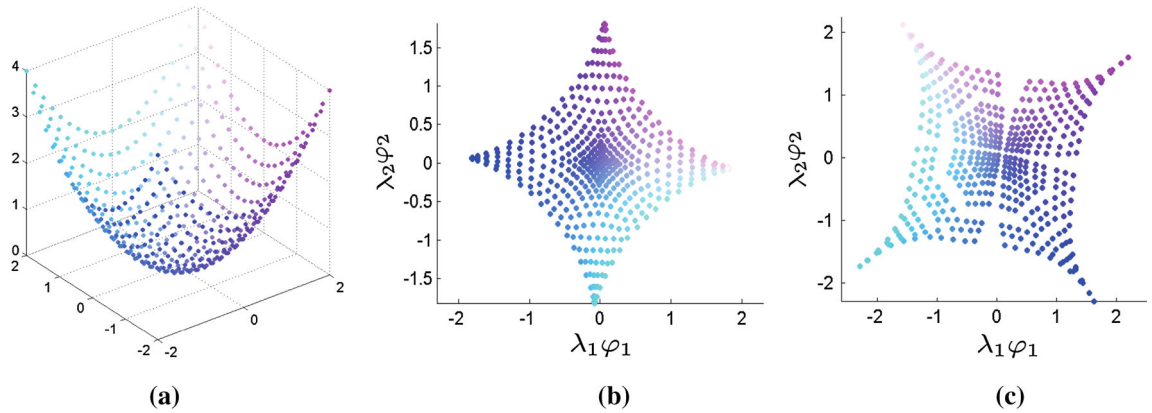

Fig. 3 Simulated results for nonlinear dimensionality reduction. **a** Input 3D points from parabola surface. **b** Nonlinear mapping from the original 3D space to the 2D reduced space based on diffusion maps. **c** Embedding based on manifold forests with Gaussian affinity model

Fig. 3b, c, respectively. It can be seen that the embedding obtained based on the manifold forest achieves somewhat better representation of the data in the lower-dimensional space. This is specifically well illustrated when comparing the distribution of the embedded points representing the base and the rim of the parabola, as these points seem to be more equally distributed when random forest manifold embedding is used.

Moreover, in diffusion maps, the method of sparsifying an affinity matrix is based on retaining the k-nearest neighbourhoods among the data in the graph. However, this may cause two problems: Choosing appropriate number of nearest points is not easy since it depends on the data structure, and it can create connected edges in the graph for the points which may be outliers. Forming affinity matrix using random forest technique would efficiently solve such problems as the points are only connected if they are in the same cluster.

Figure 4 illustrates the embedding of shapes using diffusion maps and manifold forests from *cardboard* data [39] together with corresponding representative shapes extracted from 1000 training samples. The embedding results obtained by applying manifold forests seem to be more evenly distributed than points embedded using diffusion maps, especially for the shapes located along the rim of the manifold.

One of the main advantages of using random forest manifold is that it implicitly addresses one of the main difficulties in the manifold learning, namely it optimally defines the data neighbourhood structure. The optimality criterion is defined through the splitting decision used in the nodes of the trees, in this paper optimality is defined through maximisation of the information gain. Additionally, random forest implicitly selects optimal features for the data clustering, where the optimum criterion is defined by the node decision rule. In the case of the random forest implemented in this paper the maximum information gain decision rule is used which favours features for which data splitting at a node leads to compact class distributions. This may explain a better performance of the random forest manifold when compared to the original diffusion maps with the neighbourhoods defined by the k-nearest neighbours. This can be seen in Fig. 4 (cardboard

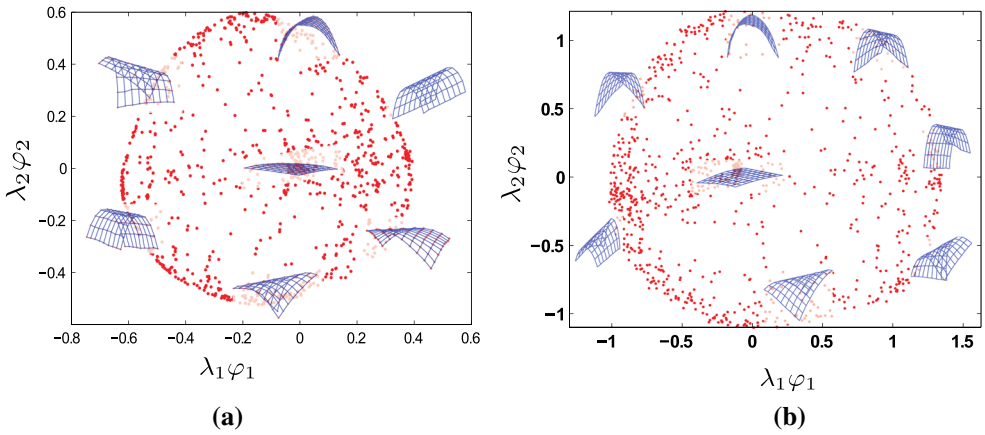

Fig. 4 The reduced space of *cardboard* dataset. **a** The embedding of shapes using diffusion maps only. **b** The embedding of shape using manifold forests

dataset) where the mapping to the lower-dimensional space, using the random forest manifold, produces visibly more uniformly distributed points.

The uniformity of the data points distribution can be also measured quantitatively by estimating the entropy characterising the data distribution in the parametric space. To start, for the parametric space, a histogram representing areas of all the Delaunay triangles for the training dataset is calculated. For the well-distributed data, this histogram is expected to be compact as all triangles would have similar area. The compactness of the histogram is measured here using entropy, with entropy equal to 6.85 and 5.08 for the data embedding in 2D space using diffusion maps and manifold forests, respectively. Both, the subjective and objective measures show that using the random forest manifold, produces more uniformly distributed points and therefore could supports more accurate data interpolation.

5 Random forests in 3D reconstruction

Once the manifold has been built from the training dataset, the shape reconstruction can be obtained from the learned shape manifold and the observed 2D measurements. In this section, an overview of the proposed manifold-based reconstruction algorithm is given first, followed by a description of out-of-sample and inverse mapping problems.

As known from [41], enforcing only the rotation constraints cannot guarantee the unique solution for the camera motion and the basis shapes. To solve this, the designed shape prior can help to attract a shape towards the manifold and therefore avoid incorrect reconstructions.

A summary of the algorithm for the shape recovery of a non-rigid object and estimation of camera motion is given in Algorithm 1. Initial shapes $\mathcal{S}'$ and camera motion $\mathcal{R}'$ are estimated by running a few iterations of the optimisation process using linear basis shapes model [35]. The method is not significantly sensitive to the initial solution as it can iteratively update the shapes by projecting them on the learned manifold until convergence. For each initial shape, Nyström extension [3] is used for embedding these new samples into the reduced space. Intuitively, if the points in the reduced space are relatively close, the corresponding shapes in the high-dimensional space should represent similar shapes. Based on this observation, the reconstructed shape at each frame can be represented as weighted sum of $n + 1$ basis shapes from the learned manifold. The coefficients of corresponding basis shapes are calculated as barycentric coordinates of $n+1$ neighbouring points from Delaunay triangulation of the training dataset. 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 (see Eq. 15). However, the quality of the reconstruction depends 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 predefined 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.

5.1 Mapping out-of-sample points

The manifold forests method briefly described in Sect. 4 is used to find a meaningful representation of the data, but the mapping Ψ is only able to provide an embedding for the data present in the given training set. In our algorithm, it is necessary to calculate embedding for shapes which are not presented in the training set. Suppose a new shape $\mathbf{S}_t \in \mathbb{R}^N$ becomes available after the manifold had been learned, instead of re-learning the manifold, as it is too com-

Algorithm 1 Outline of manifold forest-based reconstruction

Input: Stream of 2D observations, manifold forest Ψ of training dataset $\mathcal{X}$ (Sect. 4.3)

Output: 3D deformable shapes $\mathcal{S}$ and camera motion $\mathcal{R}$ for each frame.

1: Initialisation of estimating initial shapes $\mathcal{S}'$ and camera motion $\mathcal{R}'$.
2: **while** $(\|r\| > r_T)$ *or* $\left(\|r_t\| - \|r_{t-1}\| > 10^{-3}\right)$ **do**
3: Shape projection onto manifold (shape Embedding) (Sect. 5.1)
4: Find $n+1$ closest points $\mathbf{b}_l, l = 1 \ldots n+1$ in low-dimensional space, where n is the dimensionality of the reduced space.
5: Shape update (Sect. 5.2)
6: Nonlinear optimisation by minimising 2D measurement error and shape smooth term to obtain updated shapes $\mathbf{S}_t$ and camera motion $\mathbf{R}_t, t = 1 \ldots F$ (Sect. 5.3)
7: **end while**

putationally expensive, an efficient way is to interpolate the shape based on the already learned manifold. For each new shape, such embedding is calculated based on the Nyström extension. Knowing that for every sample in the training dataset:

$$\forall \mathrm{X}_i \in \mathcal{X}, \sum_{\mathrm{X}_j \in \mathcal{X}} G(\mathrm{X}_i, \mathrm{X}_j)\varphi_k(\mathrm{X}_j) = \lambda_k \varphi_k(\mathrm{X}_i), k = 1 \ldots n \tag{13}$$

Having a shape $\mathbf{S}_t$, not present in the training set $\mathcal{X}$, an embedding

$$\mathbf{S}_t \mapsto \left(\hat{\Psi}_1(\mathbf{S}_t), \ldots, \hat{\Psi}_n(\mathbf{S}_t)\right)$$

of this new shape is calculated from:

$$\hat{\Psi}_k(\mathbf{S}_t) = \sum_{\mathbf{X}_j \in \mathcal{X}} G(\mathbf{S}_t, \mathbf{X}_j)\varphi_k(\mathbf{X}_j), k \in 1 \ldots n \tag{14}$$

where $G(\mathbf{S}_t, \mathbf{X}_j)$ is calculated in the same way as the diffusion operator (see Sect. 4.1). The distance between unseen sample and the training samples is calculated using the binary affinity model, (11). In principle these calculations still could be expensive as the summation in (14) is done over all training samples. Random forest approach provides very effective way for implementing the out-of-sample mapping. For this the unseen sample is put through the forest, and subsequently only the training samples from leafs where the unseen sample ended up are used in the sum in (14).

5.2 Inverse mapping for shape update

Given a point $\mathbf{s}_t \in \mathbb{R}^n$ in the reduced space, finding its inverse mapping $\mathbf{S}_t = \Psi^{-1}(\mathbf{s}_t)$ from the reduced space back to the input space is a typical pre-image problem. As claimed in [3], the exact pre-image might not exist if the shape $\mathbf{S}_t$ is not included in the training set. However, according to the properties of diffusion maps, if the points in the reduced space are relatively close, the corresponding shapes in high-dimensional space should represent similar shapes since they have small diffusion distances. Based on this, the point $\mathbf{s}_t$ can be approximated as a linear combination of its weighted neighbouring points in reduced space, such that $\mathbf{s}_t = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{x}_{tl}$, where $\mathbf{x}_{tl}$ is the l^{th} neighbouring points of $\mathbf{s}_t$ obtained by computing Delaunay triangulation [5] on the training dataset, and the weights θ_{tl} are computed as the barycentric coordinates of $\mathbf{s}_t$. Once the weights are estimated, the shape $\mathbf{S}_t$ can be calculated as well based on a set of weighted training samples $\mathbf{S}_t = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{X}_{tl}$, where the training samples $\mathbf{X}_{tl}$ are the pre-images of $\mathbf{x}_{tl}$, and are equivalent to the basis shapes in (2).

5.3 Nonlinear refinement

The cost function $E()$ to be minimised consists of the reprojection error, shape smoothing term and rotation constraint,

$$\begin{aligned} E(\{\mathbf{R}_t\}, \{\theta_{tl}\}) = &\sum_{t=1}^{F} \|\mathbf{Y}_t - \mathbf{P} \cdot \mathbf{R}_t \cdot \mathbf{S}_t\|^2 + \gamma_{\mathbf{S}} \sum_{t=2}^{F} \|\mathbf{S}_t - \mathbf{S}_{t-1}\|^2 \\ &+ \gamma_{\mathbf{R}} \sum_{t=1}^{F} \varepsilon_{rot} \\ &\text{with} \sum_{l=1}^{n+1} \theta_{tl} = 1, \quad 0 \le \theta_t \le 1 \end{aligned} \tag{15}$$

where $\varepsilon_{rot} = \left\|\mathbf{R}_t \cdot \mathbf{R}_t{}^{\mathrm{T}} - \mathbf{I}\right\|^2$ penalises deviation from orthonomality of all $\mathbf{R}_t$. $\gamma_{\mathbf{S}}$ and $\gamma_{\mathbf{R}}$ are regularisation constants, and $\mathbf{S}_t$ is expressed as linear combination of weighted neighbouring training shapes $\mathbf{X}_{tl}$ (see Sect. 5.2). A nonlinear optimisation using Levenberg–Marquardt algorithm is applied to minimise the cost function with analytically calculated Jacobian.

However, the underlying problem is that the quality of the optimisation result strongly depends on the accuracy of initial shapes. To avoid this, we update the basis shapes in each iteration until 2D measurement error is less than the defined threshold (10^{-3} in our case) and the error between two adjacent frames is relatively small. This effectively means that for any given t the basis shapes $\{X_{tl}\}$ can change during the iteration minimising $E()$.

5.4 Nonlinear refinement with reduced training set

Building a dense manifold requires large number of training samples. In practice it may be difficult to obtain such dense training set. To address this problem, this section briefly describes a variant of the proposed method which can handle situation when only limited training data is provided. The idea is to modify the cost function with additional term.

In this case the basis shapes will be estimated, rather than matched to the local training samples. The similar method was introduced in [32], but manifold was learned without using the random forest.

The main idea is to partition a set of estimated shapes into K clusters, in which the shapes have similar structure, with each shape cluster denoted by $\mathcal{T}_i, i \in 1 \ldots K$. The clusters are obtained by performing the Delaunay triangulation in the reduced space. The points in the reduced space belong to the same Delaunay triangle (i.e. cluster), can be modelled in the same linear manifold embedded in $\mathbb{R}^N$, and therefore all corresponding reconstructed shapes (represented by that cluster) can be approximated by a linear combination of the same set of unknown but fixed basis shapes. Thus all the shapes in the cluster i can be represented as $\mathbf{S}_t = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{B}_l^i, \forall t \in \mathcal{T}_i$, where a set of basis shapes $\mathcal{B}^i = \left\{\mathbf{B}_1^i \ldots \mathbf{B}_{n+1}^i\right\}$ is spanning the tangent linear subspace representing all the shapes from the cluster i.

The parameters $\theta_{tl}, \mathbf{B}_l^i$ and $\mathbf{R}_t$ are optimised simultaneously by minimising the following modified cost function,

$$E(\{\mathbf{R}_t\}, \{\mathbf{B}_l^i\}, \{\theta_{tl}\}) = \sum_{t\in\mathcal{T}_i} \left\| \mathbf{Y}_t - \mathbf{P}\cdot\mathbf{R}_t \sum_{l=1}^{n+1} \theta_{tl}\mathbf{B}_l^i \right\|^2 + \gamma_B \varepsilon_{bs}^i + \gamma_R \sum_{t\in\mathcal{T}_i} \varepsilon_{rot} \tag{16}$$

where the additional constraint applied to the i^{th} set of basis shapes is,

$$\varepsilon_{bs}^i = \sum_{l=1}^{n+1} \left\| \mathbf{B}_l^i - \mathbf{X}_l^i \right\|^2, \mathbf{X}_l^i \in \mathcal{X} \tag{17}$$

Figure 5 shows the embedding using reduced number of training samples. 40 shapes are randomly selected from the *cardboard* dataset. In the figure, the result of the Delaunay triangulations is visualised by blue line segments. One frame from the testing sequence is chosen to demonstrate how the shape is updated in each iteration. Yellow lines illustrate the trajectory of the embedding of the shape moving through different triangles. Red and black dots represent the embedding of reconstructed and the ground truth shape respectively.

6 Reconstruction with missing data and outliers

The fact that the measurements can be affected by outliers and may not be complete, means that the reconstruction algorithm must be robust to measurement data deficiencies in real application. The algorithm described so far assumes the measurements $\mathcal{Y}$ are complete, all the feature points are identified in all the images in the sequence. In real sequences, some

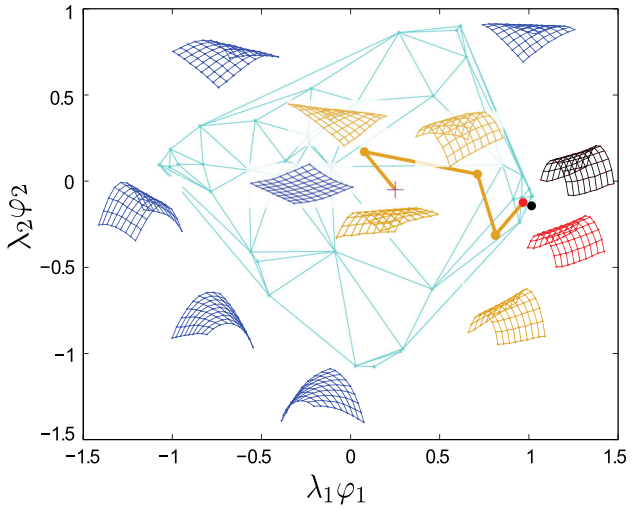

Fig. 5 The embedding of the reduced number of samples from the *cardboard* dataset with corresponding shapes

of the points cannot be detected in all the images due to the occlusions, feature detection problems, or tracking failures and therefore acquiring complete set of measurements is unlikely, and the measurement may be affected by outliers due to errors in the correspondence search.

Two methods are introduced in this section, for solving missing data and outliers problems, respectively. A nonlinear approach for missing data is presented first. This method efficiently solves the problem by simultaneously optimising the missing entries, shape and motion. In the second part, a method based on applying robust M-estimator reduces the effect of outliers by replacing the L_2 norm in cost function (Eq. 15) by Cauchy function.

6.1 Nonlinear approach for missing data

If the point tracks are not visible in all images and the object has relatively small deformations, it has been proposed in [35] that instead of considering more complex and time-consuming optimisation algorithms, using a linear method based on principal component analysis (PCA) can recover the missing entries before shape and motion estimation. The major benefit of this imputation algorithm is its simple implementation and fast computation. The PCA as a linear method, is only able to cope well with simple deformations. Although the method is not suitable when the deformations are relatively large or complex, it still can be used for providing a starting point for the optimisation when the following nonlinear approach is applied.

The diffusion maps based method can be easily extended to handle the case with missing data. To facilitate this, Eq. 15 is modified with the cost function rewritten as $E(\{\mathbf{R}_t\}, \{\theta_{tl}\}, \{\mathbf{Y}_t{}^*\})$, depending explicitly on the missing

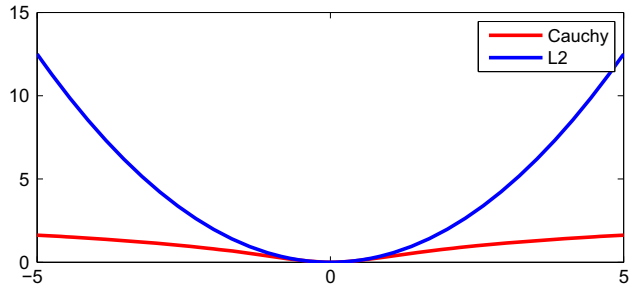

Fig. 6 Graphic representations of L_2 estimator and Cauchy function

observations $\mathbf{Y}_t{}^*$. As the results, the cost function is simultaneously minimised with respect to rotation, shape coefficients and the missing observations. It should be pointed out that only the missing observation $\mathbf{Y}_t{}^*$ are optimised not all 2D measurements $\mathbf{Y}_t$.

6.2 Robust estimator for data with outliers

It is well known that least-squares methods are sensitive to outliers as even a single outlier in the observations can strongly influence the values of the estimated parameters. Assuming that in the proposed model outliers affect only observations, the analysis can be restricted to the 2D re-projection error (the first term in Eq. 15). Defining the mismatch between the real observation (measurement) and the projection of the shape estimate at time t as a residual $\mathbf{E}_t$:

$$\mathbf{E}_t = \mathbf{Y}_t - \mathbf{P} \cdot \mathbf{R}_t \cdot \sum_{l=1}^{n+1} \theta_{tl} \mathbf{X}_{tl} \tag{18}$$

the camera motion and 3D shape are calculated by minimising the sum of squared residuals, $\sum_{t=1}^{F} \sum_{i=1}^{2} \sum_{j=1}^{P} \mathbf{E}_t^2(i, j)$, where i is the image coordinate index and j is a feature point index with P reconstructed points. In practice, the outliers will be presented in the observations and therefore the use of L_2 norm may lead to significant reconstruction errors. In this paper it is proposed to address this problem by using a function which penalises the large residual errors less than the L_2 norm and in this way "desensitise" the re-projection error with respect to outliers. The adopted approach uses the so-called M-estimators, in this case the new re-projection error is defined as: $\sum_{t=1}^{F} \sum_{i=1}^{2} \sum_{j=1}^{P} f(\mathbf{E}_t(i, j))$, where $f(\cdot)$ is Cauchy function $f(x) = \frac{c^2}{2} \log(1+(x/c)^2)$. The results presented in the next section were obtained for $c = 1$. Figure 6 illustrates the L_2 estimator and Cauchy function. The L_2 estimator is non-robust as it strongly depends on large errors, while Cauchy function considerably reduce their influence which make it less sensitive to outliers.

7 Results and discussion

A number of experiments were carried out to evaluate the proposed method. Several state-of-the-art algorithms were evaluated and compared in these experiments:

RF: The proposed random forest method;
DM: The diffusion map-based method. The DM method is similar to the RF except the manifold learning was implemented without random forest. [34];
PTA: The discrete cosine transform (DCT)-based point trajectory approach [2];
CSF: The column space fitting method [14];
KSFM: The kernel non-rigid structure from motion approach [15];
IPCA: The incremental principal components analysis-based method [35].

The testing data used for evaluation include: two articulated face sequences, *surprise* and *talking*, both captured using 3D scanner with 3D tracking of 83 facial landmarks and two surface models, *cardboard* and *cloth* [39]. This paper does not focus on feature detection and tracking. In the experiments described here the 3D points are known and these were projected onto the image sequences under the orthographic camera model and subsequently used as features. Diffusion maps require training process, so training dataset for face sequences were taken from the BU-3DFE [43] and for surface sequences the data were obtained from [39]. All the training data have been rigidly co-registered, and the same testing data have been used with the methods which do not require training.

During the experiments ten trials are run for each level of noise, missing data and outliers in different sequences. The reconstructed shapes are aligned using a single global rotation based on Procrustes alignment [2]. For evaluating the results, the same procedure as in [15] is used. The normalised means of the 3D error are compared over all frames and all points:

$$e = \frac{1}{\Delta F P} \sum_{t=1}^{F} \sum_{p=1}^{P} e_{tp}, \;\; \Delta = \frac{1}{3F} \sum_{t=1}^{F} (\Delta_{tx} + \Delta_{ty} + \Delta_{tz}) \tag{19}$$

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 point p at frame t in the reconstructed and ground truth data.

7.1 Comparison with previous methods

Table 1 shows the 3D reconstruction error for PTA, CSF, KSFM, IPCA, DM and RF. The manifold-based method DM

Table 1 Relative normalised mean reconstruction 3D error for PTA, CSF, KSFM, IPCA, DM and RF methods

	PTA	CSF	KSFM	IPCA	DM	RF		
						Initial	No opt.	Opt.
Surprise	0.037(12)	0.040(3)	0.038(4)	0.129	0.035(10)	0.315	0.293	**0.024**(15)
Talking	0.087(10)	0.057(3)	0.050(4)	0.099	0.035(10)	0.966	0.084	**0.034**(10)
Cardboard	0.289(8)	0.324(3)	0.275(2)	0.245	0.106(10)	0.267	0.161	**0.094**(10)
Cloth	0.353(6)	0.261(6)	0.181(2)	0.191	0.029(7)	0.297	0.173	**0.025**(7)
Walking	0.395(2)	0.168(2)	0.103(5)	0.326	**0.027**(9)	0.350	0.163	0.037(15)
IndianDance	0.485(13)	0.337(7)	0.234(7)	0.344	0.098(10)	0.297	0.128	**0.056**(15)
Capoeira	0.513(6)	0.365(4)	0.238(7)	0.406	0.026(9)	0.406	0.292	**0.005**(10)
Stretch	0.109(12)	0.071(8)	0.074(12)	0.192	0.069(6)	0.262	0.171	**0.059**(10)
Dance	0.294(5)	0.268(2)	0.237(4)	0.306	0.168(7)	0.261	0.153	**0.117**(15)

The optimal number of bases n, for which the 3D errors are shown in the table, is given in brackets for each tested method
The best results for each sequence are in bold

and RF on average provide better results than other trajectory-based methods. The relative normalised means (Eq. 19) of the 3D error [14] are compared over all frames and all points. For RF method the initialisation error and the error produced by the proposed algorithm with and without nonlinear refinement are given. The errors shown in the table correspond to the selected optimal dimensionality parameter n, in case of RF method this corresponds to the dimensionality of the estimated manifold. This selection is achieved by running the trials with n varying from 2 to 15. The best selected n for each tested method is shown in brackets. IPCA uses different number of basis shapes for constructing offline and online shapes, thus n is not provided in the table. As shown in the table, the previously proposed methods are able to provide comparable results, to the DM and RF nonlinear manifold-based methods, for objects with small deformations, e.g. faces. 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 deformable shape sequences (e.g. *Cloth*). As expected, the proposed RF method delivers the most accurate reconstruction in all tested cases. This mainly, can be explained as being due to the fact that the estimated affinities generate relatively uniform distribution of the training shapes in the reduced space. This subsequently effects the interpolation of the new shapes in the manifold, leading to more accurate reconstructions. Note that even though the initial error is big, 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 n has to be independently estimated for each sequence. Choosing too big n may lead to an ill-conditioned problem, but the point trajectory cannot be comprehensively represented if n is too small, while the results from the proposed method are more predictable. However, it should be noticed that the comparison for the *Walking, IndianDance, Capoeira, Stretch and Dance* sequences between the proposed method and the other methods may be seen as unfair, as better reconstruction accuracy of the proposed method comes at the cost of required availability of a representative training dataset.

7.2 The influence of embedding dimensionality

The accuracy of reconstruction is affected by the dimensionality of the reduced space n, corresponding to number of shape basis. The test described in this section looked at the relation between manifold dimensionality and the shape reconstruction error. The four sequences are tested individually with various dimensionalities ($n = 3, 5, 7, 10, 15$) of the reduced space. The forests have been trained with 600 trees. The results in Fig. 7a show that the shape reconstruction error is reduced with increasing dimension n of the reduced space. As expected, a higher number of bases is required to describe a complex shape deformation, e.g. *cloth* sequences.

Figure 7b shows the results obtained on the *cloth* sequence, comparing performance of the proposed method against previously proposed methods. The error calculated for PTA, CSF and KSFM varies with the number of bases, and overfitting occurs when $n > 10$ which indicates that the problem becomes ill-conditioned. DM and RF methods are "more stable" as the solution is strongly constrained by the requirement that it belongs to the manifold.

7.3 Sensitivity to noise and missing data

In order to assess the performance of the reconstruction algorithms when the observed data is corrupted by noise, the next experiment compared the RF method against previously proposed methods in terms of shape reconstruction error expressed as a function of level of noise in the observed data. We follow the process in [41] to simulate the noisy data, for which the measurements were perturbed by Gaussian

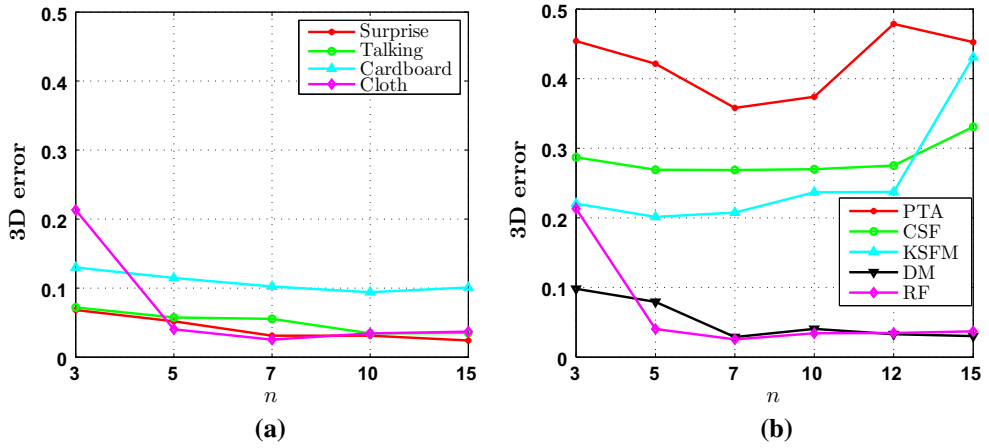

Fig. 7 Reconstruction 3D error as a function of the number of bases n. **a** Errors produced by RF with different sequences; **b** comparison results on *cloth* sequence

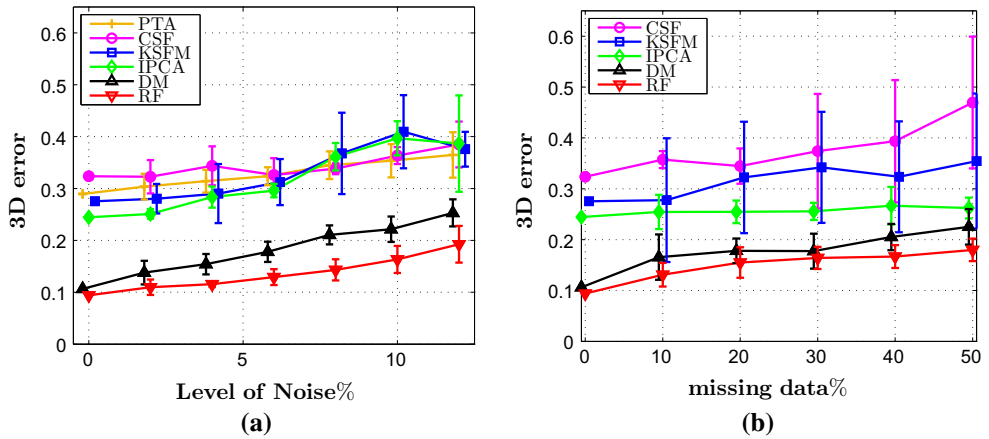

Fig. 8 Reconstruction results on *cardboard* data **a** sequence with Gaussian noise, **b** comparison results of missing data

noise. The noise level is represented by the ratio between the Frobenius norm of the noise and the measurement matrix. The reconstruction errors are evaluated for different level of noise, set to 2, 4, 6, 8, 10 and 12 %. The results are shown in Fig. 8a. It can be noticed that although the 3D reconstructed error of all five algorithms increases with the higher level of noise, the nonlinear method RF are obviously superior and achieve much smaller standard deviations, whereas others are quite sensitive with large mean error and error dispersion.

Missing data problem happens very often in real cases due to feature points track loss or occlusion. To simulate the measurement with missing data, 10, 20, 30, 40 and 50 % of the 2D entries in $\mathcal{Y}$ were randomly discarded. The results shown in Fig. 8b are calculated with the missing data ratio of up to 50 %, the average (maximum) 3D error using RF method is 0.1798 (0.2018) which is still acceptable. As the missing data problem is not addressed in [2], PTA is not used for comparison in this experiment.

Observing that in practice, measurements are likely to be affected by missing data and noise at the same time, the following experiment aims to evaluate the performance with measurement noise and different percentage of missing data. Figure 9a shows the results obtained by the nonlinear approach presented in Sect. 6.1. For comparison Fig. 9b shows results of the same experiment using the adaptive linear approach [35].

7.4 Sensitivity to outliers in the measurements

For real reconstruction cases, outliers present in the measurements is an inevitable problem that many previous algorithms did not address. It is well known that the outliers can severely affect the results. A small number of outliers may lead to completely meaningless reconstructed shapes. The robust method for mitigating the affect has been presented in Sect. 6.2. The experiment described in this section is designed to test the robustness of the proposed method using both original (Eq. 15) and the improved cost function applying robust influence function (M-estimator) for the reprojection error. To simulate the data corrupted by the outliers, a number of

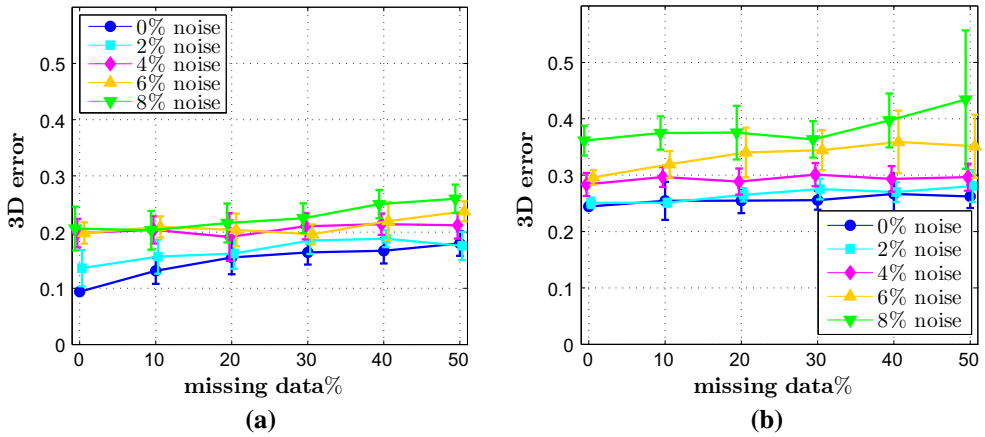

Fig. 9 Reconstruction results on *Cardboard* sequences. **a** Results for missing data and noise using nonlinear RF approach. **b** Results for missing data and noise using adaptive linear approach IPCA

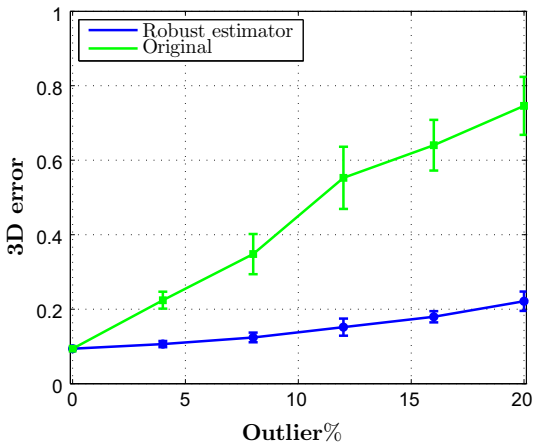

Fig. 10 Comparison results using original and improved RF methods on *Cardboard* sequence with outliers

the feature points are randomly selected in a test sequence as percentage of the total number of points, subsequently these points are replaced with randomly selected points in the same image frame. The effects of outliers on the reconstruction accuracy were tested at 4, 8, 12, 16 and 20 %. Figure 10 shows the 3D reconstructed errors as a function of the percentage of outliers given by the improved method using robust estimator and the original RF method. The robust method is able to handle the outliers and provide relatively accurate reconstructed shapes. Note that the outliers may also corrupt the training data, but the training is usually done offline, and therefore it is easier to filter the data and remove any outliers in this case.

7.5 Effect of forest model parameters

The analysis in this section is focused on how different choices of forests design parameters effect reconstructed results. The model has been evaluated in terms of two parameters: Tree depth and the forest size. For each experiment, the same testing sequences *Cardboard* and *Cloth* are used to represent small and large deformation shapes respectively.

The effect of varying tree depth has been investigated first. Here, all the forests have been trained with the fixed number of trees $T = 500$, and the maximum tree depths are varying from 2 to 7. As the forest size is sufficiently large, the variability due to randomness of parameter selection for each tree is averaged out, therefore we only show the results from one trial for each tree depth in Table 2 as the repeated experiments produce very similar results. The results show that in general increasing the tree depth decreases the error. Larger trees can better separate the shapes, thus the shapes ending up at the same leaf node are more similar. However for the data with small deformation (e.g.*cardboard*), the 3D error levels off and does not strongly depend on the tree depth. This is because that data exhibit relatively small deformations, and is governed by a small number of degrees of freedom, thereby even a small tree can well separate the shapes. It should be noticed that although larger trees can improve the results, it may lead to costly computation.

The proposed method has been also tested with respect to varying number of trees (T =10, 50, 100, 500, 1000) with fixed maximum tree depth of 5 in each experiment. The results for two sequences are shown in Table 3. The average 3D error and the standard deviation were estimated based on ten trials. As observed from the table, for both sequences applying more trees in the training process produces more accurate reconstructed results. Increasing number of trees in the forest can help the affinity Matrix $\mathbf{W}$ to better approximate to the true pairwise graph affinity.

7.6 Qualitative evaluation

The objective of this section is to provide a qualitative evidence for the assessment of the shape reconstruction quality. Following on the experiment summarised in Table 2, for com-

Table 2 The 3D reconstruction error as a function of varying tree depth

Sequence	Depth					
	2	3	4	5	6	7
Cardboard	0.1049	0.0907	0.0957	0.0940	0.0922	0.0899
Cloth	0.1462	0.0697	0.0457	0.0372	0.0254	0.0276

Table 3 The 3D reconstruction error as a function of different number of trees in the forest

Sequence	Tree number				
	10	50	100	500	1000
Cardboard					
3D error	0.1890	0.1425	0.1252	0.0942	0.0929
Max error	0.2650	0.1975	0.1352	0.1029	0.0997
SD	0.0464	0.0280	0.0044	0.0048	0.0046
Cloth					
3D error	0.1415	0.1170	0.0689	0.0255	0.0256
Max error	0.2956	0.2940	0.1532	0.0311	0.0349
SD	0.0825	0.0702	0.0389	0.0026	0.0042

parison Fig. 11 shows three randomly selected reconstructed shapes from the *Cloth* sequence using PTA, CSF, KSFM, IPCA, DM and RF methods.

The proposed algorithms have been also tested on a real video sequence showing paper being bended. A frames' sample from this video is shown in the top row of Fig. 12. In the video, 81 features were tracked along 61 frames showing approximately two periods of bending movement. The training data in this experiment is the cardboard dataset obtained from [39], which is the same as used in the previous evaluations when using *cardboard* sequence. The results show a comparison of our reconstructed shapes with the results obtained from MP, PTA, KSFM methods. Figure 13 shows 3D reconstruction results obtained for a facial expression sequence. The first row indicates the input images with

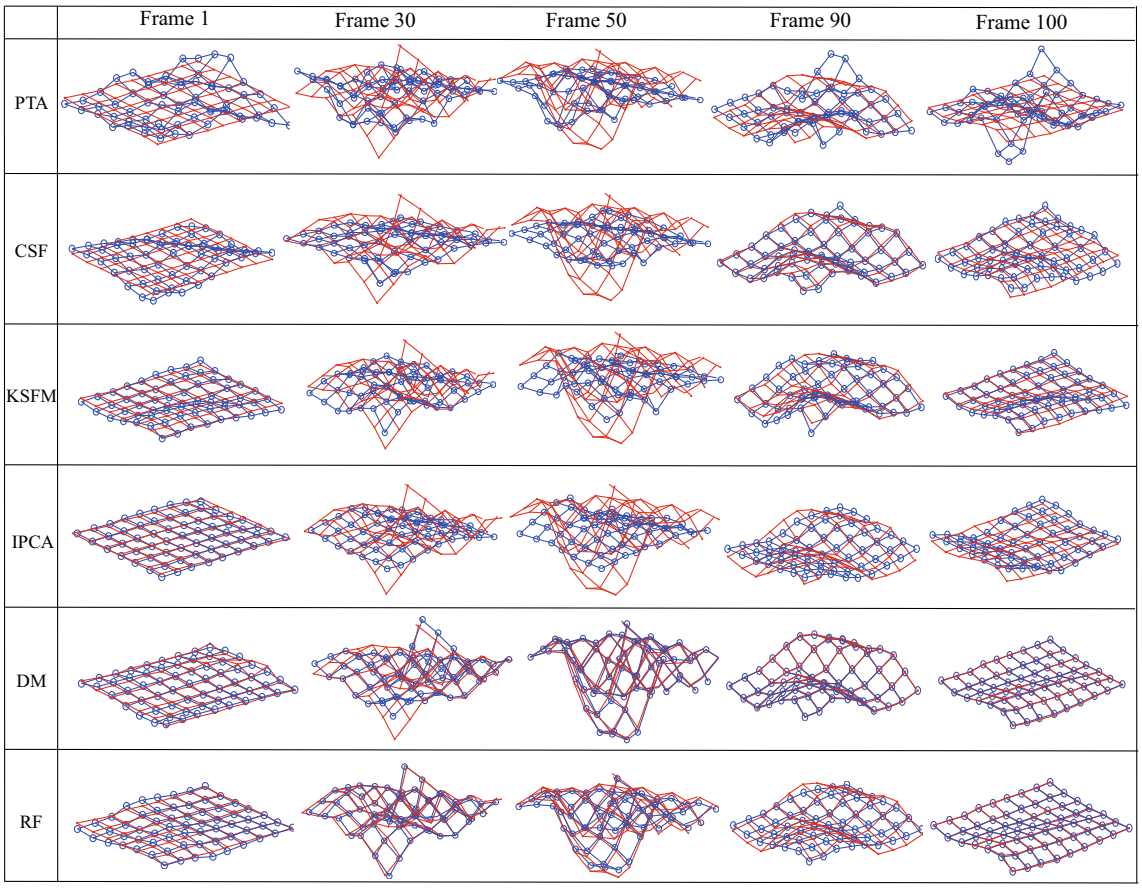

Fig. 11 Reconstruction results on the *Cloth* sequence. Reconstructed 3D shapes (*blue*), with ground truth (*red*) are shown (colour figure online)

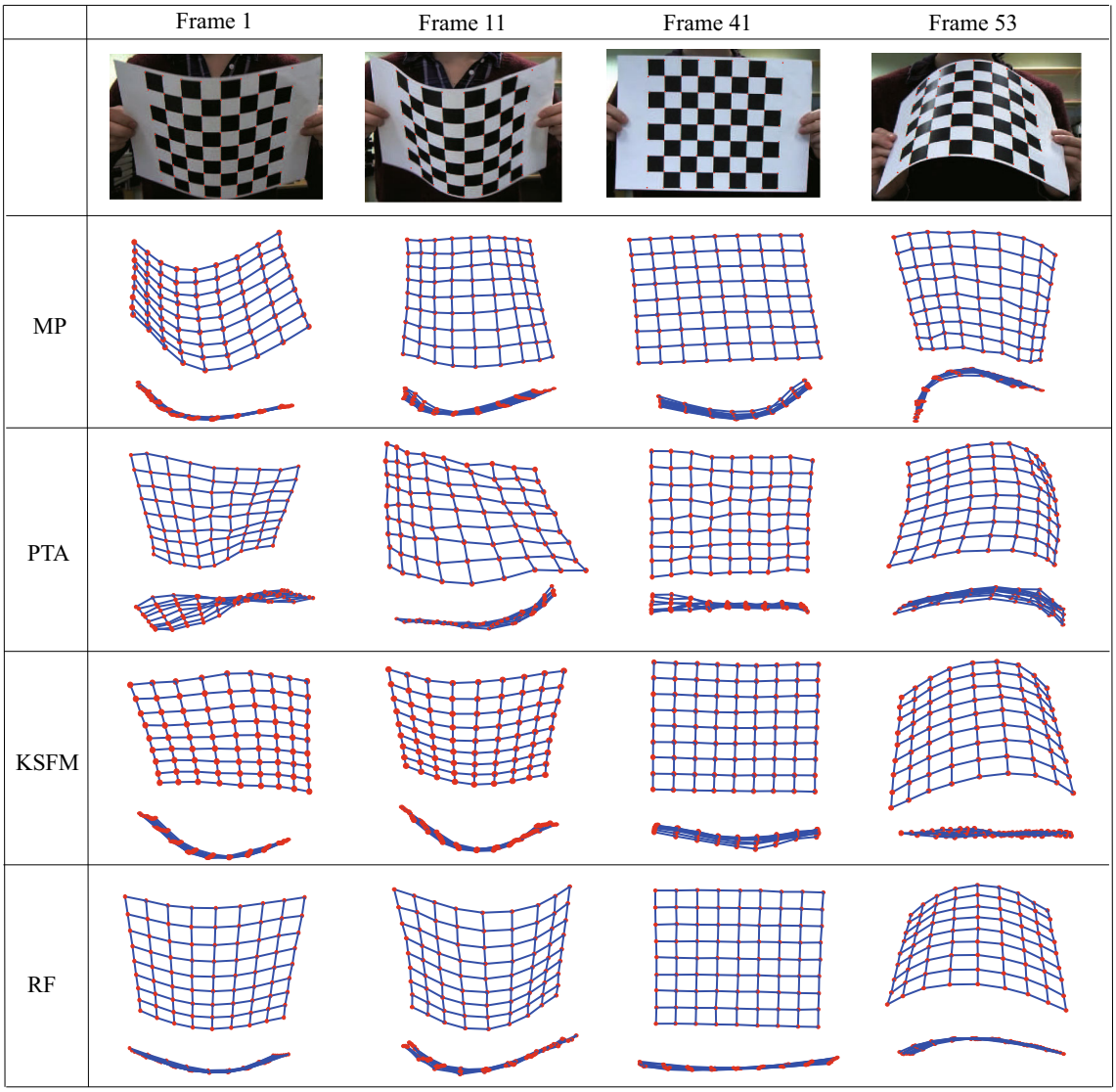

Fig. 12 Selected 2D frames from the video sequence of a paper bending. *Front* and *top views* of the corresponding 3D reconstructed results using the proposed method (RF), MP, PTA and KSFM

tracked feature points; the second and third row show the front and side views of the reconstructed results using the proposed RF method.

7.7 Results of using reduced training set

This section presents experiments of 3D reconstruction when only small number of training samples are being used. The method is labelled as **RF2** to distinguish it from RF. Two sequences are used for testing, they are: *cloth* and an articulated human motion sequence *IndianDance* from CMU motion capture database.[1] Since no separate training data are provided for human motion sequence, every 10th frame is selected to build training dataset (e.g. frame 1,11,21…are selected), whereas frames 5,15,25…are selected as a testing set. It should be stressed that the results provided for the *IndianDance* sequence are indicative only, as an objective tests should use independent training and test data, which are currently not available. In this experiment, the number of training shapes for RF2 and RF for the *cloth* sequence are 100 and 1000 respectively, and 100 for *IndianDance* sequence when RF2 is used. For the cloth sequence the reconstruction error for RF2 with the optimal selection of the number of basis shapes (9 in this case) is 4.71 %. This should be compared with the results reported in Table 1. It can be concluded that RF and RF2 are somewhat comparable in terms of the reconstruction accuracy for *Cloth* sequence. As expected, RF outperform the RF2, but RF2 uses much smaller training set than RF.

[1] The data were obtained from http://mocap.cs.cmu.edu.

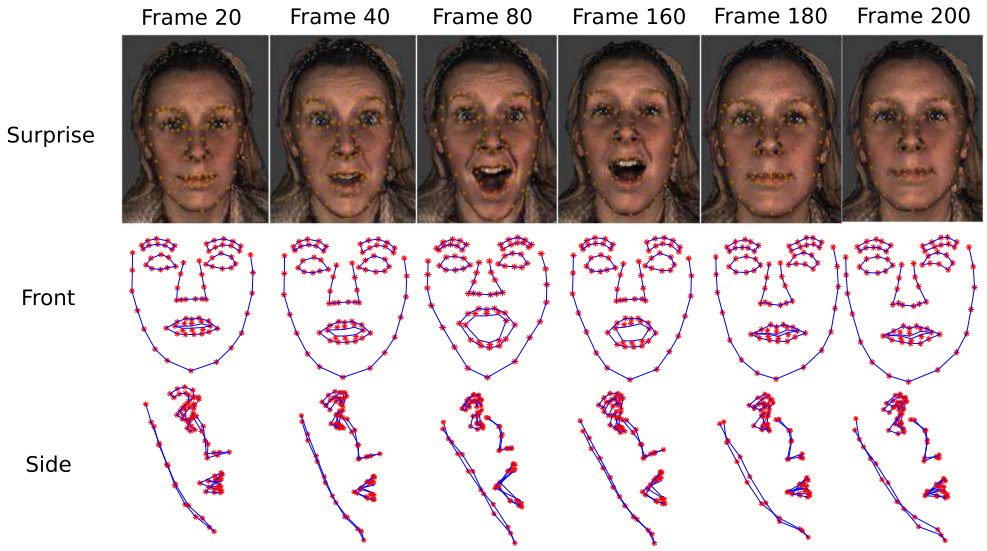

Fig. 13 Selected frames from a video sequence of a "surprise" facial expression. *First row* Input images with tracked feature points. *Second* and *third rows* Front and side views of the 3D reconstruction using the proposed RF method

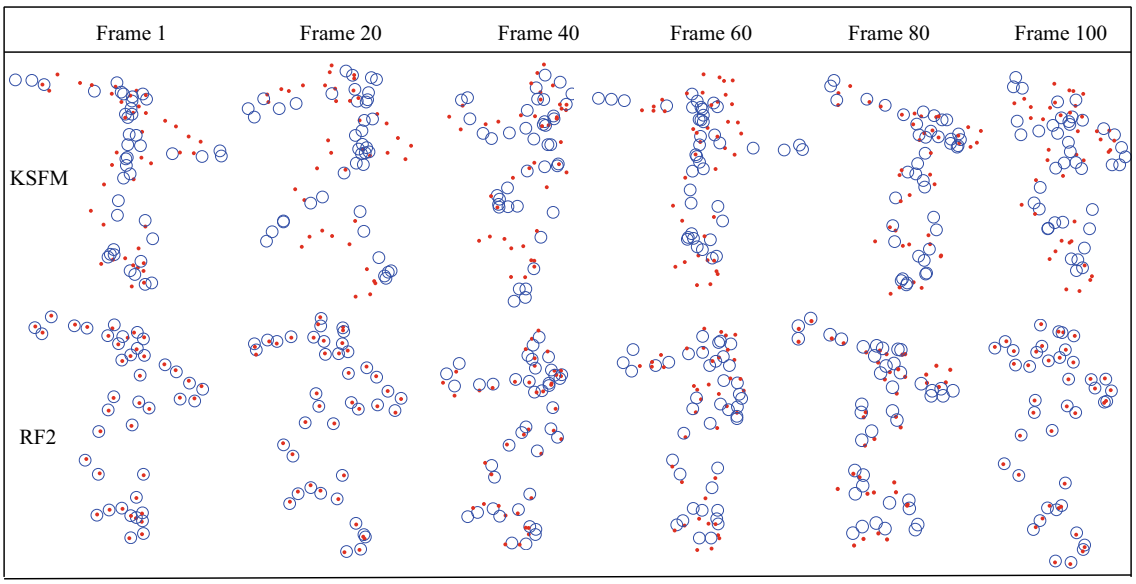

Fig. 14 Reconstruction results on the *IndianDance* sequence. Reconstructed 3D shapes (*circles*), with ground truth (*dots*) are shown

The performance of the RF2 method strongly depends on the selection of the shapes included in the reduced training set. It is beneficial when selected shapes generate well-shaped triangles in the Delaunay triangulation. In the performed test on the *IndianDance* sequence the selection of the training samples was not optimised with respect to results of the Delaunay triangulation. As human movement contains large number of degrees of freedom, the reconstruction results are affected if corresponding shapes are being clustered in badly shapes triangles (e.g. 'skinny triangles') in the reduced space. Even so the reconstructed error for *IndianDance* is 12.95 % obtained with seven basis shapes, which is still acceptable.

The visualised results of reconstructed shape extracted from the *IndianDance* sequence using KSFM (58.86 % overall 3D reconstruction error) and RF2 methods are illustrated in Fig. 14.

Limitation

While the proposed method is able to reconstruct complex deformable shapes with some success, some limitations still

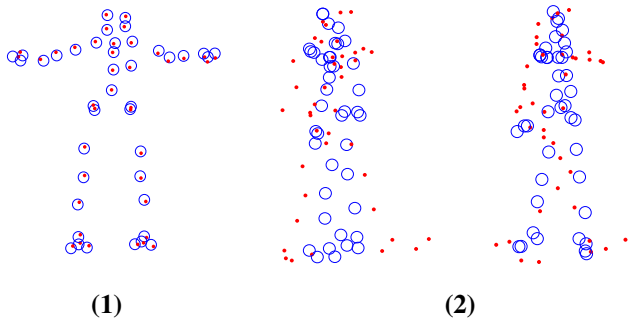

Fig. 15 The reconstruction success (*1*) and failures (*2*) examples obtained for the *Capoeria* sequence. The ground truth and the reconstructed shapes are shown as *dots* and *circles*, respectively

remain. For example, the method could fail when the shape to be reconstructed is not adequately represented in the available training dataset. Figure 15 shows examples of the successful and failed reconstructions. In this case, the *Capoeria* sequence is used for testing, whereas the above mentioned *IndianDance* sequence is used for training. The shapes are accurately reconstructed when they are sufficiently well represented by the training samples, but the reconstruction can fail when the true shape is significantly different from the shapes in the training set (see Fig. 15).

8 Conclusions

A new approach for recovery non-rigid shape based on manifold forests is described in the paper. The nonlinear manifold has been build upon diffusion maps with random forests used to estimate local manifold neighbourhood topology. The method achieves good performance especially for large and complex deformable objects, when compared with the existing approaches.

In many practical applications there are only limited number of shape examples to be used for training. To address this problem, a modification of the described RF method has been also proposed which is able to accurately reconstruct shapes even though only a small number of training shapes could be available. Additionally, the existence of outliers in the observations could be a limitation imposed by practical applications on some of the previously proposed reconstruction methods, as often outliers are not modelled explicitly. To address this problem, a further extension of the prosed RF method has been described in this paper, with a robust cost function used to measure the re-projection error. The evaluation results on simulated and real data presented in the paper demonstrate the validity of the proposed methods.

It should be mentioned that the comparison of the proposed method with respect to the other methods may be seen as unfair, as better reconstruction accuracy of the proposed method comes at the cost of required availability of a representative training dataset. 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.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Akhter, I., Sheikh, Y., Khan, S.: In defense of orthonormality constraints for nonrigid structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534–1541 (2009)
2. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Trajectory space: a dual representation for nonrigid structure from motion. IEEE Trans. Pattern Anal. Mach. Intell. **33**(7), 1442–1456 (2011)
3. Arias, P., Randall, G., Sapiro, G.: Connecting the out-of sample and pre-image problems in kernel methods. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
4. Bartoli, A., Gay-Bellile, V., Castellani, U., Peyras, J., Olsen, S., Sayd, P.: Coarse-to-fine low-rank structure-from-motion. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
5. Berg, M., Cheong, O., Kreveld, M., Overmars, M.: Computational Geometry: Algorithms and Applications. Springer, Berlin (2008)
6. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 690–696 (2000)
7. Brunet, F., Hartley, R., Bartoli, A., Navab, N., Malgouyres, R.: Monocular template-based reconstruction of smooth and inextensible surfaces. In: Asian Conference on Computer Vision, pp. 52–66. Springer (2011)
8. Buchanan, A., Fitzgibbon, A.: Damped newton algorithms for matrix factorization with missing data. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 316–322 (2005)
9. Coifman, R., Lafon, S.: Diffusion maps. Appl. Comput. Harmonic Anal. **21**(1), 5–30 (2006)
10. Coifman, R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., Zucker, S.W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Natl. Acad. Sci. **102**(21), 7426–7431 (2005)
11. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends Comput. Graph. Comput. Vis. **7**, 81–227 (2012)
12. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: Monoslam: real-time single camera slam. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1052–1067 (2007)
13. Eriksson, A., van den Hengel, A.: Efficient computation of robust weighted low-rank matrix approximations using the l1 norm. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1681–1690 (2012)
14. Gotardo, P., Martinez, A.M.: Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 2051–2065 (2011)

15. Gotardo, P., Martinez, A.M.: Kernel non-rigid structure from motion. In: IEEE International Conference on Computer Vision, pp. 802–809 (2011)
16. Hamsici, O.C., Gotardo, P.F.U., Martinez, A.M.: Learning spatially-smooth mappings in non-rigid structure from motion. In: European Conference on Computer Vision, pp. 260–273. Springer (2012)
17. van den Hengel, A., Russell, C., Dick, A., Bastian, J., Pooley, D., Fleming, L., Agapito, L.: Part-based modelling of compound scenes from images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 878–886 (2015)
18. Marques, M., Costeira, J.: Estimating 3D shape from degenerate sequences with missing data. Comput. Vis. Image Underst. **113**(2), 261–272 (2009)
19. Moreno-Noguer, F., Salzmann, M., Lepetit, V., Fua, P.: Capturing 3D stretchable surfaces from single images in closed form. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1842–1849 (2009)
20. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: real-time dense surface mapping and tracking. In: 2011 10th IEEE international symposium on Mixed and Augmented Reality (ISMAR), pp. 127–136. IEEE (2011)
21. Östlund, J., Varol, A., Ngo, D.T., Fua, P.: Laplacian meshes for monocular 3D shape recovery. In: European Conference on Computer Vision, pp. 412–425. Springer (2012)
22. Paladini, M., Bue, A., Xavier, J., Stosic, M., Dodig, M., Agapito, L.: Factorization for non-rigid and articulated structure using metric projections. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2898–2905 (2009)
23. 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 Conference on Computer Vision and Pattern Recognition, pp. 2898–2905 (2009)
24. Rabaud, V., Belongie, S.: Re-thinking non-rigid structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
25. Rabaud, V., Belongie, S.: Linear embeddings in non-rigid structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2427–2434 (2009)
26. Rey, W.: Introduction to Robust and Quasi-robust Statistical Methods, vol. 983. Springer, New York (1983)
27. Rodola, E., Rota Bulò, S., Windheuser, T., Vestner, M., Cremers, D.: Dense non-rigid shape correspondence using random forests. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4177–4184. IEEE (2014)
28. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
29. Salzmann, M., Fua, P.: Linear local models for monocular reconstruction of deformable surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 931–944 (2011)
30. Salzmann, M., Moreno-Noguer, F., Lepetit, V., Fua, P.: Closed-form solution to non-rigid 3D surface registration. In: European Conference on Computer Vision, pp. 581–594. Springer (2008)
31. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. The MIT Press, Cambridge (2002)
32. Tao, L., Matuszewski, B.J.: 3D deformable shape reconstruction with diffusion maps. In: British Machine Vision Conference (2013)
33. Tao, L., Matuszewski, B.J.: Deformable shape reconstruction from monocular video with manifold forests. Comput. Anal. Images Patterns 28–36 (2013)
34. Tao, L., Matuszewski, B.J.: Non-rigid structure from motion with diffusion maps prior. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1530–1537 (2013)
35. Tao, L., Mein, S.J., Quan, W., Matuszewski, B.J.: Recursive non-rigid structure from motion with online learned shape prior. Comput. Vis. Image Underst. **117**(10), 1278–1289 (2013)
36. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. Int. J. Comput. Vision **9**(2), 137–154 (1992)
37. Torresani, L., Yang, D.B., Alexander, E.J., Bregler, C.: Tracking and modeling non-rigid objects with rank constraints. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001, vol. 1, pp. I–493. IEEE (2001)
38. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment: a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R., (eds.) Vision Algorithms: Theory and Practice, pp. 298–372. Springer, Berlin, Heidelberg (2000)
39. Varol, A., Salzmann, M., Fua, P., Urtasun, R.: A constrained latent variable model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2248–2255 (2012)
40. Vidal, R., Tron, R., Hartley, R.: Multiframe motion segmentation with missing data using powerfactorization and GPCA. Int. J. Comput. Vision **79**(1), 85–105 (2008)
41. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. In: European Conference on Computer Vision, pp. 573–587. Springer (2004)
42. Yan, J., Pollefeys, M.: A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. IEEE Trans. Pattern Anal. Mach. Intell. **30**(5), 865–877 (2008)
43. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3D face expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition, pp. 211–216 (2006)
44. Zaheer, A., Akhter, I., Baig, M.H., Marzban, S., Khan, S.: Multiview structure from motion in trajectory space. In: IEEE International Conference on Computer Vision, pp. 2447–2453 (2011)
45. Zhu, Y., Huang, D., De La Torre, F., Lucey, S.: Complex non-rigid motion 3D reconstruction by union of subspaces. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1542–1549. IEEE (2014)

Lili Tao is a research associate at the Visual Information Lab in the Department of Computer Science at the University of Bristol. Her research interests include human motion analysis, 3D deformable object reconstruction, and facial expression analysis. Tao received a PhD in computer vision from the University of Central Lancashire.

Bogdan J. Matuszewski is a professor of computer vision in the College of Science and Technology at the University of Central Lancashire and head of the Robotics and Computer Vision Research Laboratory at the School of Engineering. His research interests include use of Bayesian methodology for modelling, tracking, and recognition; deformable models, variational, and partial differential equation-based methods for image analysis applied to data registration, segmentation, and interpretation; and biomedical and industrial applications of computer vision and machine learning. He received a PhD in electronics from Wroclaw University of Technology, Poland.