Non-Rigid Structure-From-Motion on Degenerate Deformations With Low-Rank Shape Deformation Model

Zhong Zhou, *Member, IEEE*, Feng Shi, Jiangjian Xiao, *Member, IEEE*, and Wei Wu

***Abstract*—Non-rigid structure-from-motion (NRSfM) is the process of recovering time-varying 3D structures and poses of a deformable object from an uncalibrated monocular video sequence. Currently, most NRSfM algorithms utilize a non-degenerate assumption for non-rigid object deformations whereby the 3D structures of a non-rigid object can be assumed to be a linear combination of basis shapes with full rank three. Unfortunately, this assumption will produce extra degrees-of-freedom when the non-rigid object has some degenerate deformations with shape bases of rank less than three. These extra degrees-of-freedom will yield spurious shape deformations due to non-negligible noise in real applications, which will cause substantial reconstruction errors. To solve this problem, we propose a low-rank shape deformation model to represent 3D structures of degenerate deformations. When modeling degenerate deformations, the proposed model exploits the rank-deficient nature of degenerate deformations in addition to the low-rank property of non-rigid objects' trajectories, thus providing a more accurate and compact representation compared with existing models. Based on this model, we formulate the NRSfM problem as two coherent optimization problems. These problems are solved with iterative non-linear optimization algorithms. Experiments on synthetic and motion capture data are conducted. The results exhibit the significant advantages of our approach over state-of-the-art NRSfM algorithms for the 3D recovery of non-rigid objects with degenerate deformations.**

***Index Terms*—Degenerate deformations, low-rank shape deformation model, non-rigid structure from motion, 3D reconstruction.**

I. Introduction

THE RAPID development of portable video recording devices has led to the mass production of video resources, which users find increasingly convenient to access through the Internet. Against this background, a large number of algorithms and systems have been proposed for video content analysis and retrieval [28], [29]. Among these, recovering 3D structures of objects in videos is an active research area, and the recovered depth information is potentially useful in various applications, such as human-computer interaction, object/action recognition, 3D facial reconstruction and augmented reality [25]–[27], [34]. This paper focuses on the non-rigid structure-from-motion (NRSfM) problem that recovers time-varying 3D structures and poses of a deformable object from an uncalibrated monocular video sequence using feature tracking points. Compared with 3D reconstruction algorithms for non-rigid objects that require input videos from multiple and/or calibrated cameras (e.g., [30], [31]), NRSfM enables 3D reconstruction using single-camera videos; thus, it has broader applicability.

An underlying assumption of NRSfM is that the movement of points on the surface of a deforming object is not random or unpredictable; instead, the movement has inherent spatio-temporal correlations. In the seminal work on NRSfM [2], Bregler *et al.* exploited this spatio-temporal correlation to propose a low-rank shape model that represents 3D structures of a deformable object as a linear combination of a small number K of 3D basis shapes (see Fig. 1 for an illustration). This low-rank shape model [2] proved to be effective, and it was followed by many researchers [3], [4], [6], [8]–[11]. However, all of these algorithms are based on a non-degenerate assumption, where the non-rigid object is deformed with K non-degenerate shape bases of full rank 3. Therefore, these NRSfM algorithms are mainly suitable for the 3D recovery of non-rigid objects with non-degenerate deformations. Unfortunately, many objects in the real world have non-rigid deformations along 1 or 2 dimensions, which can only be compactly represented by degenerate shape bases of rank less than 3. Such non-rigid objects are referred to as *degenerate deformations* [17]. For example, if a scene contains pedestrians walking independently along straight lines, the shape bases referring to those 1D translations are degenerate and may have a rank merely equal to 1. In facial expression analysis, the deformations of a face are dominant in the horizontal and vertical directions and are relatively subtle in the depth direction, which will also cause degeneracy problems in facial recovery. Under degenerate deformations, the non-degenerate assumption in traditional NRSfM algorithms will introduce extra degrees-of-freedom (DoFs) that are usually not constrained by input data and that will end up as fitting noise. In real cases, because of various video limitations, such as low resolution, motion blur, jerky camera motion and non-uniform

Manuscript received July 10, 2014; revised September 11, 2014 and November 23, 2014; accepted December 13, 2014. Date of publication December 19, 2014; date of current version January 15, 2015. This work was supported by the National 863 Program of China under Grant 2012AA011801 and Grant 2012AA011803 and by the Natural Science Foundation of China under Grant 61170188. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo.

Z. Zhou, F. Shi, and W. Wu are with the State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: zz@buaa.edu.cn; supersf2008@hotmail.com; wuwei@buaa.edu.cn).

J. Xiao is with the Ningbo Industrial Technology Research Institute, CAS, Ningbo 315201, China (e-mail: xiaojj@nimte.ac.cn).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a video file, which shows the results of quantitative and qualitative experiments described in the paper. This material is 46.5 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2384396

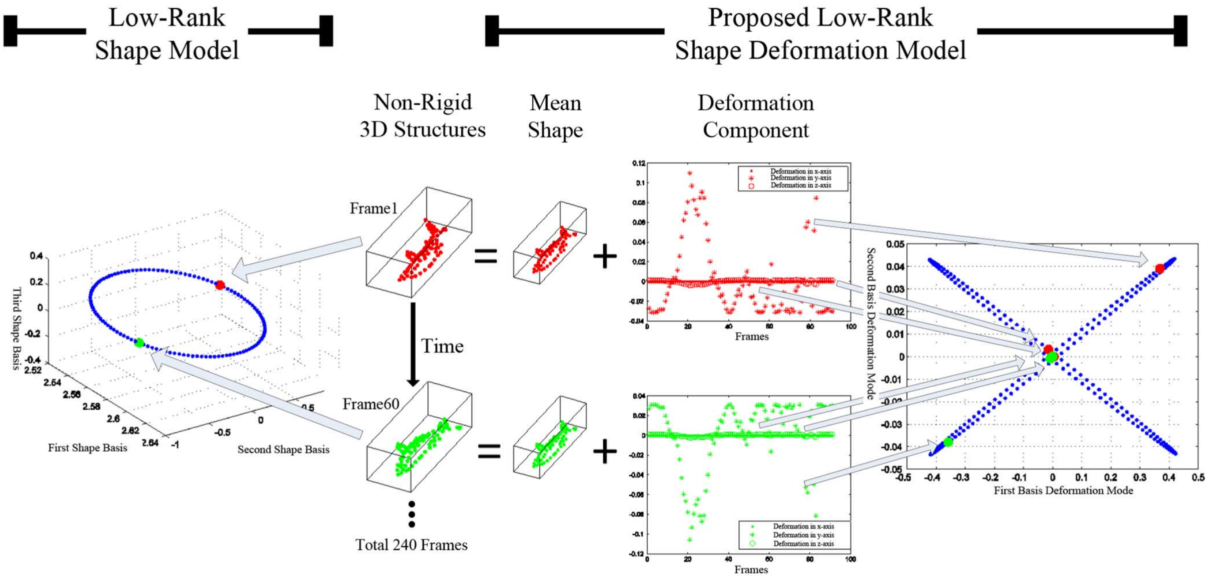


Fig. 1. Illustration of the proposed low-rank shape deformation model and the low-rank shape model [2]. The "Shark" sequence of [9] is taken as an example. In these two models, the red and green points correspond to the 3-D structures in frames 1 and 60, respectively, while the blue points refer to the 3-D structures in other frames. Note that the low-rank shape model uses three 3-D basis shapes to linearly represent non-rigid 3-D structures. Our model describes a deforming object as the mean shape plus the associated deformations, and represents the deformation component as a linear combination of two 1-D basis deformation modes.

illumination, 2D feature points tracked from a video usually include non-negligible noise, which may produce spurious shape deformations and cause significant reconstruction errors when traditional NRSfM algorithms are applied.

For the problem of the non-degenerate assumption, a low-rank shape deformation model is proposed to represent 3D structures of degenerate deformations. We first use the inherent low-rank property of trajectories to express 3D structures of a deforming object as a linear combination of K non-degenerate basis shapes. Then, the ambiguity between shape bases and coefficients is eliminated with Discrete Cosine Transform (DCT) vectors. The K shape bases are decomposed into a mean shape component and a deformation component. Finally, the rank-deficient nature of degenerate deformations is exploited to further decompose the deformation component into a linear combination of a small number of 1D basis deformation modes (see Fig. 1 for an illustration).

As shown in Fig. 1, the main difference between our model and the low-rank shape model is that our model uses a set of 1D basis deformation modes, whereas the low-rank shape model uses 3D basis shapes to linearly represent non-rigid 3D structures. This results in the flexibility of our model whereby the x, y and z components of non-rigid 3D structures can be described. More importantly, in our model, shape bases of any rank can be compactly represented using a combination of 1D basis deformation modes, whereas in the low-rank shape model, a shape basis always has to be of rank 3. Therefore, the proposed model has the ability to compactly represent 3D structures of degenerate deformations, and it can be applied when modeling special non-degenerate deformations. Subsequently, based on our proposed low-rank shape deformation model, we formulate the NRSfM problem as two coherent optimization problems: one problem is to recover the 3D structures of a deformable object, and the other problem is to estimate the object's rotations relative to the camera. Iterative non-linear optimization algorithms are then designed to solve the problems. In the experiments, we compare our method against a set of state-of-the-art NRSfM algorithms on both synthetic and motion capture data. The experimental results demonstrate that our method significantly outperforms other methods in terms of both accuracy and robustness when recovering 3D structures and poses of non-rigid objects with degenerate deformations.

The paper is organized as follows. Section II reviews related work. The background of NRSfM research and the motivation of our work are introduced in Section III. In Section IV, we propose a new low-rank shape deformation model to represent the 3D structures of degenerate deformations. Section V describes our new NRSfM algorithm based on the proposed low-rank shape deformation model and Section VI discusses the experimental results of using our algorithm with synthetic, motion capture and real sequences.

II. Related Work

Over the last two decades, a considerable number of algorithms have been proposed to recover time-varying 3D structures and poses of non-rigid objects from monocular videos. In this section, we first review shape-based and trajectory-based NRSfM algorithms. We then review existing NRSfM algorithms applied to degenerate deformations and introduce the difference between these algorithms and our algorithm.

A. Shape-Based NRSfM Algorithms

Most NRSfM algorithms are based on a low-rank shape model that represents 3D structures of a deformable object as a linear combination of a small number of 3D basis shapes. The low-rank shape model was first proposed by Bregler *et al.* [2]. Based on this model, Bregler *et al.* employed an SVD-based approach to factorize a 2D tracking matrix and then exploited the orthonormality of camera rotation matrices to recover 3D structures and poses of non-rigid objects. Subsequently, Xiao and Kanade [3] proved that enforcing the orthonormality constraint only is ambiguous and demonstrated that it can lead to incorrect solutions; in response, they introduced a uniqueness constraint on the shape bases and proved that imposing both the shape basis and the orthonormality constraints results in a closed-form solution. Brand [4] argued that this closed-form solution to NRSfM in [3] is sensitive to noise and to the selection of shape bases and proposed an optimization method for the NRSfM problem that can tolerate noisy input to some extent. Contrary to the conclusion in [3], Brand's method only uses the orthonormality constraint but can yield exact 3D reconstruction results with noiseless input. Akhter *et al.* [5] provided theoretical support for Brand's method [4] by proving that the orthonormality constraint alone is sufficient for the correct recovery of non-rigid structures. Akhter *et al.* [5] further indicated that solving the NRSfM problem based on orthonormality constraint alone will lead to a complicated non-linear optimization problem, which is difficult to solve reliably. To avoid this type of problem, other approaches introduced additional constraints to the NRSfM problem. For example, Torresani *et al.* [6] imposed a Gaussian prior on shape coefficients based on an assumption that the reconstructed 3D shapes at each frame are similar to each other. The authors also added a linear transition constraint on the shape coefficients to model the temporal dynamics in shapes. Olsen *et al.* [7] designed a temporal regularizer to constrain camera trajectories and shape coefficients to behave smoothly. They also designed a spatial regularizer to constrain neighboring image point tracks to have similar 3D spatial structures. Dai *et al.* [10] proposed exploiting inherent rank constraints on the low-rank shape model in [2] to facilitate 3D structure recovery. In [32], Del Bue introduced a known 3D shape as *a priori* information for the rigid component of a non-rigid 3D object. Tao *et al.* [33] integrated the diffusion map method to determine and apply the *a priori* shape that constrains reconstructed shapes. Agudo *et al.* [35] proposed using a modal analysis approach based on continuum mechanics to constrain the shape bases as a set of physically meaningful deformation modes.

B. Trajectory-Based NRSfM Algorithms

In contrast to the above shape-based NRSfM algorithms, Akhter *et al.* [12], [13] proposed a low-rank trajectory model that represents trajectories of a deformable object as a linear combination of several basis trajectories. Akhter *et al.* [13] proved that the low-rank trajectory model is equivalent to the low-rank shape model in [2] in terms of their representativeness. Furthermore, Akhter *et al.* [12], [13] proposed exploiting the inherent temporal smoothness of shape deformations to predefine the trajectory bases as DCT bases. This results in a considerable simplification of the underlying optimization process of NRSfM. In this way, the trajectory approach exhibits a much greater numerical stability and allows one to reconstruct more complicated non-rigid deformations with less error. Subsequently, Gotardo *et al.* [14] indicated that Akhter's low-rank trajectory model cannot describe the high-frequency components of shape deformations and that the results are often over-smoothed. To solve this problem, Gotardo *et al.* [14] proposed the 3D shape trajectory model that subsumes the low-rank shape and trajectory models by constraining shape coefficients with DCT bases. Later, Gotardo *et al.* [15] proposed the Kernel Shape Trajectory Approach (KSTA), which uses a kernel trick to capture non-linear relations between the 3D structures of a non-rigid object and its shape coefficients in the 3D shape trajectory model. As a result, KSTA is able to describe 3D structures of non-linear deformations more compactly than the 3D shape trajectory model. Based on the repetition of 3D structures of non-rigid objects, Khan [16] imposed a uniqueness constraint on shape coefficients of the 3D shape trajectory model. Consequently, the number of basis shapes required to represent the non-rigid shape is significantly reduced, making the NRSfM problem easier to solve. In addition, for solving the NRSfM problem with missing data input, the researchers in [14] and [20] proposed utilizing DCT bases to estimate missing entries in 2D input data.

C. NRSfM Algorithms Applied to Degenerate Deformations

A common attribute of the above shape- and trajectory-based NRSfM algorithms is that they all assume that shape deformations are non-degenerate. However, degenerate deformations often occur in the real world. Under degenerate deformations, Xiao and Kanade [17] proved that the ill-posed nature of the NRSfM problem is more serious; therefore, they introduced a positive semi-definite constraint to achieve the desired results. To solve the problem of sequential NRSfM in a more convenient manner, Paladini *et al.* [18] proposed the 3D implicit low-rank shape model that represents the 3D structures of degenerate deformations using a linear combination of PCA basis vectors. In [10], Dai *et al.* proposed the block matrix method, where a stronger but more meaningful low-rank constraint is imposed on the 3D structures of non-rigid objects. In this way, the non-degenerate assumption on shape deformations can be relaxed to some extent. Although failing to provide a solution, Angst *et al.* [19] demonstrated that recovering 3D structures of degenerate deformations is a non-trivial problem in NRSfM.

A new NRSfM algorithm that aims to recover 3D structures and poses of non-rigid objects with degenerate deformations is presented in this study. In our approach, we seek to address the fundamental problem of degenerate deformation recovery. Therefore, rather than introducing additional constraints similarly to [10] and [17], we propose a novel low-rank model to represent 3D structures of non-rigid objects with degenerate deformations. The main advantage of our model over the 3D implicit low-rank shape model [18] is that our model exploits the inherent low-rank property of trajectories of non-rigid objects when modeling degenerate deformations and thus provides a more compact representation. Then, based on the proposed model, we employ the factorization technique to recover 3D structures and poses of non-rigid objects. The factorization technique is the most common framework used in NRSfM [2]–[4], [8]–[10], [12], [14], [15], [17], and it originated from the factorization method proposed by Tomasi and Kanade for recovering rigid 3D structures [1]. Because of its more accurate and compact representation, our NRSfM algorithm achieves a much better performance when applied to degenerate deformations compared with existing NRSfM algorithms.

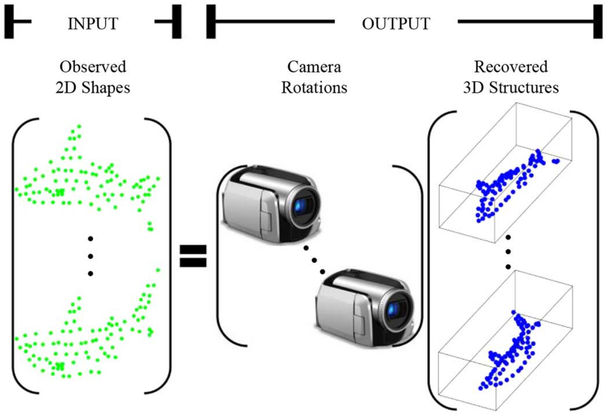


Fig. 2. Illustration of the NRSfM problem under the orthographic camera model. The "Shark" sequence of [9] is taken as an example.

III. Background and Motivation

A. Notation

We use uppercase letters (e.g., U) to denote matrices; the exceptions are F (the number of video frames), P (the number of feature points), and K (the number of shape bases or trajectories bases). Bold lowercase letters (e.g., $\mathbf{u}$) and the usual lowercase letters (e.g., u) denote vectors and scalars, respectively. $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the matrix Frobenius norm and the vector L_2 norm, respectively. $\odot$ and $\otimes$ represent the component-wise product and Kronecker product of matrices, respectively. I_m denotes an identity matrix of order m. $vec(\cdot)$ denotes the column-wise matrix vectorization operator. $U \in \mathbb{R}^{m\times n}$ indicates that U is a matrix with m rows and n columns.

B. Problem Statement

We assume that a deformable object consists of P time-varying 3D points $\mathbf{s}_f^p = (x_f^p, y_f^p, z_f^p)^T$, where p is the index over the points and f is the index over the image frames. These time-varying 3D structures represent object deformation in a local coordinate system. Assuming that at each instant f, these points undergo a rigid motion and orthographic projection to 2D

$$(u_f^p, v_f^p)^T = R_f(\mathbf{s}_f^p + \mathbf{d}_f) \tag{1}$$

where $(u_f^p, v_f^p)^T$ is the 2D projection of the point p at time f, $R_f \in \mathbb{R}^{2\times 3}$ is the first two rows of the orthographic projection matrix at time f, and $\mathbf{d}_f \in \mathbb{R}^{3\times 1}$ is a displacement vector from the origin of the local coordinate system to the origin of the camera coordinate system at time f.

Given 2D feature points of the non-rigid object tracked from an uncalibrated monocular video with F frames, $\{(u_f^p, v_f^p)^T | p = 1,\ldots,P; f = 1,\ldots,F\}$. Our goal is to recover the object's 3D structures, $\{\mathbf{s}_f^p = (x_f^p, y_f^p, z_f^p)^T | p = 1,\ldots,P; f = 1,\ldots,F\}$, and its rotations relative to the camera, $\{R_f | f = 1,\ldots,F\}$ (see Fig. 2 for an illustration). In the remainder of this paper, the relative rotations between non-rigid objects and the camera is called "camera rotation" for brevity. We assume that the camera rotation is sufficient for the successful recovery of a non-rigid object's 3D structures.

C. Background of Non-Rigid Factorization

Clearly, without any *a priori* information and constraints, the above problem would be under-constrained and could not be solved. In the low-rank shape model proposed by Bregler *et al.* [2], the 3D structures of non-rigid objects are assumed to be represented as a linear combination of a small number K of 3D basis shapes (see Fig. 1 for an illustration),

$$S_f = \sum_{i=1}^{K} c_f^i B_i, f = 1,\ldots,F \tag{2}$$

where $S_f = (\mathbf{s}_f^1,\ldots,\mathbf{s}_f^P) \in \mathbb{R}^{3\times P}$ is the 3D structure of a non-rigid object at frame f, $B_i \in \mathbb{R}^{3\times P}$ is the ith shape basis, and c_f^i is the ith shape coefficient of S_f.

Instead of using the shape basis representation, Akhter *et al.* [13] proposed the use of a set of basis trajectories to linearly express trajectories of a non-rigid object. Specifically, they defined the x, y and z trajectories of the pth point of a non-rigid object as $\mathbf{t}_x(p) = (x_1^p,\ldots,x_F^p)^T$, $\mathbf{t}_y(p) = (y_1^p,\ldots,y_F^p)^T$ and $\mathbf{t}_z(p) = (z_1^p,\ldots,z_F^p)^T$. Then, a shape matrix $S^* \in \mathbb{R}^{F\times 3P}$ of the non-rigid object can be formed by stacking all trajectories vertically

$$S^* = \begin{pmatrix} x_1^1 & \cdots & x_1^P & y_1^1 & \cdots & y_1^P & z_1^1 & \cdots & z_1^P \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_F^1 & \cdots & x_F^P & y_F^1 & \cdots & y_F^P & z_F^1 & \cdots & z_F^P \end{pmatrix}. \tag{3}$$

From (2), the 3D structures of non-rigid objects are constrained to lie within a linear space spanned by K shape bases. It follows that the shape matrix S^* should have rank K. Thus, one can obtain

$$\begin{aligned} \mathbf{t}_x(p) &= \sum_{j=1}^{K} a_j^p(x)\,\boldsymbol{\theta}^j, p = 1,\ldots,P \\ \mathbf{t}_y(p) &= \sum_{j=1}^{K} a_j^p(y)\,\boldsymbol{\theta}^j, p = 1,\ldots,P \\ \mathbf{t}_z(p) &= \sum_{j=1}^{K} a_j^p(z)\,\boldsymbol{\theta}^j, p = 1,\ldots,P \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}^j \in \mathbb{R}^{F\times 1}$ is the jth trajectory basis, and $a_j^p(x)$, $a_j^p(y)$ and $a_j^p(z)$ are the jth trajectory coefficients of $\mathbf{t}_x(p)$, $\mathbf{t}_y(p)$ and $\mathbf{t}_z(p)$, respectively.

Akhter *et al.* [13] proved that (2) and (4) are dual to each other, i.e., the bases in one are equivalent to the coefficients in the other and vice versa. Therefore, the representations of these two models for a given non-rigid object are equal. Additionally, by exploiting the inherent temporal smoothness in trajectories of most naturally deforming objects, the trajectory bases in (4) can be predefined as DCT bases. In the remainder of this paper, we set $\boldsymbol{\theta}^j$ in (4) as the jth DCT basis, and its ith component is denoted by θ_i^j as follows:

$$\theta_i^j = \frac{\sigma_j}{\sqrt{F}} \cos\left(\frac{\pi(2i-1)(j-1)}{2F}\right), i = 1,\ldots,F \tag{5}$$

where $\sigma_1 = 1$ and $\sigma_j = \sqrt{2}$ for $j \geq 2$.

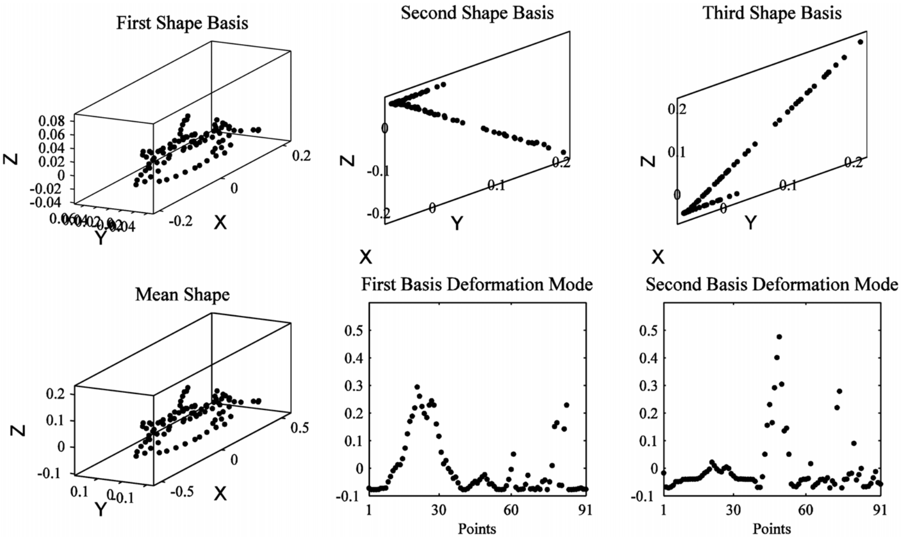


Fig. 3. First row: the shape bases of the "Shark" sequence of [9]. The shape bases are computed by PCA. Second row: mean shape and basis deformation modes of our model for the "Shark" sequence.

D. Motivation

In the low-rank shape model in (2), the deformation of a non-rigid object is assumed to be non-degenerate, i.e., all of its shape bases are assumed to be of full rank 3. However, in practice, the degenerate deformations are deformed not only with non-degenerate shape bases of full rank 3 but also with degenerate shape bases of rank less than 3. The degenerate shape bases usually correspond to 1D or 2D deformations of the non-rigid object. The first row in Fig. 3 illustrates the shape bases of the "Shark" sequence of [9]. In this sequence, the shark's deformation occurs only in a 2D plane and is thus degenerate. As is visible, the "Shark" sequence has 3 shape bases. The first basis, B_1, is a non-degenerate shape basis of full rank 3. The second and third bases, B_2 and B_3, both correspond to 2D deformations of the shark and are thus degenerate shape bases.

Under degenerate deformations, because of the non-degenerate assumption, the low-rank shape model in (2) actually uses the non-degenerate shape basis to approximate degenerate shape bases and introduces extra DoFs. In addition, the low-rank trajectory model in (4) also lacks the ability to compactly represent degenerate deformations because it is dual to the low-rank shape model in (2). During reconstruction, these extra DoFs are usually not constrained by input data and will end up fitting the non-negligible noise in real applications. As a result, the results of NRSfM algorithms based on (2) or (4) will usually contain spurious shape deformations and substantial reconstruction errors. Furthermore, when modeling non-rigid objects with higher degrees of degeneracy, (2) and (4) will be more unreliable; thus, the NRSfM algorithms based on these two models are prone to greater errors.

The main objective of this paper is to solve the degenerate deformation problem in two steps. First, we propose a new low-rank shape deformation model that has the ability to compactly represent 3D structures of degenerate deformations. An illustration of our model is shown in Fig. 1. Second, based on the proposed model, we propose a new NRSfM algorithm to recover camera rotations and 3D structures of non-rigid objects. Fig. 4 shows the flowchart of our NRSfM algorithm.

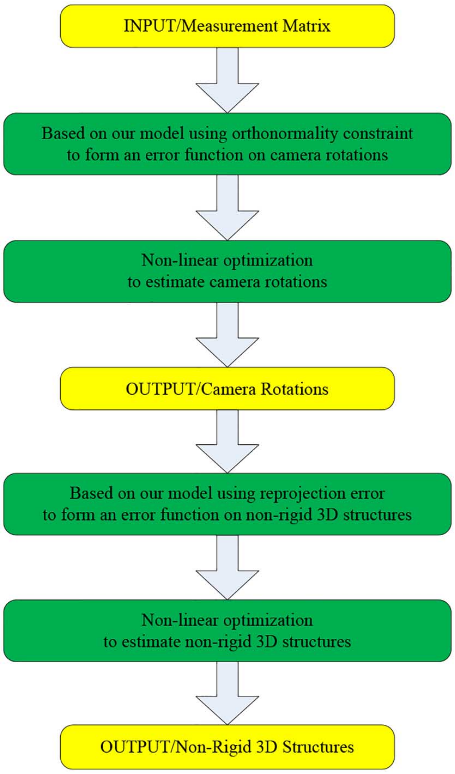

INPUT/Measurement Matrix
Based on our model using orthonormality constraint to form an error function on camera rotations
Non-linear optimization to estimate camera rotations
OUTPUT/Camera Rotations
Based on our model using reprojection error to form an error function on non-rigid 3D structures
Non-linear optimization to estimate non-rigid 3D structures
OUTPUT/Non-Rigid 3D Structures

Fig. 4. Flowchart of our NRSfM algorithm.

IV. New Low-Rank Shape Deformation Model

In this section, we propose a novel low-rank shape deformation model to compactly represent the 3D structures of a non-rigid object with degenerate deformations. A matrix is first formed by horizontally stacking the 3D structure of this non-rigid object at each frame

$$S = \begin{pmatrix} x_1^1 & \cdots & x_1^P \\ y_1^1 & \cdots & y_1^P \\ z_1^1 & \cdots & z_1^P \\ \vdots & \ddots & \vdots \\ x_F^1 & \cdots & x_F^P \\ y_F^1 & \cdots & y_F^P \\ z_F^1 & \cdots & z_F^P \end{pmatrix}. \tag{6}$$

For simplicity and without confusion, we also name $S \in \mathbb{R}^{3F\times P}$ in (6) the shape matrix of the non-rigid object. Then, by utilizing the trajectory model in (4), S can be factorized as follows:

$$S = \left(\boldsymbol{\theta}^1 \otimes I_3 \cdots \boldsymbol{\theta}^K \otimes I_3\right) \underbrace{\begin{pmatrix} a_1^1(x) & \cdots & a_1^P(x) \\ a_1^1(y) & \cdots & a_1^P(y) \\ a_1^1(z) & \cdots & a_1^P(z) \\ \vdots & \ddots & \vdots \\ a_K^1(x) & \cdots & a_K^P(x) \\ a_K^1(y) & \cdots & a_K^P(y) \\ a_K^1(z) & \cdots & a_K^P(z) \end{pmatrix}}_{A}. \tag{7}$$

By partitioning the matrix A in (7) into K sub-matrices, $A = (A_1^T \quad \cdots \quad A_K^T)^T, A_i \in \mathbb{R}^{3\times P}$, (7) can be rewritten as follows:

$$S = \begin{pmatrix} \theta_1^1 I_3 & \cdots & \theta_1^K I_3 \\ \vdots & \ddots & \vdots \\ \theta_F^1 I_3 & \cdots & \theta_F^K I_3 \end{pmatrix} \begin{pmatrix} A_1 \\ \vdots \\ A_K \end{pmatrix}. \tag{8}$$

Subsequently, from (8), the 3D structure of the non-rigid object at each frame can be represented as follows:

$$S_f = \sum_{i=1}^{K} \theta_f^i A_i, f = 1, \ldots, F. \tag{9}$$

According to the duality between (2) and (4), it can be concluded that the role of A_i in (9) is identical to that of B_i in (2). It follows that A_i is the ith shape basis of the non-rigid object, and the components of the vectors $\boldsymbol{\theta}^j$ are the corresponding shape coefficients. Under degenerate deformations, suppose that of K shape bases k_1 bases are of rank 1, k_2 are of rank 2, and k_3 are of rank 3. Then, the matrix A is of rank $k_d = 3k_3 + 2k_2 + k_1$. From (7), we can see that the rank of S should also be k_d.

Next, from (8), we can obtain

$$\begin{pmatrix} A_1 \\ \vdots \\ A_K \end{pmatrix} = \begin{pmatrix} \theta_1^1 I_3 & \cdots & \theta_F^1 I_3 \\ \vdots & \ddots & \vdots \\ \theta_1^K I_3 & \cdots & \theta_F^K I_3 \end{pmatrix} \begin{pmatrix} S_1 \\ \vdots \\ S_F \end{pmatrix}. \tag{10}$$

Then, noting that $\theta_f^1 = 1/\sqrt{F}, f = 1, \ldots, F$ from (5), A_1 is

$$A_1 = \theta_1^1 S_1 + \cdots + \theta_F^1 S_F = \sqrt{F}\bar{S} \tag{11}$$

where $\bar{S} = (\sum_{f=1}^{F} S_f)/F \in \mathbb{R}^{3\times P}$ is the mean shape of the non-rigid object over F frames. Subsequently, by combining (7) and (11), we can obtain

$$\begin{pmatrix} S_1 - \bar{S} \\ \vdots \\ S_F - \bar{S} \end{pmatrix} = (\,\boldsymbol{\theta}^2 \otimes I_3 \quad \cdots \quad \boldsymbol{\theta}^K \otimes I_3\,) \begin{pmatrix} A_2 \\ \vdots \\ A_K \end{pmatrix}. \tag{12}$$

From (11) and (12), we observe that the first shape basis A_1 and the other shape bases $A_i, i = 2, \ldots, K$ have different meanings in terms of modeling the non-rigid object. Specifically, A_1 represents the mean shape of the non-rigid object, while $A_i, i = 2, \ldots, K$ can be interpreted as the deformation components of A_1. Generally, the mean shape of a deforming object over a certain time period is approximately the same as its shape at rest. Thus, A_1 is of full rank 3, and the case of rank deficiency only exists in $A_i, i = 2, \ldots, K$. Therefore, we set $\hat{A} = (A_2^T \quad \cdots \quad A_K^T)^T \in \mathbb{R}^{(3K-3)\times P}$, and the matrix $\hat{A}$ is of rank $k_d - 3$.

By utilizing SVD, $\hat{A}$ can be further decomposed into a sub-unitary matrix $G \in \mathbb{R}^{(3K-3)\times(k_d-3)}$, a diagonal matrix $D \in \mathbb{R}^{(k_d-3)\times(k_d-3)}$, and a sub-unitary matrix $V \in \mathbb{R}^{P\times(k_d-3)}$

$$\hat{A} = GDV^T. \tag{13}$$

Let $\hat{A}' \in \mathbb{R}^{(k_d-3)\times P}$ denote DV^T; then, $\hat{A} = G\hat{A}'$. By denoting the ith row of $\hat{A}'$ as $\hat{\mathbf{a}}_i'$, (12) can be reformulated as

$$\begin{pmatrix} S_1 - \bar{S} \\ \vdots \\ S_F - \bar{S} \end{pmatrix} = (\,\boldsymbol{\theta}^2 \otimes I_3 \quad \cdots \quad \boldsymbol{\theta}^K \otimes I_3\,)\, G \begin{pmatrix} \hat{\mathbf{a}}_1' \\ \vdots \\ \hat{\mathbf{a}}_{k_d-3}' \end{pmatrix}. \tag{14}$$

From (14), the 3D structure of the non-rigid object at each frame can be represented as

$$S_f = \bar{S} + \sum_{i=1}^{k_d-3} \left(\sum_{j=1}^{K-1} \theta_f^{j+1} \hat{\mathbf{g}}_j^i \right) \hat{\mathbf{a}}_i, f = 1, \ldots, F \tag{15}$$

where $\hat{\mathbf{g}}_j^i = (g_j^{3i-2}, g_j^{3i-1}, g_j^{3i})^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$ and g_j^i is the ith row and the jth column element of G.

Equation (15) is our proposed novel low-rank shape deformation model for degenerate deformations. Fig. 1 illustrates an example of our model as applied to the "Shark" sequence of [9]. Our model describes the 3D structures of a deforming object as the mean shape plus the associated deformations. Furthermore, in our model, the deformation component of a non-rigid 3D object is represented as a linear combination of a few basis deformation modes, $\hat{\mathbf{a}}_i', i = 1, \ldots, k_d - 3$. By modeling degenerate shape bases with a combination of these 1D vectors $\hat{\mathbf{a}}_i'$, our model provides the ability to compactly represent 3D structures of degenerate deformations. The second row in Fig. 3 shows the mean shape and basis deformation modes of our model for the "Shark" sequence of [9]. In particular, when S denotes a non-degenerate deformation, its rank k_d will equal $3K$. At this time, G will become an identical matrix, and our model will be equivalent to the low-rank trajectory model in (4).

V. NRSfM With Low-Rank Shape Deformation Model

Given 2D inputs, all we need to compute are the camera rotations, the matrix G, the mean shape $\bar{S}$ and the basis deformation modes $\hat{\mathbf{a}}'_i$. In this section, we first estimate the camera rotations, $\{R_f | f = 1, \ldots, F\}$, and compute G, $\bar{S}$, and $\hat{\mathbf{a}}'_i$. Then, the 3D structures of the non-rigid object can be recovered by our model in (15).

A. Estimating the Camera Rotations

First, by stacking all 2D inputs of a non-rigid object horizontally, a measurement matrix $W \in \mathbb{R}^{2F \times P}$ is formed as follows:

$$W = \begin{pmatrix} u_1^1 & \cdots & u_1^P \\ v_1^1 & \cdots & v_1^P \\ \vdots & \ddots & \vdots \\ u_F^1 & \cdots & u_F^P \\ v_F^1 & \cdots & v_F^P \end{pmatrix}. \tag{16}$$

Following the standard framework of [2], we then compute a mean column vector of W and subtract it from each column of W to obtain a registered measurement matrix $\hat{W}$. In the case of incomplete 2D input we use the Column Space Fitting algorithm in [14], referred to as CSF0, to fill in the missing entries. According to (1), $\hat{W}$ can be decomposed as the product of a camera matrix $R \in \mathbb{R}^{2F \times 3F}$ and a shape matrix S

$$\hat{W} = RS = \begin{pmatrix} R_1 & & \\ & \ddots & \\ & & R_F \end{pmatrix} \begin{pmatrix} x_1^1 & \cdots & x_1^P \\ y_1^1 & \cdots & y_1^P \\ z_1^1 & \cdots & z_1^P \\ \vdots & \ddots & \vdots \\ x_F^1 & \cdots & x_F^P \\ y_F^1 & \cdots & y_F^P \\ z_F^1 & \cdots & z_F^P \end{pmatrix}. \tag{17}$$

By exploiting our model in (15), $\hat{W}$ can be further decomposed into the product of $\Lambda \in \mathbb{R}^{2F \times k_d}$ and $\tilde{A} \in \mathbb{R}^{k_d \times P}$

$$\hat{W} = \underbrace{R\left(\boldsymbol{\theta}^1 \otimes I_3 \cdots \boldsymbol{\theta}^K \otimes I_3\right) \begin{pmatrix} I_3 & 0 \\ 0 & G \end{pmatrix}}_{\Lambda} \underbrace{\begin{pmatrix} \sqrt{F}\bar{S} \\ \hat{A}' \end{pmatrix}}_{\tilde{A}}. \tag{18}$$

Under the assumption of non-degenerate camera motion, we can conclude from (17) that the rank of $\hat{W}$ equals the rank of S, i.e., k_d. Thus, (18) is a full-rank factorization of $\hat{W}$. By utilizing SVD, we can obtain another full-rank factorization of $\hat{W}$ as $\hat{W} = \Lambda' \tilde{A}'$. There exists a non-singular matrix $Q \in \mathbb{R}^{k_d \times k_d}$ satisfying

$$\Lambda' Q = \Lambda, Q^{-1}\tilde{A}' = \tilde{A}. \tag{19}$$

Let us denote the first triple columns of Q by $\mathbf{q}^{123}$. From (18) and (19), we have

$$\Lambda' \mathbf{q}^{123} = \begin{pmatrix} \theta_1^1 R_1 \\ \vdots \\ \theta_F^1 R_F \end{pmatrix}. \tag{20}$$

By exploiting the orthonormality of the camera rotation matrices $R_f, f = 1, \ldots, F$, an error function on $\mathbf{q}^{123}$ can be formed as follows:

$$\sum_{f=1}^{F} \|\Lambda'_{2f-1:2f}\mathbf{q}^{123}(\mathbf{q}^{123})^T(\Lambda'_{2f-1:2f})^T - \left(\theta_f^1\right)^2 I_2\|_F^2 \tag{21}$$

where $\Lambda'_{2f-1:2f}$ is the two rows of Λ' at positions $2f-1$ and $2f$. Equation (21) is a quartic polynomial on $\mathbf{q}^{123}$. We utilize the Levenberg-Marquardt (LM) algorithm in [21] to compute $\mathbf{q}^{123}$ and subsequently compute the camera rotation matrices with (20).

B. Estimating the Non-Rigid 3D Structures

In this subsection, an error function of the 3D structures of the non-rigid object is first designed. Then, an iterative nonlinear optimization algorithm is presented to minimize the error function and estimate the non-rigid 3D structures. We define a mask matrix $\Gamma \in \mathbb{R}^{2F \times P}$ to indicate which entries in the measurement matrix W are missing, i.e., Γ has entries of only 1 (for observed entries in W) or 0 (for missing entries in W). Only the observed entries in W are used to estimate the non-rigid 3D structures. Based on the Euclidean distance between the input 2D feature points and the estimated feature locations from (18), an error function of G, $\bar{S}$, and $\hat{\mathbf{a}}'_i$ can be formed as follows:

$$\left\| \Gamma \odot \left(\hat{W}R\left(\left(\boldsymbol{\theta}^1 \cdots \boldsymbol{\theta}^K\right) \otimes I_3\right) \begin{pmatrix} I_3 & 0 \\ 0 & G \end{pmatrix} \begin{pmatrix} \sqrt{F}\bar{S} \\ \hat{A}' \end{pmatrix} \right) \right\|_F^2. \tag{22}$$

Through minimizing (22), G, $\bar{S}$ and $\hat{\mathbf{a}}'_i$ can be computed, and subsequently, the non-rigid 3D structures can be recovered by our model in (15). Next, noting that $\hat{A} = (\sqrt{F}\bar{S}^T, \hat{A}'^T)^T$, we recast the minimization of (22) as a bilinear optimization problem on G and $\hat{A}$. To solve this problem, we first introduce several notations to rewrite the error function in (22) in a simpler form without Γ. Without loss of generality, the number of observed entries in the pth column of $\hat{W}$ is set to $2F^P (F^P \leq F)$, as denoted by $\mathbf{w}^p \in \mathbb{R}^{2F^P \times 1}$. Then, a row-truncated identity matrix $\Pi^p \in \mathbb{R}^{2F^P \times 2F}$ is defined such that $\Lambda^p = \Pi^p \Lambda \in \mathbb{R}^{2F^P \times k_d}$ includes the rows in Λ that correspond to the rows of entries in $\mathbf{w}^p$. We use $\tilde{\mathbf{a}}^p \in \mathbb{R}^{k_d \times 1}$ to denote the pth column of $\tilde{A}$. Therefore, (22) can be rewritten as follows:

$$f\left(G, \tilde{A}\right) = \sum_{p=1}^{P} \|\mathbf{w}^p - \Lambda^p \tilde{\mathbf{a}}^p\|_2^2 = \sum_{p=1}^{P} (\mathbf{r}^p)^T \mathbf{r}^p \tag{23}$$

where $\mathbf{r}^p = \mathbf{w}^p - \Lambda^p \tilde{\mathbf{a}}^p$ is the residual between the observed and estimated 2D trajectory of the pth feature point.

For computing G and $\tilde{A}$, we present an iterative nonlinear optimization algorithm, as described in Algorithm 1, to minimize (23); its initialization is given below. In each iteration of Algorithm 1, we alternately fix one factor and update another factor to minimize the error function in (23). Therefore, each iteration can reduce the error in (23); thus, the optimum solution around the initialization values can be obtained when the algorithm converges. Moreover, to achieve both convergence and computational efficiency, we adopt the Gaussian-Newton and least squares methods to compute G and $\tilde{A}$, respectively.

Algorithm 1 Iterative non-linear optimization algorithm for minimizing (23).

```
    Input: The initial estimation of $G$, $\delta = 10^{-4}$.
1. Repeat
2.     Fix $G$, compute $\tilde{A}$ with (24).
3.     Fix $\tilde{A}$, compute gradient $\mathbf{v} = df/dG$ and Hessian
       $H = d^2f/dG^2$ with (25).
4.     Repeat
5.         $\delta = 10\delta$.
6.         Compute $G$ with $vec(G) = (H + \delta I)^{-1}\mathbf{v}$.
7.     Until $f(G - G, \tilde{A}) < f(G, \tilde{A})$.
8.     $G = G - G, \delta = 0.01\delta$.
9.     Orthogonalize the columns of $G$
10. Until convergence
    Output: The optimized $G$ and $\tilde{A}$.
```

In Algorithm 1, I is an identity matrix of order $((k_d - 3)(3K - 3))$. Equations (24) and (25) have the following form:

$$vec(\tilde{A}) = (\Psi^T\Psi)^{-1}\Psi^T vec(\hat{W}) \tag{24}$$

where $\Psi \in \mathbb{R}^{(2\sum_{p=1}^{P} F^p)\times k_d P}$ is a block diagonal matrix that is formed by $\Lambda^p, p = 1, \ldots, P$. $vec(\hat{W}) = ((\mathbf{w}^1)^T, \ldots, (\mathbf{w}^P)^T)^T \in \mathbb{R}^{(2\sum_{p=1}^{P} F^p)\times 1}$

$$\mathbf{v} = -2\sum_{p=1}^{P}(J^p)^T\mathbf{r}^p, H = 2\sum_{p=1}^{P}(J^p)^T J^p \tag{25}$$

where $J^p = ((\tilde{\mathbf{a}}^p_{4:K_d})^T \otimes (\Pi^p R\Omega)) \in \mathbb{R}^{2F^p\times((k_d-3)(3K-3))}$ is the Jacobian matrix of the residual $\mathbf{r}^p$ with respect to G: $d\mathbf{r}^p = -J^p vec(dG)$. $\tilde{\mathbf{a}}^p_{4:K_d} \in \mathbb{R}^{(k_d-3)\times 1}$ is the fourth to the last rows of $\tilde{\mathbf{a}}^P$. $\Omega = (\boldsymbol{\theta}^2 \otimes I_3 \quad \cdots \quad \boldsymbol{\theta}^K \otimes I_3) \in \mathbb{R}^{3F\times 3(K-3)}$.

After G, $\bar{S}-$, and $\hat{\mathbf{a}}'_i$ are computed, the 3D structures of the non-rigid object can be recovered by our model in (15).

Initialization. By denoting Ω as $(\boldsymbol{\theta}^2 \otimes I_3 \quad \cdots \quad \boldsymbol{\theta}^K \otimes I_3)$, (18) can be written as

$$\hat{W} = R\big(\bar{S}^T \cdots \bar{S}^T\big)^T + R\Omega G\hat{A}'. \tag{26}$$

Based on the assumption that the mean shape is the dominant rigid component of a non-rigid 3D object, $\bar{S}$ can be initialized as the best rank-3 approximation of $\hat{W}$. Specifically, we approximate $\bar{S}$ as $((R_1{}^T \quad \cdots \quad R_F{}^T)^T)^+\hat{W}$, where $+$ denotes the Moore-Penrose pseudo-inverse [22]. Then, $G\tilde{A}'$ can be approximated as

$$G\hat{A}' \approx (\Omega)^+R^+(\hat{W} - R(\bar{S}^T \cdots \bar{S}^T)^T). \tag{27}$$

Finally, we perform SVD on $G\tilde{A}'$: $O_1\Sigma O_2^{\mathrm{T}}$, and initialize G with the first $k_d - 3$ columns of O_1.

VI. Experiments

A. Experimental Setup

In this section, we first quantitatively evaluate our method as applied to synthetic and motion capture data. In the quantitative experiments, the performance of our method is compared with that of the state-of-the-art NRSfM algorithms, which include the following: (1) the 3D point trajectory approach (PTA) [13]; (2) the Column Space Fitting method, which explicitly models complementary rank-3 spaces (CSF2) [14]; (3) the Kernel Shape Trajectory Approach (KSTA) [15]; (4) the block matrix method (BMM) [10]; (5) the Metric Projections (MP) algorithm [8]; (6) the EM algorithm based on the linear dynamics model (EM-LDS) [6]; and (7) the sequential NRSfM algorithm based on the 3D-implicit low-rank shape model (SLR) [18]. We do not consider the shape NRSfM algorithm in [17] because its results are highly inferior to those of PTA. In addition, we qualitatively evaluate the performance of our method.

Following the methodology in [13]–[15], we perform PTA, CSF2, KSTA, BMM, MP and EM-LDS with different values of $K \in \{2, \ldots, 13\}$ and report the best results. The number of DCT bases in CSF2 and KSTA is set to $0.1F$ for the "Shark", "Handshake" and "High-five" sequences. And, it is set to $0.3F$ for the "Face1" and "Face2" sequences. The dimension of the shape space in KSTA is set to $h = 2$. To provide a fair comparison, we perform our method with different values of $k_d \in \{4, \ldots, 39\}$ and report the best results. Another parameter,, in our method is set to $0.1F$ for the "Shark", "Handshake" and "High-five" sequences. And, K is set to $0.3F$ for the "Face1" and "Face2" sequences. In SLR, the width of the sliding window is set to 5 frames, the reprojection threshold is set to 0.018 pixels, and the number of starting frames used to estimate the mean shape is set to $0.2F$.

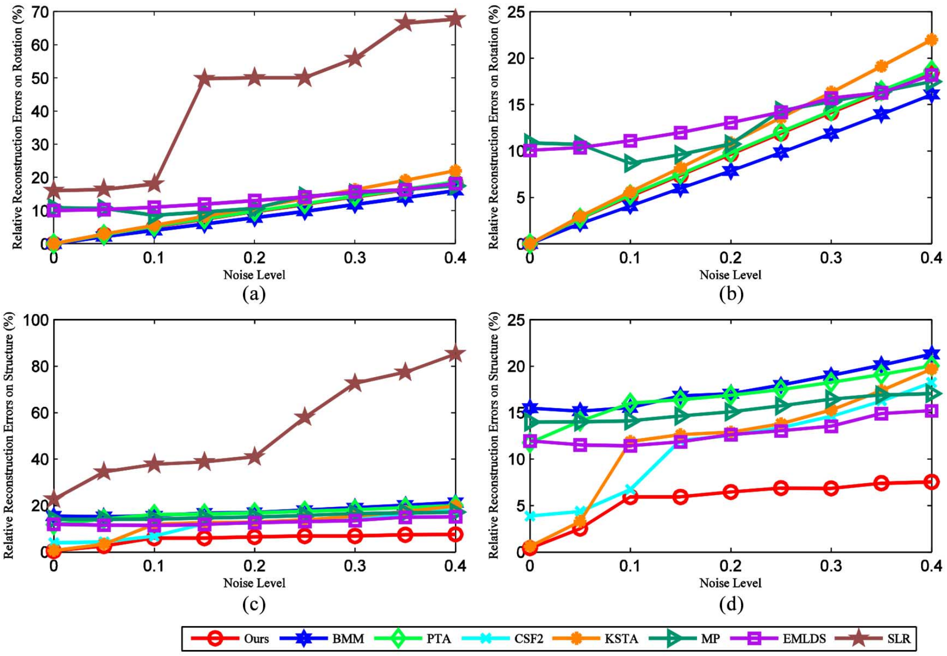
To evaluate the reconstruction quality of the NRSfM algorithms, we first compute the relative reconstruction errors for the camera rotations and 3D structures at each frame and subsequently record the average value of the errors over all frames. We thus set the error metrics as follows:

$$\begin{aligned} e_{3D} &= \frac{1}{F}\sum_{f=1}^{F}\frac{\sum_{p=1}^{p}\|\mathbf{s}^p_f - \check{\mathbf{s}}^p_f\|_F}{\sum_{p=1}^{p}\|\mathbf{s}^p_f\|_F}, \\ e_R &= \frac{1}{F}\sum_{f=1}^{F}\frac{\|R_f - \check{R}_f\|_F}{\|R_f\|_F} \end{aligned} \tag{28}$$

where e_R and e_{3D} are the average relative reconstruction errors for the camera rotations and 3D structures, respectively; R_f and $\check{R}_f$ are the ground-truth and reconstructed camera rotation matrices at time f, respectively; and S_f and $\check{S}_f$ are the ground-truth and reconstructed 3D structures of the non-rigid object at time f, respectively.

B. Quantitative Evaluation Using Synthetic Data

Here, we use the "Shark" sequence (240/91) from [9] to perform our quantitative evaluation, where (F/P) denotes the number of frames (F) and the number of feature points (P). Based on Fig. 3, the deformation in this sequence is degenerate



(a) (b) (c) (d)

Fig. 5. Reconstruction errors for the "Shark" sequence when the noise level increases from 0 to 0.4. (a) and (b) show reconstruction errors for camera rotations; (c) and (d) show reconstruction errors for 3-D structures. The reconstruction errors for camera rotations of CSF2 are not plotted in (a) and (b) because they are identical to those of KSTA. (a) and (c) compare all methods discussed here, and (b) and (d) provide close observations for (a) and (c) by excluding SLR.

and deformed with one non-degenerate shape basis and two degenerate shape bases.

We first evaluate the reconstruction accuracies and robustness of the NRSfM algorithms when different levels of noise are added to the registered measurement matrix $\hat{W}$. Noise is assumed to be Gaussian, and its level is computed as the ratio between the Frobenius norm of the noise and the registered measurement, i.e., $||noise||_F/||\hat{W}_F||$. Fig. 5(a) and (c) show the relative reconstruction errors for camera rotations and 3D structures, respectively, of the eight NRSfM algorithms when the noise level increases from 0 to 0.4. The performance of SLR is significantly lower than that of other NRSfM algorithms because SLR addresses a more difficult problem, i.e., incremental 3D reconstruction, compared with the other methods, which are all batch methods [18]. Thus, to provide a better visual comparison, we plot the reconstruction errors of all algorithms expect for SLR in Fig. 5(b) and (d). As Fig. 5 shows, among all of the compared algorithms, the overall performance of our algorithm is the best. Specifically, when the noise level is 0, our method recovers the exact camera rotations and 3D structures with zero error. Furthermore, our method exhibits superior robustness to noise when recovering 3D structures. As a result, even when contaminated by noise as high as level 0.4, the reconstructed 3D structures of our method are still satisfactory and significantly better than those of other NRSfM algorithms (see Fig. 6).

We then evaluate the performance of the NRSfM algorithms when some feature points are missing in some frames. We do not test PTA and BMM because they lack the ability to handle incomplete 2D input. In the field of NRSfM, the missing data case is important to test because it is very common in real tracking that a point cannot be tracked during an entire video sequence. We simulate missing data by randomly discarding a particular percentage of entries in the measurement matrix W. Fig. 7 illustrates relative reconstruction errors of the six NRSfM algorithms when the percent of missing entries in W increases from 0% to 60%. As is visible in the figure, the reconstruction accuracies of our method, CSF2 and KSTA show little variation over all the tested levels of missing entries and are significantly better than those of MP, EM-LDS and SLR.

Finally, we evaluate the performance of the NRSfM algorithms when they are used to recover dynamic scenes with different numbers of degenerate shape bases. Five dynamic scenes are generated, and they are, respectively, composed of 2 to 6 sharks of the "Shark" sequence. In each scene, the sharks move independently along different straight lines in the xz-plane and with different velocities. There are $n+2$ degenerate shape bases in the scene containing n sharks: two bases come from the deformation of the shark itself, and the others refer to the 1D translation of the n sharks. For each scene, we generate synthetic camera rotations that are 2 degrees per frame around the z-axis and subsequently project the 3D data using these rotations to obtain the 2D inputs. Fig. 8(a) and (b) show the relative reconstruction errors of all algorithms when applied to the five dynamic scenes. As expected, our method provides exact reconstructions with zero error for all the tested dynamic scenes, whereas the performances of SLR, BMM, MP and EM-LDS are very poor. Meanwhile, we note that PTA, CSF2 and KSTA also provide reasonable results for all dynamic scenes. The good performances of PTA, CSF2 and KSTA should be attributable to the noise-free CG data of the five sequences. Furthermore, we compare our method with PTA, CSF2 and KSTA when applied to the sequences with an added noise level of 0.01 and show the results in Fig. 8(c) and (d). Even with very small amount of noise, the performances of PTA, CSF2 and KSTA quickly

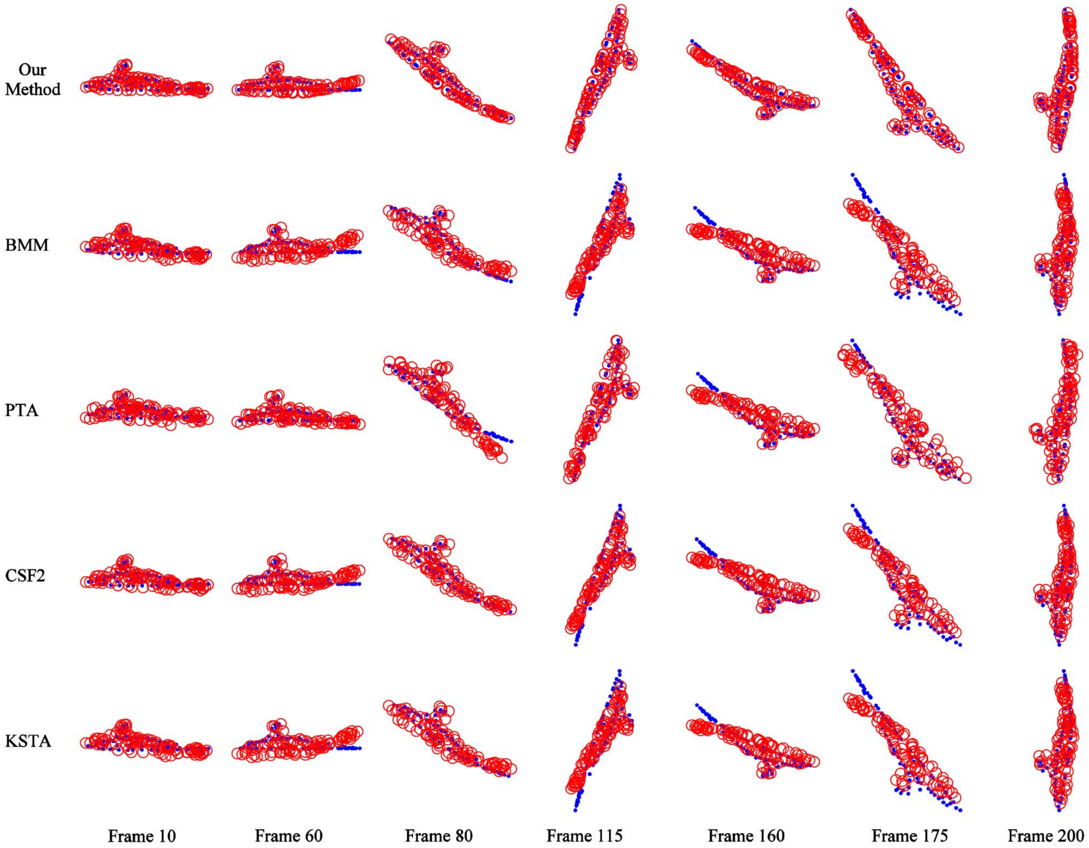
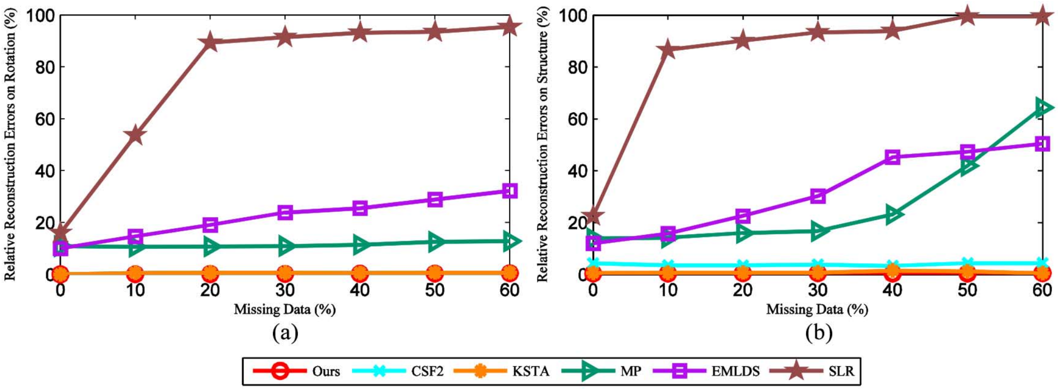


Fig. 6. 3D reconstruction results of the "Shark" sequence when level 0.4 noise is added. The first, second, third, fourth, and fifth rows show the ground-truth (blue dots) and the reconstruction results (red circles) generated by our method, BMM, PTA, CSF2, and KSTA, respectively.



(a)

(b)

Fig. 7. Reconstruction errors for the "Shark" sequence when the percent of missing entries in W increases from 0% to 60%. (a) shows reconstruction errors for camera rotations; (b) shows reconstruction errors for 3-D structures. The reconstruction error for camera rotations of CSF2 is not plotted in (a) because it is identical to that of KSTA. The values of relative errors greater than 100% are truncated to 100% to provide a better visual comparison.

degrade as the number of degenerate shape bases increases. In contrast, our method still provides high-quality reconstruction results for all tested sequences.

C. Quantitative Evaluation Using Motion Capture Data

To test the power of our method when applied to degenerate deformations, we first introduce two typical degenerate deformations from the CMU motion capture database:[1] "Handshake" (303/82) and "High-five" (233/82). These two sequences both describe two people walking independently along straight lines. When they meet, they shake hands or high five each other. Then, two face sequences that have been used in previous studies are tested: "Face1" (316/40) from [6] and "Face2" (74/37) from [8]. As discussed in Section I, facial expressions are close to being degenerate. Among the four motion capture datasets, "Face1" and "Face2" are rotating themselves in the sequences; therefore, we obtain 2D inputs by extracting the x and y coordinates of 3D marker measurements. For the "Handshake" and "High-five" datasets, we generate synthetic camera rotations and project 3D data using these rotations to obtain 2D inputs. Following the

[1] [Online]. Available: http://mocap.cs.cmu.edu

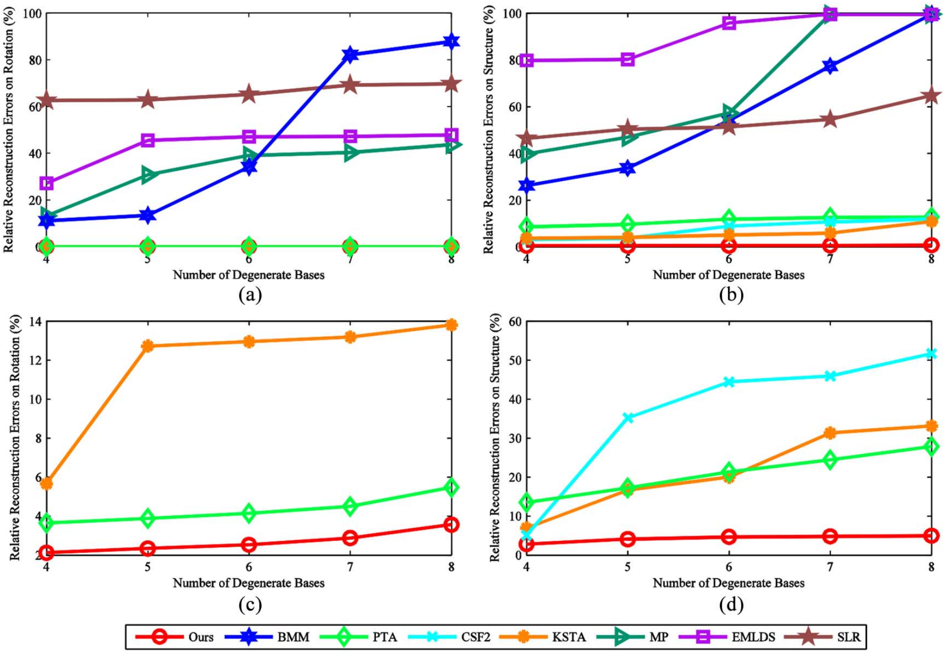


Fig. 8. Reconstruction errors for five dynamic scenes that contain four to eight degenerate shape bases. (a) and (c) show reconstruction errors for camera rotations; (b) and (d) show reconstruction errors for 3-D structures. (a) and (b) show reconstruction errors when the noise level is 0; (c) and (d) show reconstruction errors when the noise level is 0.01. The reconstruction errors for camera rotations of CSF2 are not plotted in (a) and (c) because they are identical to those of KSTA. The values of relative errors greater than 100% are truncated to 100% to provide a better visual comparison.

TABLE I
Quantitative Comparison of Our Method With PTA [13], CSF2[14], KSTA [15], BMM [10], MP [8], EM-LDS [6], and SLR [18] When Applied to Motion Capture Data. (a) and (b), Respectively, Show Comparison on the Original Data and the Original Data Added With Level 0.4 Noise

Method	Face1	Face2	Handshake		High-five	
	e_{3D}(%)	e_{3D}(%)	e_R(%)	e_{3D}(%)	e_R(%)	e_{3D}(%)
Ours	**1.11**	**2.93**	**9.50**	**9.67**	**9.61**	**10.04**
PTA	2.92	8.47	11.78	13.57	15.73	17.21
CSF2	1.93	3.46	11.78	13.11	9.71	11.64
KSTA	2.12	4.26	11.78	14.43	9.71	12.62
BMM	1.82	3.87	**9.09**	14.62	11.21	21.35
MP	2.79	6.59	12.26	16.62	10.44	21.94
EM-LDS	2.26	3.42	17.96	141.10	17.33	128.47
SLR	3.54	3.91	33.02	127.68	48.92	103.15

(a)

Method	Face1	Face2	Handshake		High-five	
	e_{3D}(%)	e_{3D}(%)	e_R(%)	e_{3D}(%)	e_R(%)	e_{3D}(%)
Ours	**12.14**	**10.20**	**31.69**	**18.78**	**31.51**	**21.53**
PTA	18.92	24.50	46.48	39.53	42.91	45.74
CSF2	23.39	29.73	55.47	39.04	58.39	42.54
KSTA	26.70	32.10	55.47	40.21	58.39	44.02
BMM	18.94	25.69	**27.20**	57.02	38.58	49.70
MP	19.51	30.84	**29.31**	46.27	**25.44**	54.97
EM-LDS	34.47	21.40	**26.57**	146.48	**22.67**	138.48
SLR	133.76	55.92	53.02	94.45	55.07	97.06

(b)

methodology in [13], the camera rotation is set to 2 degrees per frame around the z-axis, while the overall camera motion is oscillatory with a pan of ±45 degrees.

Table I shows a quantitative comparison between our method and other NRSfM algorithms when applied to the four motion capture sequences. As observed from Table I(a), when the original data are tested, our method achieves the best overall performance among all the compared algorithms. Specifically, the reconstruction quality of our method for 3D structures is the best, while for camera rotations, it is only slightly inferior to BMM's reconstruction quality for the "Handshake" sequence. It can be further observed from Table I(b) that when a noise of level 0.4 is added, the advantage of our method over other NRSfM algorithms becomes more evident. The average improvement of our method on shape recovery increases from 1.37% to 27.24% for the "Face1" sequence, from 1.92% to 21.25% for the "Face2" sequence, from 39.06% to 47.36% for the "Handshake" sequence, and from 35.12% to 45.97% for the "High-five" sequence. The average improvement of our method on camera rotation recovery increases from 5.88% to 10.24% for the "Handshake" sequence, and from 7.97% to 11.55% for the "High-five" sequence. The results in Table I clearly demonstrate that our method has significant advantages over state-of-the-art NRSfM algorithms when applied to degenerate deformations. From Table I, it can be observed that our reconstruction errors for camera rotations are comparable to but slightly higher than BMM's errors for the "Handshake" sequence. Note that regardless of the addition of noise, our reconstruction accuracies are distinctly better than BMM's accuracies for camera rotations in the "High-five" sequence. The difference between the "Handshake" and "High-five" sequences is that the persons' arm actions in the latter are faster and more substantial, and they result in high-frequency deformations. Therefore, we suspect that BMM's preferable results for camera rotations in the "Handshake" sequence might be attributed to the relatively slower and smoother deformations contained in this sequence. Generally, a slower and smoother deformation can be represented more effectively by linear models; thus, it fits the rank-3 constraint used by BMM to

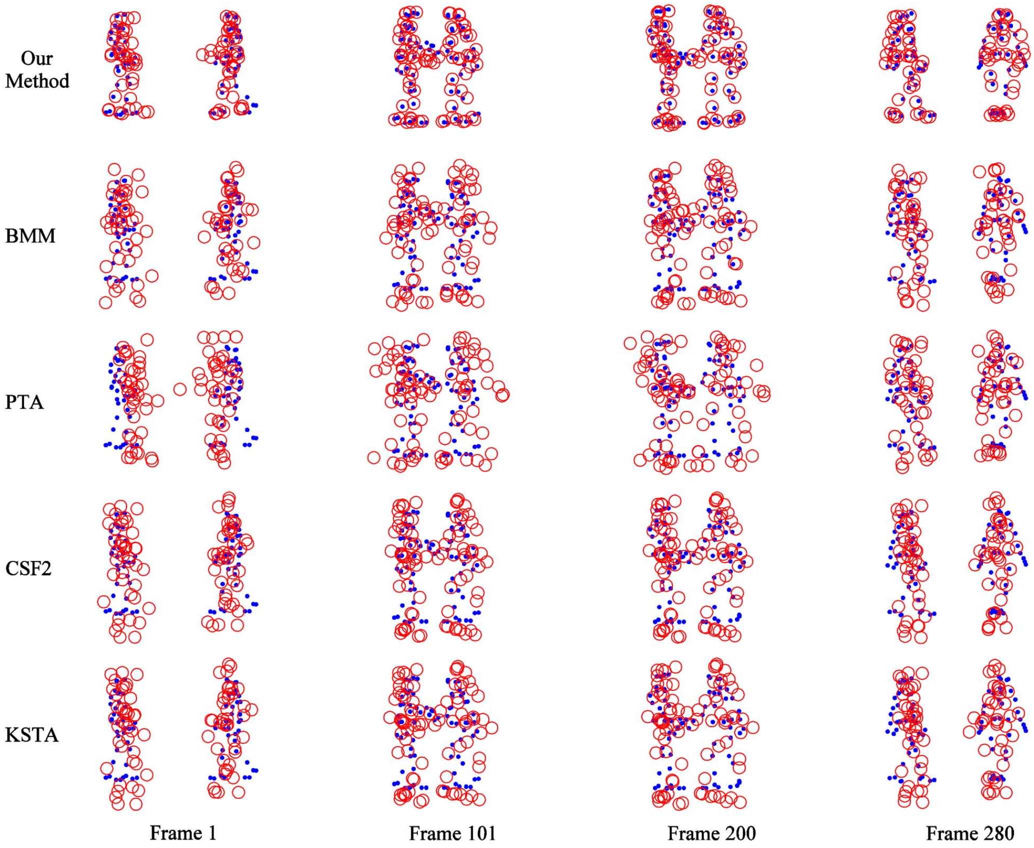
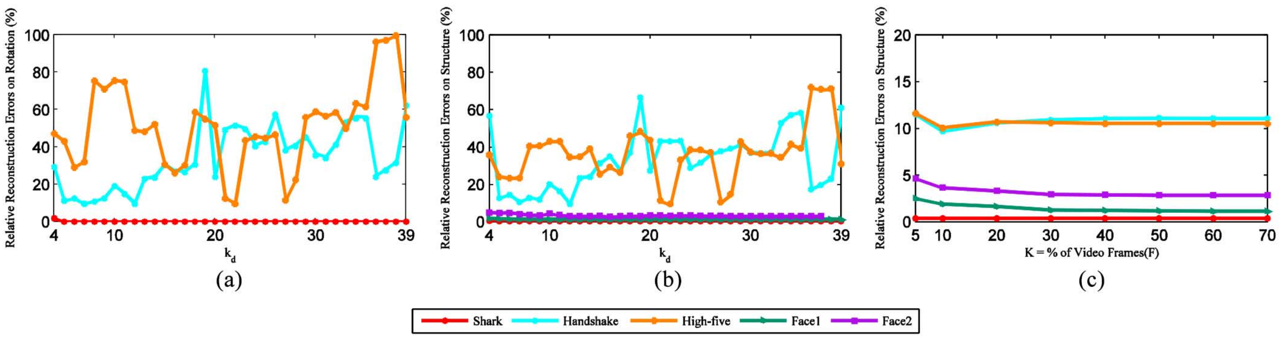


Fig. 9. 3-D reconstruction results of the “Handshake” sequence when level 0.4 noise is added. The first, second, third, fourth, and fifth rows show the ground-truth (blue dots) and the reconstruction results (red circles) generated by our method, BMM, PTA, CSF2, and KSTA, respectively.



(a)

(b)

(c)

Fig. 10. Reconstruction errors of our NRSfM algorithm using different values of K and k_d applied to the synthetic and motion capture data. (a) and (b) show reconstruction errors for camera rotations and 3-D structures, respectively, when k_d increases from 4 to 39; (c) shows reconstruction errors for 3-D structures when K increases from $0.05F$ to $0.7F$.

estimate camera rotations better. From Table I(b), we can also observe that our reconstruction accuracies for camera rotations are slightly inferior to MP’s and EM-LDS’s accuracies for the “Handshake” and “High-five” sequences. The performances of MP and EM-LDS in the estimation of camera rotations are mainly from the iterative refinement between camera rotations and 3D structures [6], [8], while our method does not involve this process for reasons of efficiency.

Compared with other NRSfM algorithms, our method is computationally efficient. Taking the “Face1” sequence (without added noise) as an example, the runtime of our method averaged over 100 runs is 7.23 seconds ($K = 6$, $k_d = 6$), which is slightly slower than the 3.65 seconds of PTA ($K = 5$) and faster than the 8.69 seconds of MP ($K = 5$), 42.40 seconds of EM-LDS ($K = 3$), 86.50 seconds of CSF2 ($d = 100$, $K = 5$), 90.14 seconds of KSTA ($d = 100$, $K = 4$), 198.34 second of BMM ($K = 7$), and 215.43 seconds of SLR. All the time data were collected on a PC laptop with a 2.6 GHz Intel Core i5 processor and 4 GB of RAM.

Fig. 9 shows the 3D reconstruction results of our method, PTA, BMM, CSF2 and KSTA for the “Handshake” sequence when noise with a level of 0.4 is added. The reconstructed 3D structures of our method are significantly better than those of the other NRSfM algorithms.

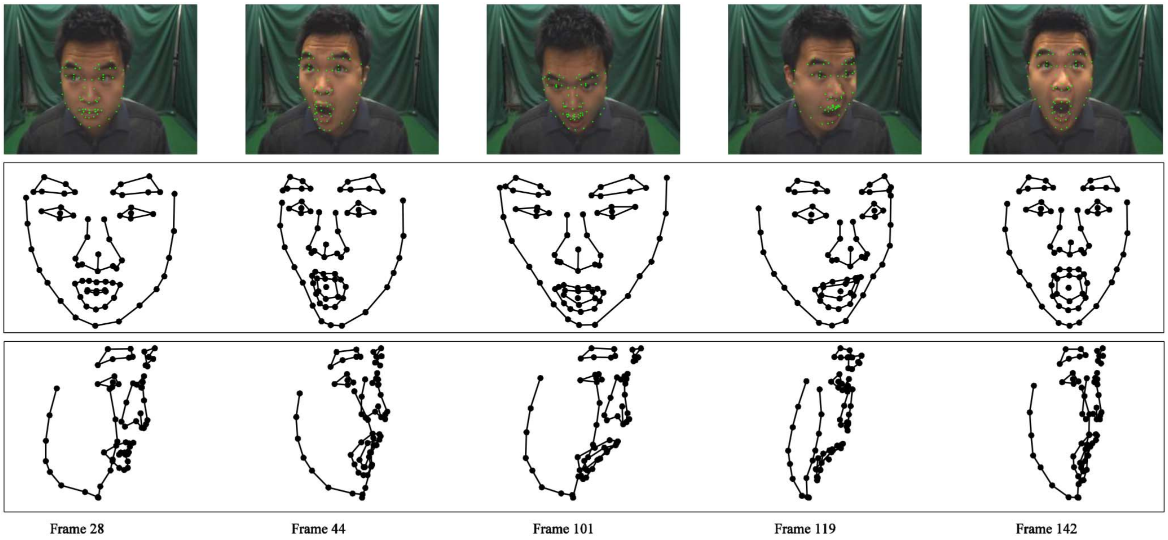


Fig. 11. 3-D reconstruction results of our method on the recorded "Face" sequence. First row: five out of 150 frames of the "Face" sequence. In each frame, the tracked 2-D feature points are overlaid in green. Second and third rows: two orthogonal views of the recovered 3-D structures. The lines are not part of our model; they are shown for visualization purposes only.

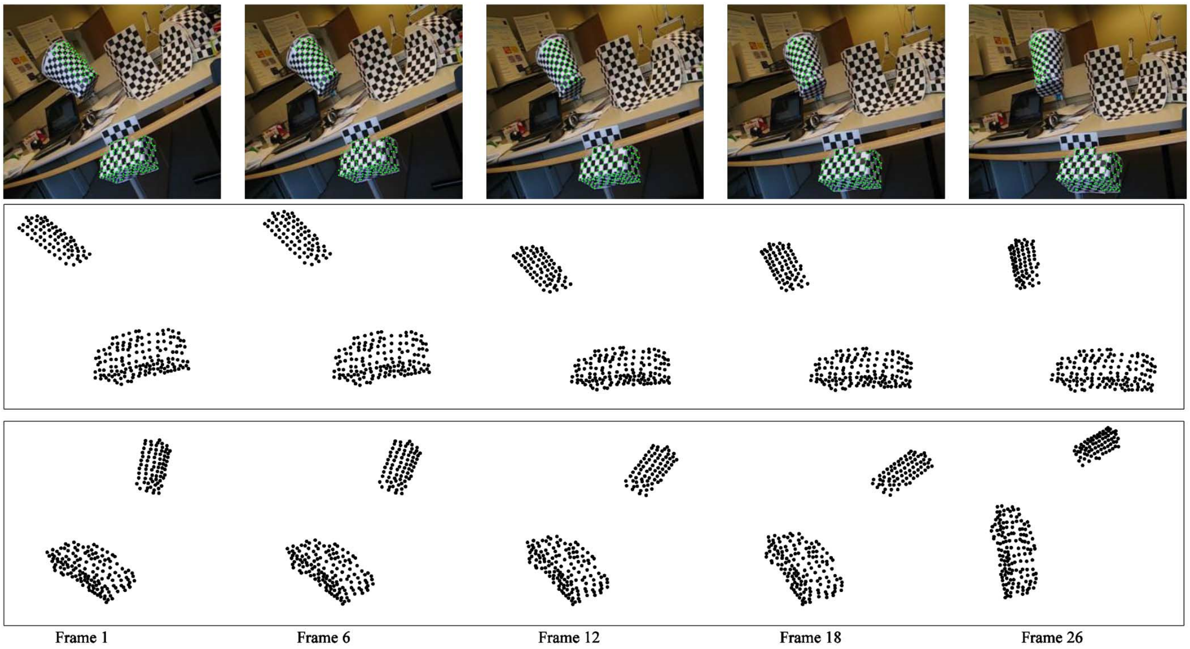


Fig. 12. 3-D reconstruction results of our method on the "1R2TCR" sequence in the Hopkins155 dataset. First row: five out of 26 frames of the "1R2TCR" sequence. In each frame, the 2-D feature points are overlaid in green. Second and third rows: two orthogonal views of the recovered 3-D structures.

D. Quantitative Evaluation on Different Parameters

We next test the relation between the reconstruction quality of our NRSfM algorithm and two parameters of our model. We run our method on the "Shark", "Face1", "Face2", "Handshake" and "High-five" sequences by changing one parameter while keeping the other fixed. Fig. 10 shows our reconstruction errors as a function of the parameters K and k_d. We do not plot the correlation between the estimation errors of the camera rotations and K because the process of estimating camera rotations is independent of K in our method.

From Fig. 10(a) and (b), the reconstruction quality of our method applied to sequences such as "Handshake" and "High-five" is significantly influenced by the choice of k_d. In essence, the parameter k_d equals the rank of the registered measurement matrix $\hat{W}$. This finding may be further exploited to compute k_d using the spatio-temporal regularity of the observed 2D shapes. From Fig. 10(c), we observe that when K is below a particular threshold (e.g., $0.3F$ for "Face1" or $0.1F$ for "Handshake"), the reconstruction accuracies of our method are relatively poor. This is because if the chosen K is too small, our model will not be able to fully capture the data variability. In contrast, when K exceeds the threshold, the reconstruction accuracies of our method remain consistently stable. The robustness should be mainly attributed to the DCT constraint imposed on the deformation coefficients in our model in (15). During the process of estimating the non-rigid 3D structures, the increasing DoFs from larger K will be constrained by a set of predefined DCT basis vectors. As a result, the overfitting problem typically caused by parameter overestimation can be effectively suppressed.

E. Qualitative Evaluation of Real Video Sequences

Finally, we qualitatively evaluate the proposed NRSfM algorithm using two real video sequences. The first sequence is a close-up video of a face that was recorded in our laboratory. The "Face" sequence contains 150 frames, and each frame contains 68 feature points that are tracked by the 2D Active Appearance Model (AAM) [23]. With $K = 18$ and $k_d = 6$, the solution of our method for this sequence has a mean (maximum) 2D reprojection error of 0.0216 (0.5473) pixels. The second sequence is the "1R2TCR" sequence from the Hopkins155 dataset [24], which consists of 26 frames. In this sequence, the deformations formed by two foreground objects are degenerate: one object translates along a straight line and forms a rank-1 degenerate shape basis, and the other object rotates around an axis and approximately forms a rank-2 degenerate shape basis. Therefore, we only consider the two foreground objects, which contain 219 feature points. The 2D coordinates of these feature points have been provided by the author of the Hopkins155 dataset. With $K = 3$ and $k_d = 6$, the solution of our method for the "1R2TCR" sequence has a mean (maximum) 2D reprojection error of 0.0675 (0.4497) pixels. Fig. 11 and 12 show various 3D reconstruction results of the proposed NRSfM algorithm on the "Face" and "1R2TCR" sequences, respectively. Our method produces reasonable reconstructions for these two real video sequences.

VII. Conclusion

In this study, we proposed a new NRSfM algorithm to address the problem of degenerate deformation recovery. In most existing NRSfM algorithms, the deformations of non-rigid objects are assumed to be non-degenerate. However, we find that such an assumption causes these algorithms' reconstructions to be inaccurate and sensitive to noise when applied to degenerate deformations. We thus propose a novel low-rank shape deformation model to represent 3D structures of degenerate deformations. The main advantage of our model over the previous model on degenerate deformations, i.e., the 3D implicit low-rank shape model [18], is that our model exploits the inherent low-rank property of trajectories of non-rigid objects when modeling degenerate deformations and can thus provide a more compact representation. Moreover, our model is the first in batch NRSfM to provide the ability to represent 3D structures of degenerate deformations, whereas the 3D implicit low-rank shape model is proposed to solve the sequential NRSfM problem. Then, based on the proposed low-rank shape deformation model, we formulate the NRSfM problem as two coherent optimization problems and solve them with iterative non-linear optimization algorithms. We perform a series of comparison experiments between our method and state-of-the-art NRSfM algorithms using synthetic and motion capture data. The experimental results illustrate that our method has significant advantages over state-of-the-art NRSfM algorithms in terms of both accuracy and robustness when recovering degenerate deformations. In the future, we plan to relax the assumption of non-degenerate camera motion in our method and aim to reconstruct 3D scenes in which the overall rigid motion of deformable shapes relative to the camera is small. Another future direction is to generalize the camera model in our algorithm for applicability to full perspectives.

References

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, pp. 137–154, 1992.

[2] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2000, vol. 2, pp. 690–696.

[3] J. Xiao and T. Kanade, "A closed-form solution to non-rigid shape and motion recovery," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 196–209.

[4] M. Brand, "A direct method for 3D factorization of nonrigid motion observed in 2D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 2, pp. 122–128.

[5] I. Akhter, Y. Sheikh, and S. Khan, "In defense of orthonormality constraints for nonrigid structure from motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1534–1541.

[6] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 878–892, May 2008.

[7] S. Olsen and A. Bartoli, "Implicit non-rigid structure-from-motion with priors," *J. Math. Imag. Vis.*, vol. 31, no. 2–3, pp. 233–244, 2008.

[8] M. Paladini, A. D. Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito, "Factorization for non-rigid and articulated structure using metric projections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 2898–2905.

[9] V. Rabaud and S. Belongie, "Linear embeddings in non-rigid structure from motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 2427–2434.

[10] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2018–2025.

[11] H. Zhou, X. Li, and A. H. Sadka, "Nonrigid structure-from-motion from 2-D images using Markov chain Monte Carlo," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 168–177, Jun. 2012.

[12] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Nonrigid structure from motion in trajectory space," in *Proc. Neural Inf. Process. Syst.*, 2008, pp. 41–48.

[13] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1442–1456, Jul. 2011.

[14] P. F. U. Gotardo and A. M. Martinez, "Non-rigid structure from motion with complementary rank-3 spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 3065–3072.

[15] P. F. U. Gotardo and A. M. Martinez, "Kernel non-rigid structure from motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 802–809.

[16] I. Khan, "Non-rigid structure-from-motion with uniqueness constraint and low rank matrix fitting factorization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1350–1357, Aug. 2014.

[17] J. Xiao and T. Kanade, "Non-rigid shape and motion recovery: Degenerate deformations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun.–Jul. 2004, vol. 1, pp. 668–675.

[18] M. Paladini, A. Bartoli, and L. Agapito, "Sequential non-rigid structure-from-motion with the 3D-implicit low-rank shape model," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–28.

[19] R. Angst and M. Pollefeys, "A unified view on deformable shape factorizations," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 682–695.

[20] P. F. U. Gotardo and A. M. Martinez, "Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2051–2065, Oct. 2011.

[21] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[22] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd ed. New York, NY, USA: Wiley, 1999.

[23] I. Matthews and S. Baker, “Active appearance models revisited,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[24] R. Tron and R. Vidal, “A benchmark for the comparison of 3D motion segmentation algorithms,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[25] L. Ding and A. M. Martinez, “Modelling and recognition of the linguistic components in american sign language,” *Image Vis. Comput.*, vol. 27, no. 12, pp. 1826–1844, 2009.

[26] F. Rengier, A. Mehndiratta, H. von Tengg-Kobligk, C. M. Zechmann, R. Unterhinninghofen, H.-U. Kauczor, and F. L. Giesel, “3D printing based on imaging data: Review of medical applications,” *Int. J. Comput. Assisted Radiol. Surgery*, vol. 5, no. 4, pp. 335–341, 2010.

[27] J. Zhu, S. C. Hoi, and M. R. Lyu, “Real-time non-rigid shape recovery via active appearance models for augmented reality,” in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 186–197.

[28] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 2, pp. 1470–1477.

[29] P. Ochs, J. Malik, and T. Brox, “Segmentation of moving objects by long term video analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.

[30] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh, “3D reconstruction of a moving point from a series of 2D projections,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 158–171.

[31] A. Zaheer, I. Akhter, M. H. Baig, S. Marzban, and S. Khan, “Multiview structure from motion in trajectory space,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2447–2453.

[32] A. D. Bue, “Adaptive non-rigid registration and structure from motion from image trajectories,” *Int. J. Comput. Vis.*, vol. 103, no. 2, pp. 226–239, 2013.

[33] L. Tao and B. J. Matuszewski, “Non-rigid structure from motion with diffusion maps prior,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1530–1537.

[34] H. Liang, J. Yuan, and D. Thalmann, “Parsing the hand in depth images,” *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1241–1253, Aug. 2014.

[35] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel, “Good vibrations: A modal analysis approach for sequential non-rigid structure from motion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1558–1565.

Zhong Zhou (M'10) received the B.S. degree from Nanjing University, Nanjing, China, in 1999, and the Ph.D. degree from Beihang University, Beijing, China, in 2005.

He is an Associate Professor and Ph.D. Adviser with the State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. His main research interests include augmented virtual environment, natural phenomena simulation, distributed virtual environment, and Internet-based VR technologies.

Dr. Zhou is a member of the ACM and CCF.

Feng Shi received the M.S. degree in mathematics from the University of Science and Technology Beijing, Beijing, China, in 2006, and is currently working toward the Ph.D. degree in computer science at the State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China.

His research interests include representation and modeling of dynamic objects in videos, detection human actions from videos, and motion segmentation.

Jiangjian Xiao (S'03–M'05) received the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA, in 2004.

He is a Professor with the Ningbo Industrial Technology Research Institute, CAS, Ningbo, China. His research interests include computer vision, computer graphics, and visualization.

Dr. Xiao is a member of the ACM. He is an Associate Editor of the *Machine Vision and Application Journal*.

Wei Wu received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1995.

He is a Professor with the State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China, and he is currently the Chair of the CCF Technical Committee on VRV. His current research interests include virtual reality, virtualization, and distributed interactive simulation.