sensors

MDPI

Article

Capturing Complex 3D Human Motions with Kernelized Low-Rank Representation from Monocular RGB Camera

Xuan Wang [1,2,3,4], **Fei Wang** [1,2,3,4,*] **and Yanan Chen** [1,2,3,4]

[1] The Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an 710048, China; xwang.cv@gmail.com (X.W.); chenyanan@stu.xjtu.edu.cn (Y.C.)

[2] The School of Software Engineering, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an 710048, China

[3] National Engineering Laboratory for Visual Information Processing and Application, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an 710048, China

[4] Shaanxi Digital Technology and Intelligent System Key Laboratory, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an 710048, China

* Correspondence: wfx@mail.xjtu.edu.cn

Received: 5 July 2017; Accepted: 24 August 2017; Published: 3 September 2017

Abstract: Recovering 3D structures from the monocular image sequence is an inherently ambiguous problem that has attracted considerable attention from several research communities. To resolve the ambiguities, a variety of additional priors, such as low-rank shape basis, have been proposed. In this paper, we make two contributions. First, we introduce an assumption that 3D structures lie on the union of nonlinear subspaces. Based on this assumption, we propose a Non-Rigid Structure from Motion (NRSfM) method with kernelized low-rank representation. To be specific, we utilize the soft-inextensibility constraint to accurately recover 3D human motions. Second, we extend this NRSfM method to the marker-less 3D human pose estimation problem by combining with Convolutional Neural Network (CNN) based 2D human joint detectors. To evaluate the performance of our methods, we apply our marker-based method on several sequences from Utrecht Multi-Person Motion (UMPM) benchmark and CMU MoCap datasets, and then apply the marker-less method on the Human3.6M datasets. The experiments demonstrate that the kernelized low-rank representation is more suitable for modeling the complex deformation and the method consequently yields more accurate reconstructions. Benefiting from the CNN-based detector, the marker-less approach can be applied to more real-life applications.

Keywords: 3D human pose estimation; monocular reconstruction; non-rigid structure from motion; kernel low-rank representation

1. Introduction

Most of the video or image data used nowadays are captured by a single RGB sensor, such as cell phone cameras, which is one of the most widely used sensors. Recovering the 3D structure from such data is an active research field in computer vision and computer graphics communities. In this work, we focus on tackling the problem of reconstructing the complex 3D motions, especially for the 3D human poses, from the monocular image sequence. This is an inherently ill-posed problem since the same projection can be captured from different 3D structures. To resolve the ambiguities, several approaches relying on additional priors have been proposed. To be specific, lots of methods have been presented to solve the 3D human pose estimation problem, which leverage more priors from human bodies.

The seminal factorization [1] of Non-Rigid Structure from Motion (NRSfM) techniques was proposed to tackle the non-rigid problem via extending its rigid version [2]. In Xiao et al. [3], the shape basis constraints is presented to resolve the inherent ambiguities and derived the closed-form solution. Akhter et al. [4,5] showed the dual approach which modeled 3D trajectories under basis constraints. In addition, the trajectory-space method was proposed in Gotardo and Martinez [6]. In Akhter et al. [7], it proved that, even though there is an ambiguity in shape bases or trajectory bases, non-rigid shapes can still be recovered uniquely. Based on this, a prior-free method [8] was introduced to estimate the 3D non-rigid structures and camera rotations by only exploiting the low-rank shape assumption. In Wang et al. [9], they use the low-rank assumption in a similar way, but an Accelerated Proximal Gradient (APG) algorithm solver is employed to solve the resulting problem. Furthermore, the method in Gotardo and Martinez [10] combined the shape basis model and trajectory basis model, and revealed trajectories of the shape basis coefficients. The Procrustean Normal Distribution (PND) model was proposed in Lee et al. [11], where 3D shapes are aligned and fit into a normal distribution. Simon et al. [12] exploited the Kronecker pattern in shape-trajectory (spatial-temporal) priors. Then, Zhu and Lucey [13] combined the convolutional sparse coding technique with NRSfM by using point trajectory. Nevertheless, it requires learning an over-complete basis of trajectories. Most recently, a novel part-based method has been proposed in Lee et al. [14]. For most existing approaches mentioned above, the performance highly depends on the complexity of the 3D non-rigid motions. Generally, the correspondences between different frames are assumed to be given, by e.g., infrared markers. As a consequence, they usually fail to handle the reconstruction problem in uncontrolled scenarios with complex non-rigid motions.

The human body is an important subject in nonrigid reconstruction since reconstruction of 3D human structures has broad applications, such as athletic training, computer animation and gait analysis. Using multi-camera systems [15,16] can easily resolve the ambiguities in non-rigid reconstruction problems, and the requirement for dedicated equipment prevents the methods from applying in more practical scenarios. Using a depth camera can also avoid the ambiguities in monocular 3D non-rigid reconstruction problems [17]. Since 3D information can be directly obtained by depth-cameras, the problem of recovering 3D human poses can be formulated as problem of 3D tracking [18]. However, massive existing videos are captured by the single RGB camera. Therefore, estimating 3D human structures from monocular image sequence is still an essential task. In Rádlová et al. [19], a model-based approached was presented. According to Sigal [20], estimating the accurate pose on single frame is an ill-posed problem, and exploiting all available information across the sequence can promote performance [21,22]. Latent Variable Models (LVMs) are often used in the literature [23,24]. Tian et al. [25] proposed a discriminative approach that leverages LVMs, and successfully addressed the over-fitting and poor generalization problem. In Zhu et al. [26], the human body is classified into several parts; a pre-defined human model is fitted to the input images; and the dense reconstruction is yielded. Ek et al. [27] presented a method relying on the Gaussian process latent variable models (GPLVM), in which a parametric mapping from pose to latent space is learned to enforce a one-to-one correspondence. In the work of Tekin et al. [28], they employed two convolutional neural networks to align the bounding boxes of the human in consecutive frames, then created the data volume and reconstructed the 3D structure from the volume with Kernel Ridge Regression (KRR) and Kernel Dependency Estimation. With the advent of deep learning, it is obvious that using CNN to obtain the human joint detection or to estimate the 3D structures directly attracts more and more attention [29]. For more details, a comprehensive survey can be found in Gong et al. [30].

In this paper, we present an NRSfM method to tackle the problem of recovering complex 3D non-rigid motions from a monocular sequence. The term ’complex’ here means that the entire motions are composed of several different ’primitive’ or ’simple’ actions. The proposed method empirically shows that the assumption, i.e., 3D shapes lie on the union of nonlinear shape subspaces, is better than the linear version [31] in modeling complex 3D non-rigid motions. We formulate this assumption as kenerlized low-rank representation and model the motion recovery as solving an optimization problem.

The experiments demonstrate that our method has better performance than the state-of-the-art methods, especially for tackling the complex motions. Moreover, using the outputs, called joint heightmaps in this paper, of a CNN-based human joint detector [32] can remove the dependency on markers. As a result, our method can be applied to more practical scenarios. We can conveniently switch between the joint heightmap based data term and the traditional re-projection term. This yields two versions of our method—marker-based and marker-less versions—and ensures the method can be applied more broadly.

2. Motivations

Before we introduce the proposed method, let us give a brief review of two lines of research that are most relevant to our method. One is the NRSfM based on low-rank shape constraints, and the other is low-rank representation.

2.1. NRSfM with Low-Rank Priors

In Dai et al. [8], an NRSfM method only relying on low-rank priors was proposed. Given the tracked feature points, e.g., human joints in images, the reconstruction is modeled as the following optimization problem:

$$\begin{array}{rl} \min_{\mathbf{X}} & \frac{1}{2}\|\mathbf{W}-\mathbf{R}\mathbf{X}\|_F^2+\mu\|\mathbf{X}^{\#}\|_* \\ s.t. & \mathbf{X}^{\#}=g(\mathbf{X}), \end{array} \tag{1}$$

where $\mathbf{X} \in \Re^{3N_P \times N_F}$ is the 3D structure matrix to be estimated, which stacks 3D coordinates of all the N_P feature points at all the N_F frames. The first term in Equation (1) is a re-projection term that encourages projections of the estimated 3D structures to be close to the observations in $\mathbf{W}$. In this term, $\mathbf{W} \in \Re^{2N_P \times N_F}$ and $\mathbf{R} \in \Re^{2N_F \times 3N_F}$ are observation matrix and rotation matrix, which encode the camera rotations and 2D positions of tracked feature points (human joints for human pose estimation), respectively. The second term encourages the 3D structures at all the frames to be low-rank. In other words, all the 3D structures can be represented by linear combinations of a few 3D structure bases, the number of which is much smaller than the number of the frames. Note that the shape basis is not obtained explicitly here. By contrast, the nuclear norm used in the first term ensures that estimated 3D structures should be as low-rank as possible. The relative weight μ is to control how strong this prior influences the resulting reconstruction. Therefore, this makes the original monocular non-rigid reconstruction problem well-constrained. However, when motions are complex, 3D structures are quite different from each other, and the low-rank assumption is broken since 3D structures don't lie on a single low-rank linear space. In this case, either a high-rank shape space or multiple subspaces are needed to represent the 3D structures.

2.2. Low-Rank Representation

Low-rank representation (LRR) [33] was proposed to seek the desired low-rank affinity matrix $\mathbf{Z}$ via solving the optimization problem in Equation (2):

$$\begin{array}{rl} \min_{\mathbf{Z},\mathbf{E}} & \|\mathbf{Z}\|_*+\lambda\|\mathbf{E}\|_{2,1} \\ s.t. & \mathbf{X}=\mathbf{X}\mathbf{Z}+\mathbf{E}, \end{array} \tag{2}$$

where $\mathbf{X}$ is data matrix and $\mathbf{E}$ is error matrix. In particular, LRR uses the self-expressiveness property of data, i.e., each data point in a union of subspaces can be efficiently represented as a linear combination of other points. When pursuing the low-rank affinity matrix, the data points are assigned to different subspaces implicitly. This property ensures LRR can be an effective regularization to handle the complex motions in NRSfM problem [31]. To this end, the formulation is as follows:

$$\begin{array}{rl} \min_{\mathbf{X},\mathbf{Z},\mathbf{E}} & \|\mathbf{Z}\|_*+\lambda_1\|\mathbf{X}\|_*+\lambda_2\|\mathbf{E}\|_l \\ s.t. & \mathbf{X}=\mathbf{X}\mathbf{Z},\ \mathbf{W}=\mathbf{R}\mathbf{X}^{\#}+\mathbf{E}. \end{array} \tag{3}$$

Nevertheless, the deformation of non-rigid object, e.g., human body, is often nonlinear. To better model the nonlinear deformation, the traditional approach is mapping the data point to a higher-dimension space in which the linear low-rank structure can be obtained. We empirically find that the kernerlized low-rank representation (KLRR) method [34] has better performance on human motion clustering than original LRR. To this end, we introduce the KLRR as a regularization to the NRSfM problem.

3. Formulations

In order to recover the highly complex 3D motions (also named structures, shapes or deformations) from a monocular image sequence, we map the 3D structures to a space with higher dimension, as shown in Equation (4)

$$\begin{array}{rl} \min_{\mathbf{X},\mathbf{Z},\mathbf{E}} & \|\mathbf{Z}\|_* + \lambda_1\|\mathbf{X}\| + \lambda_2\|\mathbf{E}\|_1 + \lambda_3\|\Phi(\mathbf{X})(\mathbf{I}-\mathbf{Z})\|_F^2 \\ s.t. & \mathbf{W} = \mathbf{R}\mathbf{X}^{\#} + \mathbf{E}, \end{array} \tag{4}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,N_F} \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{2,N_F} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{N_P,1} & \mathbf{x}_{N_P,2} & \cdots & \mathbf{x}_{N_P,N_F} \end{bmatrix}, \quad \mathbf{X}^{\#} = \begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{2,1} & \cdots & \mathbf{x}_{N_P,1} \\ \mathbf{x}_{1,2} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{N_P,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{1,N_F} & \mathbf{x}_{2,N_F} & \cdots & \mathbf{x}_{N_P,N_F} \end{bmatrix}. \tag{5}$$

We follow the definition of notation above. Assuming the object is human body whose motions are captured by an orthographic camera, the aim of the proposed method is to recover the 3D position of its joints from the given 2D location of joints only. $\mathbf{X} \in \Re^{3N_P \times N_F}$ is the 3D motion matrix; here, the 3D coordinates are defined in the camera coordinate system. In both matrices $\mathbf{X}$ and $\mathbf{X}^{\#}$, vector $\mathbf{x}_{p,f}$ stacks the 3D coordinates of p_{th} joint at frame f, but they have different arrangement of these vectors. $\mathbf{R}$ is the $\Re^{2N_F \times 3N_F}$ block-diagonal matrix of N_F orthographical camera matrices. The given 2D tracks of all the joint points are in the observation matrix $\mathbf{W} \in \Re^{2N_F \times N_P}$, and $\mathbf{E}$ is an error matrix with the same dimensions of $\mathbf{W}$. The $\Phi(\mathbf{X})$ is the defined mapping, such that the mapped $\mathbf{X}$ resides in multiple linear subspaces. In addition, the matrix $\mathbf{Z} \in \Re^{N_F \times N_F}$ is the affinity matrix, i.e., the self-expressive coefficient matrix. Note that the noise in this problem usually occurs at the tracking or joint detection process. In this work, the part of tracking or detection is assumed to be given by other existing methods. Certainly, such tracking or detection approaches could not yield the perfectly accurate results. Thus, when dealing with the re-projection constraint, we use the matrix E to model the possible re-projection error. Furthermore, we penalise the more robust L_1 norm rather L_2 norm of this matrix. It ensures that the method is relatively robust to the outliers of observation.

The low-rank shape priors is preserved in our framework. In Equation (4), $\lambda_1\|\mathbf{X}\|_*$ is to exploit the low-rank nature of non-rigid objects, and the constraint $\mathbf{W} = \mathbf{R}\mathbf{X}^{\#} + \mathbf{E}$ is used to penalize the re-projection error. Note that nuclear norm, a convex approximation of matrix rank, of $\mathbf{X}$, rather than $\mathbf{X}^{\#}$, is minimized here, as described in Dai et al. [8]. This is because the rank of $\mathbf{X}$ is bound by $\min(N_F, 3N_P)$, whereas the rank of $\mathbf{X}^{\#}$ is bound by $\min(3N_F, N_P)$. Minimizing the rank of $\mathbf{X}$ is preferable as it attempts to directly learn redundancies among frames.

The information revealing the low-rank subspaces structure exists in the affinity matrix $\mathbf{Z}$. By applying the spectral clustering to $\mathbf{Z}$, the explicit structure can be obtained. More details about this process can be found in [33,34]. Nevertheless, our method doesn't need the explicit subspace structure. It only needs to ensure that the 3D structures from a lower rank subspaces are preferred by minimizing the nuclear norm of $\mathbf{Z}$. If the low-rank subspaces assumption holds, this term will be a proper regularization for NRSfM problem with complex 3D motions. Unfortunately, LRR may not yield satisfied results when the 3D structures are actually from the nonlinear subspaces, since it is originally designed to handle the linear case.

Due to this drawback of LRR, we exploit the kernel-induced mapping $\mathbf{X} \rightarrow \Phi(\mathbf{X})$. We denote the column in $\mathbf{X}$ as $\{\mathbf{x}_i\}_{i=1}^{N_F}$, where $\mathbf{x}_i \in \Re^{3N_P}$, define the kernel matrix $\mathbf{K} \in \Re^{N_F \times N_F}$, and then the elements of K are calculated as follows:

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \ \ \forall i, j = 1, \ldots, N_F, \tag{6}$$

where $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j)$ is the kernel function. Defining $\Phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_{N_F})]$, we have $\mathbf{K}(\mathbf{X}) = \Phi(\mathbf{X})^{\mathrm{T}}\Phi(\mathbf{X})$. In the remainder of this paper, we use $\mathbf{K}$ to represent $\mathbf{K}(\mathbf{X})$ for short. We can rewrite Equation (4) as follows:

$$\begin{array}{ll} \min_{\mathbf{X},\mathbf{Z},\mathbf{E}} & \|\mathbf{Z}\|_* + \lambda_1\|\mathbf{X}\| + \lambda_2\|\mathbf{E}\|_1 + \lambda_3 trace((\mathbf{I}-\mathbf{Z})^{\mathrm{T}}\mathbf{K}(\mathbf{I}-\mathbf{Z})) \\ s.t. & \mathbf{W} = \mathbf{R}\mathbf{X}^{\#} + \mathbf{E}. \end{array} \tag{7}$$

Equation (7) above is designed for the general purpose. When considering the special case, e.g., recovering the 3D motions of human bodies, we can leverage additional particular priors. The exploited prior is that the length between the connected joints is invariable. Given the set of all the pairs of connected joints $\mathcal{E}$, the inextensibility constraints can be written as:

$$R(\mathbf{X}, \mathbf{l}) = \textstyle\sum_{f=1}^{N_F} \sum_{(p,q)\in\mathcal{E}} (\|\mathbf{x}_{f,p} - \mathbf{x}_{f,q}\|_2 - l_{p,q})^2, \tag{8}$$

where $\mathbf{x}_{f,p}$ stacks the 3D coordinates of point p at frame f. The $\mathbf{l} = \{l_{p,q}\}$ are unknown, in which $l_{p,q}$ encodes the length between the pairs of connected joints p and q, and thus act as the auxiliary variables to be determined by our algorithm. Adding Equation (8) into Equation (7), we have the final objective function as follows:

$$\begin{array}{ll} \min_{\mathbf{X},\mathbf{Z},\mathbf{E},\mathbf{l}} & \|\mathbf{Z}\|_* + \lambda_1\|\mathbf{X}\| + \lambda_2\|\mathbf{E}\|_1 + \lambda_3 trace((\mathbf{I}-\mathbf{Z})^{\mathrm{T}}\mathbf{K}(\mathbf{I}-\mathbf{Z})) + \lambda_4 R(\mathbf{X}, \mathbf{l}) \\ s.t. & \mathbf{W} = \mathbf{R}\mathbf{X}^{\#} + \mathbf{E}. \end{array} \tag{9}$$

In this formulation, the observation matrix W is often obtained by setting markers on the subject. Nevertheless, in some daily-life scenarios, e.g., recovering 3D human motions using a camera on the mobile phone, markers are usually unavailable. To this end, we provide an optional version of our method in which an CNN-based human joint detector [32] is employed. By a small modification, Equation (9) can be conveniently converted to a marker-less version as follows:

$$\begin{array}{ll} \min_{\mathbf{X},\mathbf{Z},\mathbf{E},\mathbf{l}} & \|\mathbf{Z}\|_* + \lambda_1\|\mathbf{X}\| + \lambda_2\|\mathbf{E}\|_1 + \lambda_3 trace((\mathbf{I}-\mathbf{Z})^{\mathrm{T}}\mathbf{K}(\mathbf{I}-\mathbf{Z})) + \lambda_4 R(\mathbf{X}, \mathbf{l}) + \lambda_5 H(\hat{\mathbf{W}}) \\ s.t. & \mathbf{W} = \mathbf{R}\mathbf{X}^{\#} + \mathbf{E}, \mathbf{W} = \hat{\mathbf{W}}, \end{array} \tag{10}$$

where the function $H(\hat{\mathbf{W}})$ is yielded by the CNN-based detector. For arbitrary joint point p and frame f, the CNN-based detector provides a heightmap $h_{f,p}$ with the same size of input image. The pixel in the heightmap $h_{f,p}$ takes a non-negative value, which indicates how possible it is the projection of the joint p at frame f. Therefore, the function $H(\hat{\mathbf{W}})$ is calculated as follows:

$$H(\hat{\mathbf{W}}) = -\sum_{f=1}^{N_F}\sum_{p=1}^{N_P} h_{f,p}(\hat{\mathbf{w}}_{f,p}), \tag{11}$$

where $\hat{\mathbf{W}} = \{\hat{\mathbf{w}}_{f,p}\}$ is the auxiliary variable introduced for the convenience of optimization. The $\hat{\mathbf{w}}_{f,p}$ stacks the 2D image coordinate of the projection of 3D point $\mathbf{x}_{f,p}$.

4. Optimization

In the previous section, capturing 3D human motions from monocular images is formulated as solving the optimization problem (9). We use ALM to solve the yielding optimization problem. By introducing auxiliary variables $\mathbf{C}$ and $\hat{\mathbf{X}}$, the complete Lagrangian formulation is:

$$\begin{aligned} \min_{\mathbf{X},\hat{\mathbf{X}},\mathbf{Z},\mathbf{E},\mathbf{C},\mathbf{l}} \max_{\mathbf{\Gamma}_1,\mathbf{\Gamma}_2,\mathbf{\Gamma}_3} \mathcal{L} = \; & \|\mathbf{Z}\|_* + \lambda_1\|\hat{\mathbf{X}}\|_* + \lambda_2\|\mathbf{E}\|_1 + \lambda_3 trace(\mathbf{C}^{\mathrm{T}}\mathbf{K}\mathbf{C}) + \lambda_4 R(\mathbf{X},\mathbf{l}) \\ & + < \mathbf{\Gamma}_1, \mathbf{W} - \mathbf{R}\mathbf{X}^{\#} - \mathbf{E} > + \tfrac{\mu}{2}\|\mathbf{W} - \mathbf{R}\mathbf{X}^{\#} - \mathbf{E}\|_F^2 \\ & + < \mathbf{\Gamma}_2, \mathbf{C} - \mathbf{I} + \mathbf{Z} > + \tfrac{\mu}{2}\|\mathbf{C} - \mathbf{I} + \mathbf{Z}\|_F^2 \\ & + < \mathbf{\Gamma}_3, \mathbf{X} - \hat{\mathbf{X}} > + \tfrac{\mu}{2}\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2, \end{aligned} \tag{12}$$

where $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are the Lagrangian multiplier matrices. In iterations of Algorithm 1, the variables $\{\mathbf{X}, \mathbf{Z}, \mathbf{E}, \mathbf{C}, \mathbf{l}\}$ are alternatively solved. When solving each subproblem, the other fixed variables are regarded as constants; hence, we ignore the iteration index for these variables. Then, at the end of each iteration, the penalty factor μ and Lagrangian multiplier matrices $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ and $\mathbf{\Gamma}_3$ are updated.

Algorithm 1: Solving problem (12) by ALM.

```
**Input**: $\mathbf{W}, \mathbf{R}, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_{max}, \rho$
**Output**: $\mathbf{X}, \mathbf{Z}, \mathbf{E}, \mathbf{C}, \mathbf{l}$
initialization $\mathbf{X}^0, \mathbf{Z}^0, \mathbf{E}^0, \mathbf{C}^0, \mathbf{l}^0, \mathbf{\Gamma}_1 = \mathbf{0}, \mathbf{\Gamma}_2 = \mathbf{0}, \mu_0 = 0, k = 0$;
**while** *not converged* **do**
    $\mathbf{X}^{(k+1)} \leftarrow \arg\min_{\mathbf{X}} \mathcal{L}(\mathbf{X}^{(k)}, \hat{\mathbf{X}}, \mathbf{Z}, \mathbf{E}, \mathbf{C}, \mathbf{l})$;
    $\hat{\mathbf{X}}^{(k+1)} \leftarrow \arg\min_{\hat{\mathbf{X}}} \mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}^{(k)}, \mathbf{Z}, \mathbf{E}, \mathbf{C}, \mathbf{l})$;
    $\mathbf{Z}^{(k+1)} \leftarrow \arg\min_{\mathbf{Z}} \mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}, \mathbf{Z}^{(k)}, \mathbf{E}, \mathbf{C}, \mathbf{l})$;
    $\mathbf{E}^{(k+1)} \leftarrow \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}, \mathbf{Z}, \mathbf{E}^{(k)}, \mathbf{C}, \mathbf{l})$;
    $\mathbf{C}^{(k+1)} \leftarrow \arg\min_{\mathbf{C}} \mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}, \mathbf{Z}, \mathbf{E}, \mathbf{C}^{(k)}, \mathbf{l})$;
    $\mathbf{l}^{(k+1)} \leftarrow \arg\min_{\mathbf{l}} \mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}, \mathbf{Z}, \mathbf{E}, \mathbf{C}, \mathbf{l}^{(k)})$;
    $\mathbf{\Gamma}_1^{(k+1)} \leftarrow \mathbf{\Gamma}_1^{(k)} + \mu^{(k)}(\mathbf{W} - \mathbf{R}\mathbf{X}^{\#} - \mathbf{E})$;
    $\mathbf{\Gamma}_2^{(k+1)} \leftarrow \mathbf{\Gamma}_2^{(k)} + \mu^{(k)}(\mathbf{C} - \mathbf{I} + \mathbf{Z})$;
    $\mathbf{\Gamma}_3^{(k+1)} \leftarrow \mathbf{\Gamma}_3^{(k)} + \mu^{(k)}(\mathbf{X} - \hat{\mathbf{X}})$ ;
    $\mu^{k+1} \leftarrow min(\mu_{max}, \rho\mu^k)$;
    $k \leftarrow k + 1$;
**end**
```

4.1. *The Solution of* $\hat{\mathbf{X}}$ *and* $\mathbf{Z}$

By a simple derivation, the subproblems of solving $\mathbf{Z}$ and $\hat{\mathbf{X}}$ can be written as the following optimization problems:

$$\begin{aligned} \min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \tfrac{\mu}{2}\|\mathbf{Z} - (\mathbf{I} - \mathbf{C} - \tfrac{\mathbf{\Gamma}_2}{\mu})\|_F^2, \\ \min_{\hat{\mathbf{X}}} \lambda_1\|\hat{\mathbf{X}}\|_* + \tfrac{\mu}{2}\|\hat{\mathbf{X}} - (\mathbf{X} + \tfrac{\mathbf{\Gamma}_3}{\mu})\|_F^2, \end{aligned} \tag{13}$$

whose formulation is the same as that of the standard problem in Cai et al. [35]. For this reason, there are closed-form solutions for $\mathbf{Z}$ and $\hat{\mathbf{X}}$. In the iteration k, these variables are updated by

$$\begin{aligned} \mathbf{Z}^{(k+1)} = \mathcal{D}_{1/\mu}(\mathbf{I} - \mathbf{C}^{(k)} - \tfrac{\mathbf{\Gamma}_3^{(k)}}{\mu}), \\ \hat{\mathbf{X}}^{(k+1)} = \mathcal{D}_{\lambda_1/\mu}(\mathbf{X}^{(k+1)} + \tfrac{\mathbf{\Gamma}_3}{\mu}), \end{aligned} \tag{14}$$

where $\mathcal{D}_\tau(\cdot)$ is the Singular Value Decomposition operator. The readers can refer to Cai et al. [35] for mode details.

4.2. The Solution of $\mathbf{X}$

The subproblem of solving $\mathbf{X}$ has no closed-form solution due to the existence of non-convex terms. It is obvious that the subproblem can be converted to a nonlinear least square problem by the similar derivation, which is used to get Equation (13). Exploiting the gradient-based iterative optimization algorithm [36], we get the estimation of $\mathbf{X}$ at iteration k:

$$\begin{aligned} \mathbf{X}^{(k+1)} = \mathrm{argmin}_{\mathbf{X}} \quad & \lambda_3 trace(\mathbf{C}^{(k)^{\mathrm{T}}}\mathbf{K}\mathbf{C}^{(k)}) + \lambda_4 R(\mathbf{X}, \mathbf{l}^k) \\ & + \tfrac{\mu}{2}(\|\mathbf{X} - \hat{\mathbf{X}}^{(k)} + \tfrac{\mathbf{\Gamma}_3^{(k)}}{\mu}\|_F^2 + \|\mathbf{W} - \mathbf{R}\mathbf{X}^{\#} - \mathbf{E}^{(k)} + \tfrac{\mathbf{\Gamma}_1^{(k)}}{\mu}\|_F^2). \end{aligned} \tag{15}$$

4.3. The Solution of $\mathbf{E}$

Fixing $\mathbf{X}$, $\hat{\mathbf{X}}$, $\mathbf{Z}$, $\mathbf{C}$ and $\mathbf{l}$, the cost function is reduced to

$$\min_{\mathbf{E}} \lambda_2 \|\mathbf{E}\|_1 + \tfrac{\mu}{2}\|\mathbf{W} - \mathbf{R}\mathbf{X}^{\#} - \mathbf{E} + \tfrac{\mathbf{\Gamma}_1}{\mu}\|_F^2. \tag{16}$$

The resulting l_1 minimization problem has a closed-form solution:

$$\mathbf{E}^{k+1} = \mathcal{S}_{\lambda_2/\mu}(\mathbf{W} - \mathbf{R}\mathbf{X}^{\#^{(k+1)}} + \tfrac{\mathbf{\Gamma}_1^{(k)}}{\mu}), \tag{17}$$

where $\mathcal{S}_\tau(\cdot)$ is the element-wise shrinkage thresholding operator [37].

4.4. The Solution of $\mathbf{C}$ and $\mathbf{l}$

Since $\mathbf{K}$ is semi-definite positive matrix, both subproblems for solving $\mathbf{C}$ and $\mathbf{l}$ are convex:

$$\begin{gathered} \min_{\mathbf{C}} \lambda_3 trace(\mathbf{C}^{\mathrm{T}}\mathbf{K}\mathbf{C}) + \tfrac{\mu}{2}\|\mathbf{C} - \mathbf{I} + \mathbf{Z} + \tfrac{\mathbf{\Gamma}_2}{\mu}\|_F^2, \\ \min_{\mathbf{l}} \lambda_4 R(\mathbf{X}, \mathbf{l}). \end{gathered} \tag{18}$$

The subproblems have closed-form solutions. Setting the derivatives to zero first, we then solve the yielded linear least square problems to obtain the solutions:

$$\begin{aligned} \mathbf{C}^{(k+1)} &= (2\lambda_3 \mathbf{K}^{(k+1)} + \mu \mathbf{I}_{N_F})^{-1}(\mu(\mathbf{I} - \mathbf{Z}^{(k+1)}) - \mathbf{\Gamma}_2^{(k)}), \\ l_{p,q}^{(k+1)} &= \tfrac{1}{N_F}\textstyle\sum_{f=1}^{N_F} \|\mathbf{x}_{f,p}^{(k+1)} - \mathbf{x}_{f,q}^{(k+1)}\|_2. \end{aligned} \tag{19}$$

For the existence of non-convex term and multiple blocks in our proposed objective function, the convergence of ALM is not guaranteed. Nonetheless, the Algorithm 1 successfully converged in our experiment as shown in Figure 1. For all the experiments, we leverage the resulting $\mathbf{X}$ from PTA [5] as the initialisation.

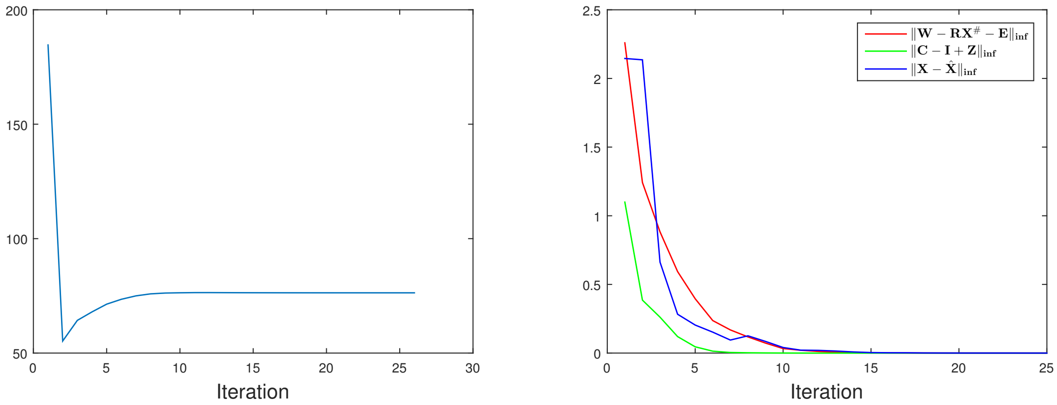

Figure 1. Convergence curve. We test our TUNS method on '56-02' sequence in CMU MoCap datasets is used here. We plot the residual of the objective function (**left**) and the residuals of three constraints (**right**).

5. Experiments

We make the quantitative evaluation on two benchmarks: (i) synthetic camera 2D projections generated from the 3D CMU Motion Capture (MoCap) dataset, (ii) real-world 2D projection from the 2D point-tracks of video in Utrecht Multi-Person Motion (UMPM) [38] dataset. Furthermore, the performance of our marker-less algorithm is illustrated on the Human3.6M. We compare our approach, which is denoted as TUNS (temporal union of nonlinear subspaces) in the remainder of this section, against six NRSfM baselines: point tracking algorithm PTA [5], the trajectory-sapce method CSF [6], the block matrix method BMM [8], the temporal union of subspaces TUS [31], the accelerated proximal gradient optimization APG [9] and the consensus NRSfM of CNR [14]. For PTA [5], CSF [6], BMM [8], CNR [14], we use authors' implementation in experiments. For PTA [5] and CSF [6], we manually set the rank of the subspace to the value yielding the best results. For TUS [31] and APG [9], since there are not publicly available implementations, our re-implementations are adopted in comparison. We test such re-implementations and get similar results to what the authors reported in [9,31].

5.1. Subspaces Analysis

We empirically demonstrate why the nonlinear low-rank representation could further help the reconstruction. To illustrate the effectiveness of utilising LRR and KLRR to recover the low-rank structures of 3D human motions, we select the motion sequence 'p1_grab_3' from the UMPM benchmark. We apply both LRR and KLRR to the 3D motions to get the affinity matrices. Meanwhile, our reconstruction method also produces an affinity matrix. The difference is that LRR and KLRR estimate the affinity matrices on the 3D data directly; however, our method jointly estimates the 3D motions and affinity matrix from the 2D projection. Applying spectral clustering algorithm to these affinities, the motion clustering results are yielded. Obviously in Figure 2, the colour in the first row of the colour bars changes more frequently. This phenomenon is not preferable, since the image sequence is temporally consecutive. In other words, the LRR approach divides quite similar motions into different categories. Nevertheless, the clustering result based on KLRR, which has high temporal consistency, is quite close to the result of our method. As shown on bottom right of Figure 2, the affinity matrix from KLRR has higher clarity of the low-rank structure than the one from LRR. This experiment demonstrates that the nonlinear low-rank representation is more suitable to model the complex 3D motions than its linear counterpart.

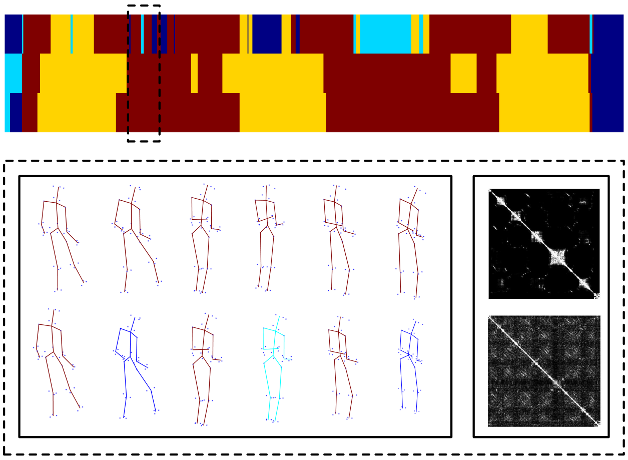

Figure 2. Subspace analysis on 3D human motions. On the top of this figure, three colour bars represent the subspace clustering results. In each bar, the same colour means same cluster. From the top down, the clustering results are obtained from the affinity matrices yielded by LRR [33], KLRR citeXiaoTNNLS15 and our method, respectively. For the first two rows, the LRR and KLRR are applied to the 3D motions 'p1_grab_3' from the UMPM benchmark. For the third row, the affinity matrix produced from the reconstruction process is exploited. On the bottom, two rows of 3D motions and the clustering in a short and consecutive interval are visualized in a dash-line box. The first row is produced by KLRR and the second one is from LRR. The corresponding affinity matrices are shown on the right side.

5.2. Quantitative Evaluation

We quantitatively compare our TUNS approach with six baselines on 20 sequences of subjects 56 and 86, which contain several primitive actions, in CMU motion capture datasets. The synthetic 2D projections are generated by a randomly rotating orthographic camera, which rotates with the y-axis of camera and is always pointing at the centre of the body. Our TUNS approach and baseline methods are employed to reconstruct the 3D motions from the synthetic 2D projections. Since the monocular reconstruction has a global scale ambiguity, a type of mean relative reconstruction error e_{mean}, which is utilised in [8], is employed in this paper. This error is computed as follows:

$$\begin{array}{c} e_{mean} = \frac{1}{\sigma N_F N_P} \sum_{f=1}^{N_F} \sum_{p=1}^{N_P} e_{fp}, \\ \sigma = \frac{1}{3F} \sum_{f=1}^{F} (\sigma_{fx} + \sigma_{fy} + \sigma_{fz}), \end{array} \tag{20}$$

where σ_{fx}, σ_{fy} and σ_{fz} are the standard deviations in x, y, and z coordinates of the original shape at frame f. For each joint p, e_{fp} is the distance between its reconstructed 3D position and its ground truth 3D position at frame f. Meanwhile, we report the median error which is computed as follows:

$$e_{med} = median(\{e_{fp} | f = 1 \dots N_F, p = 1 \dots N_P\}). \tag{21}$$

As reported in Table 1, our TUNS approach yields more accurate reconstructions than the baselines. PTA [5] can give a close-form solution that can be easily used as initialization for all the nonlinear optimization based algorithms. For CSF [6], it is not suitable for recovering the complex motion, since it has high space complexity with respect to the rank K. BMM [8] utilises no prior but the low-rank subspace. In such a method, the motions are assumed to lie in a single subspace. Consequently, it is less effective when recovering the complex motion. APG [9] inherently uses the single subspace, whereas it performs better than BMM [8] since a more effective rank-minimisation technique is employed. The part-based method CNR [14], which is more adept at handling complex shape configurations, yields better reconstruction of subject 86 than the subject 56, since sequences of subject 86 have more points than subject 56. Nonetheless, for the complex motions, using temporal union of subspaces (TUS and TUNS) can yield better reconstruction than the other methods. Benefiting from the kernel technique, our TUNS yields even more accurate reconstructions than TUS.

Meanwhile, we also evaluate the performance of our approach by reconstructing 3D motion from real-world 2D projection stemming from 2D point-tracks of videos in UMPM. Our TUNS and baselines are tested on the three sequences 'p1_grab_3', 'p3_ball_12' and 'p1_chair_2'. In Table 2, it shows that our method produces more accurate reconstructions. It demonstrates that TUNS is suitable for tackling the long-term sequence with complex motions. The reconstructions of some selected frames from UMPM datasets are shown in Figures 3–5. Please note that Figure 5 illustrates that our TUNS approach is even effective in multi-object cases.

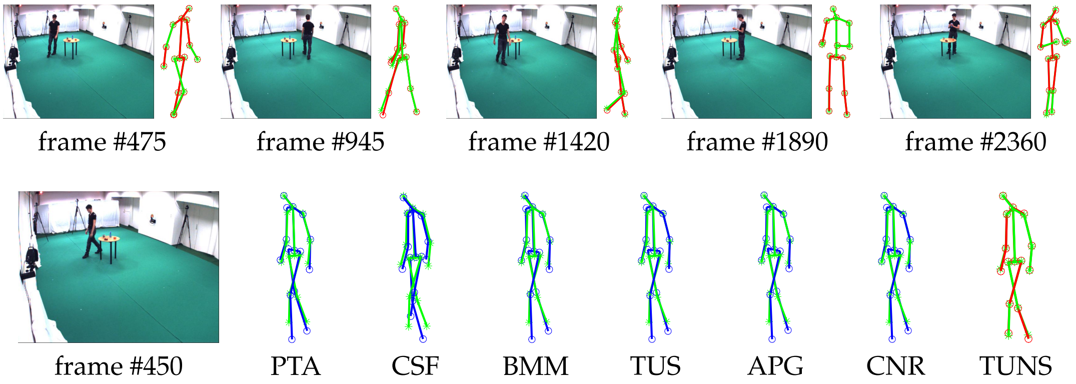

Figure 3. Evaluation on 'p1_grab_3' sequence in UMPM. For five selected frames, our reconstruction (red skeleton), ground truth (green skeleton) and the corresponding image are shown on top. On the bottom, the reconstructions, yielded by all the baselines (blue skeleton) and our TUNS (red skeleton) of one specific frame, are shown with overlapping ground truth (green skeleton).

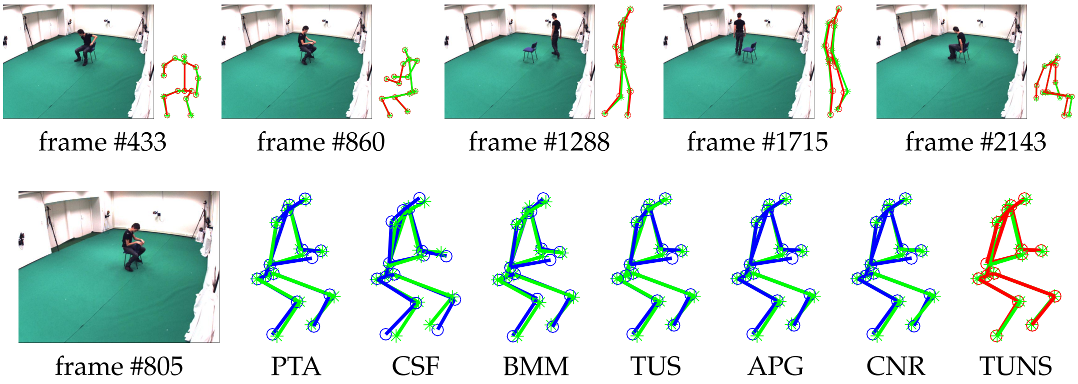

Figure 4. Evaluation on 'p1_chair_2' sequence in UMPM. For five selected frames, our reconstruction (red skeleton), ground truth (green skeleton) and the corresponding image are shown on top. On the bottom, the reconstructions, yielded by all the baselines (blue skeleton) and our TUNS (red skeleton) of one specific frame, are shown with overlapping ground truth (green skeleton).

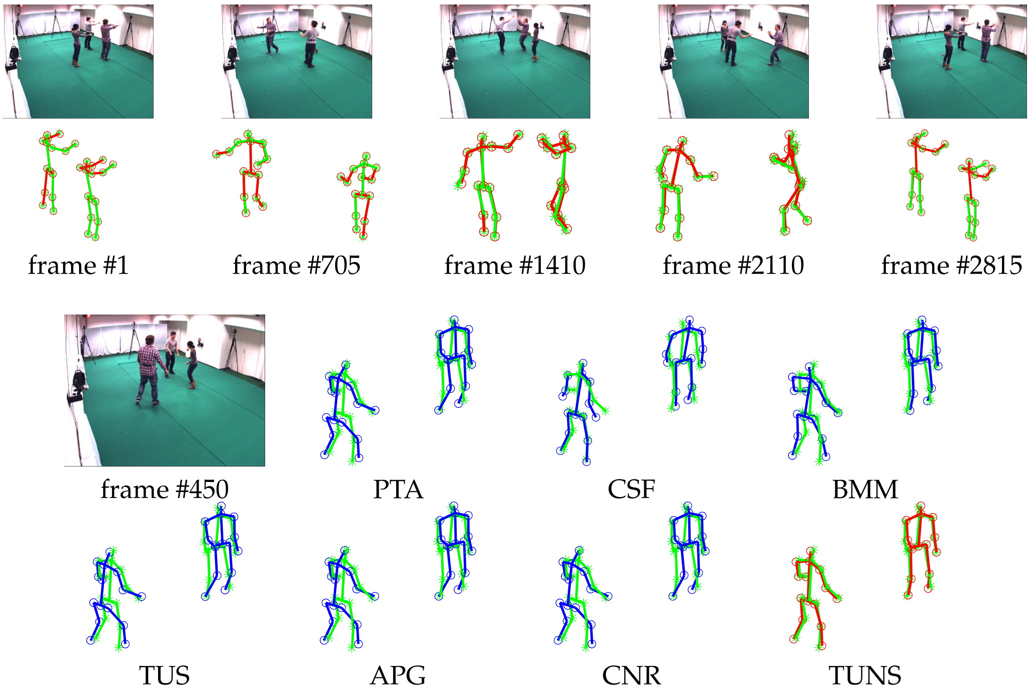

Figure 5. Evaluation on 'p3_ball_12' sequence in UMPM. For five selected frames, our reconstruction (red skeleton), ground truth (green skeleton) and the corresponding image are shown on top. On the bottom, the reconstructions, yielded by all the baselines (blue skeleton) and our TUNS (red skeleton) of one specific frame, are shown with overlapping ground truth (green skeleton).

Table 1. Quantitative evaluation on CMU motion capture sequences. The table reports the mean error e_{mean} and median error e_{med} of 3D reconstructions for the following NRSfM baselines: PTA [5], CSF [6], BMM [8], TUS [31], APG [9], CNR [14]; and our proposed method TUNS. For each sequence, the best and the second-best results of, both mean and median errors, are shown in **red** and **blue**, respectively.

Method / **Data**	**PTA** [5]	**CSF** [6]	**BMM** [8]	**TUS** [31]	**APG** [9]	**CNR** [14]	**TUNS (Ours)**
	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}
56_02	0.0227/0.0113	0.0500/0.0252	0.0235/0.0147	**0.0204**/**0.0112**	0.0215/**0.0102**	0.0341/0.0240	**0.0205**/0.0118
56_03	0.0655/0.0301	0.1309/0.0792	0.0748/0.0446	**0.0557**/**0.0254**	0.0739/0.0460	0.0605/0.0316	**0.0448**/**0.0219**
56_04	0.0720/0.0325	0.1819/0.1199	0.0843/0.0516	**0.0637**/**0.0287**	0.0792/0.0487	0.0661/0.0329	**0.0538**/**0.0278**
56_05	0.0697/0.0354	0.2056/0.1269	0.0857/0.0596	**0.0613**/**0.0312**	0.0741/0.0479	0.0629/0.0341	**0.0506**/**0.0287**
56_06	0.0951/0.0483	0.2412/0.1616	0.1085/0.0696	**0.0827**/**0.0402**	0.0975/0.0582	0.0829/0.0413	**0.0667**/**0.0321**
56_07	0.1259/0.0542	0.3446/0.2144	0.1453/0.0889	**0.0959**/**0.0424**	0.1262/0.0789	0.2066/0.0982	**0.0790**/**0.0385**
56_08	0.0807/0.0386	0.2015/0.1236	0.1158/0.0822	**0.0717**/**0.0427**	0.1415/0.1160	0.1910/0.1041	**0.0583**/**0.0303**
86_01	0.0700/0.0321	0.1591/0.0942	0.0832/0.0516	0.0619/**0.0267**	0.0642/0.0271	**0.0607**/0.0278	**0.0582**/**0.0237**
86_02	0.1817/0.0930	0.3170/0.2125	0.1716/0.0927	0.1521/0.0710	0.1586/0.0750	**0.1463**/**0.0683**	**0.1449**/**0.0646**
86_03	0.1861/0.1057	0.3696/0.2751	0.1766/0.1018	0.1554/0.0799	0.1578/0.0832	**0.1541**/**0.0795**	**0.1433**/**0.0733**
86_04	0.0934/0.0464	0.2384/0.1601	0.0975/0.0559	0.0826/0.0391	0.0854/0.0402	**0.0783**/**0.0363**	**0.0774**/**0.0351**
86_05	0.2267/0.1394	0.3387/0.2595	0.1975/0.1131	0.1821/0.0972	0.1872/0.1008	**0.1738**/**0.0938**	**0.1784**/**0.0949**
86_06	0.1765/0.0949	0.3724/0.2780	0.1745/0.0995	0.1570/0.0781	0.1617/0.0815	**0.1516**/**0.0761**	**0.1501**/**0.0734**
86_07	0.1457/0.0854	0.4838/0.3634	0.1447/0.0878	0.1293/0.0702	0.1331/0.0724	**0.1254**/**0.0677**	**0.1253**/**0.0688**
86_08	0.1333/0.0746	0.3672/0.2562	0.1366/0.0836	0.1191/0.0631	0.1227/0.0657	**0.1152**/**0.0608**	**0.1157**/**0.0609**
86_09	0.0302/0.0135	0.0728/0.0498	0.0412/0.0237	0.0306/0.0118	0.0310/0.0112	**0.0291**/**0.0107**	**0.0270**/**0.0096**
86_10	0.0636/0.0334	0.2489/0.1609	0.0681/0.0431	0.0534/0.0257	0.0552/0.0261	**0.0516**/**0.0245**	**0.0505**/**0.0232**
86_11	0.0727/0.0416	0.4362/0.3212	0.0729/0.0465	0.0605/0.0319	0.0630/0.0329	**0.0586**/**0.0306**	**0.0569**/**0.0304**
86_12	0.1190/0.0667	0.2714/0.2062	0.1225/0.0735	0.1102/**0.0596**	0.1131/0.0616	**0.1096**/0.0597	**0.1078**/**0.0584**
86_13	0.0676/0.0420	0.1261/0.0931	0.0777/0.0566	0.0586/0.0361	0.0684/0.0462	**0.0584**/**0.0348**	**0.0515**/**0.0276**
Average Err.	0.1049/0.0560	0.2579/0.1790	0.1101/0.0670	**0.0902**/**0.0456**	0.1008/0.0565	0.1008/0.0518	**0.0830**/**0.0417**
Relative Err.	1.2639/1.3429	3.1072/4.2926	1.3265/1.6067	**1.0867**/**1.0935**	1.2145/1.3549	1.2145/1.2422	**1**/**1**

Table 2. Quantitative evaluation on UMPM sequences. The table reports the mean error e_{mean} and median error e_{med} of 3D reconstructions for the following NRSfM baselines PTA [5], CSF [6], BMM [8], TUS [31], APG [9], CNR [14]; and our proposed method TUNS. For each sequence, the best and the second-best results of, both mean and median errors, are shown in **red** and **blue**, respectively.

Method Data	**PTA** [5]	**CSF** [6]	**BMM** [8]	**TUS** [31]	**APG** [9]	**CNR** [14]	**TUNS (Ours)**
	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}	e_{mean}/e_{med}
p1_grab_3	0.1036/0.0639	0.1619/0.1127	0.0976/0.0673	0.0882/0.0539	0.0908/0.0558	**0.0805**/**0.0492**	**0.0781**/**0.0560**
p1_chair_2	0.0763/0.0461	0.1889/0.1281	0.0892/0.0664	0.0736/**0.0429**	0.0736/0.0433	**0.0678**/**0.0404**	**0.0628**/0.0441
p3_ball_12	0.0431/0.0244	0.0930/0.0682	0.0709/0.0591	0.0423/0.0245	0.0414/**0.0225**	**0.0407**/0.0226	**0.0286**/**0.0168**

5.3. Qualitative Evaluation of Marker-Less Method

As described above, our method can be easily converted to the marker-less version. Cooperating with the CNN-based human joint detector, our method obtains accurate results as shown in Figures 6 and 7. Note that the employed detector is not restricted to the one proposed in [32]. Any detector that provides the heightmaps can be introduced into our framework. This proves that our method is effective in more practical scenarios.

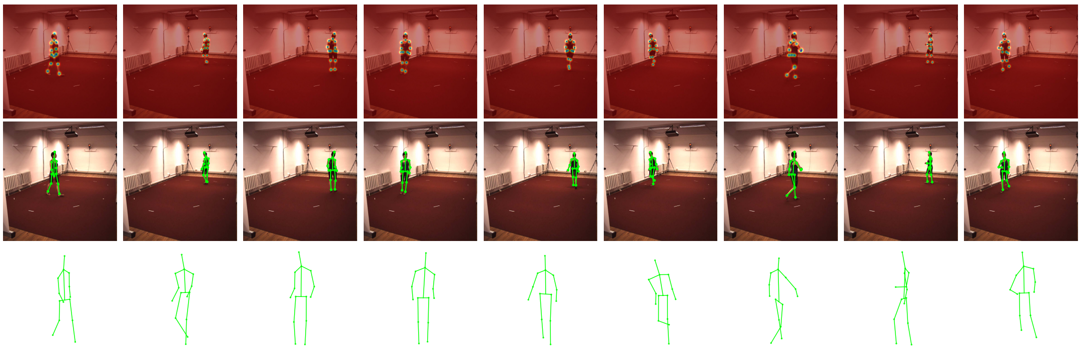

Figure 6. Evaluation on 'p3_ball_12' sequence in UMPM. For five selected frames, our reconstruction (red skeleton), ground truth (green skeleton) and the corresponding image are shown on top. On the bottom, the reconstructions, yielded by all the baselines (blue skeleton) and our TUNS (red skeleton) of one specific frame, are shown with overlapping ground truth (green skeleton).

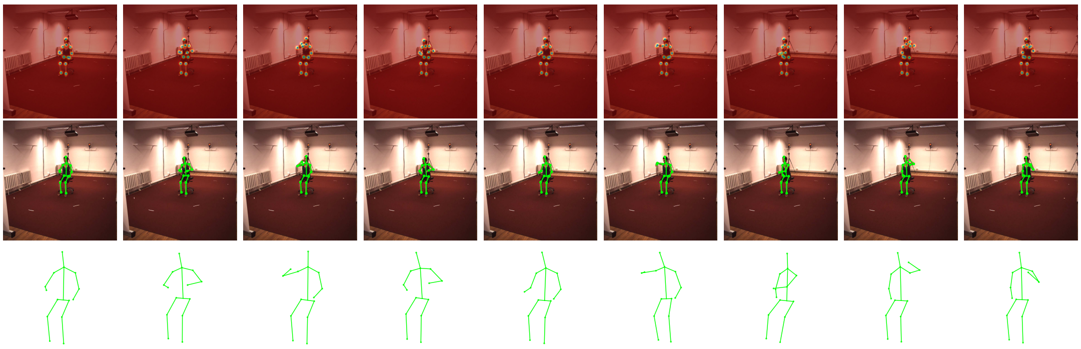

Figure 7. Evaluation of marker-less method on 'eating' sequence in Human3.6M datasets. The visualization of results from nine selected frames are shown from left to right. For each frame, the heightmaps of all the joints are displayed together in the first row, the projection of our reconstructed 3D human structure is plotted in the second row and the reconstructed 3D structure is shown in the bottom row.

5.4. Discussion

Our method is also limited by the reconstructability problem, the same as the previous NRSfM techniques. When the input sequence has a small range of viewpoints and the human body is extremely small, the method yields poor reconstructions. In such cases, more additional priors are needed to obtain the robust 3D reconstructions, such as pre-learned shape basis, trajectory basis or temporal consistency. Moreover, camera rotation estimation is not taken into account in this paper. Similar to the method in [31], how to optimize the camera rotations along with estimating the 3D structures under the assumption, i.e., union of subspaces, is still an open problem.

6. Conclusions

In this paper, an NRSfM approach is proposed to recover complex 3D human motions from monocular RGB images. Kernelized low-rank representation, which empirically proves to be more effective to represent complex human motions, is introduced to the NRSfM framework. Combining with the soft inextensibility constraint, our method produces more accurate reconstruction against the baseline approaches. The qualitative analysis illustrates that our KLRR-NRSfM method can be conveniently converted to the mark-less version without dependency on the given tracks. However, some issues are still open for future research. For example, promoting the reconstruction accuracy is a challenging task when the range of viewpoints is extremely small, and estimating the camera rotations under the union of nonlinear shape subspaces assumption is also a future work, which is crucial for applying the method to more practical real-life scenarios. Additionally, the 2D locations of joints might be lost due to the occlusion. In the marker-based method, the 2D location of joints is often obtained by tracking the infrared markers. By contrast, in the marker-less case, the CNN-based detector is relatively robust. Nonetheless, an explicit mechanism is still needed to handle the occlusion for future works.

Author Contributions: Xuan Wang presented and implemented the solution; finished and analysed experiments; Meanwhile wrote and revised the manuscript. Fei Wang proposed the research issue, the idea and the experimental design. Meanwhile he supervised the research by providing suggestions. Yanan Chen re-implemented the baselines.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bregler, C.; Hertzmann, A.; Biermann, H. Recovering non-rigid 3D shape from image streams. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, USA, 15 June 2000; pp. 690–696.
2. Tomasi, C.; Kanade, T. Shape and motion from image streams under orthography: A factorization method. *Int. J. Comput. Vis.* **1992**, *9*, 137–154.
3. Xiao, J.; Chai, J.; Kanade, T. A closed-form solution to non-rigid shape and motion recovery. *Int. J. Comput. Vis.* **2006**, *67*, 233–246.
4. Akhter, I.; Sheikh, Y.; Khan, S.; Kanade, T. Nonrigid structure from motion in trajectory space. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 41–48.
5. Ahkter, I.; Sheikh, Y.; Khan, S.; Kanade, T. Trajectory Space: A Dual Representation for Nonrigid Structure from Motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1442–1456.
6. Gotardo, P.F.U.; Martinez, A.M. Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2051–2065.
7. Akhter, I.; Sheikh, Y.; Khan, S. In defense of orthonormality constraints for nonrigid structure from motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1534–1541.

8. Dai, Y.; Li, H.; He, M. A simple prior-free method for non-rigid structure-from-motion factorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2018–2025.
9. Wang, Y.; Tong, L.; Jiang, M.; Zheng, J. Non-Rigid Structure Estimation in Trajectory Space from Monocular Vision. *Sensors* **2015**, *15*, 25730–25745.
10. Gotardo, P.F.U.; Martinez, A.M. Non-rigid structure from motion with complementary rank-3 spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3065–3072.
11. Lee, M.; Cho, J.; Choi, C.; Oh, S. Procrustean normal distribution for non-rigid structure from motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 Jun 2013; pp. 1280–1287.
12. Simon, T.; Valmadre, J.; Matthews, I.; Sheikh, Y. Separable spatiotemporal priors for convex reconstruction of time-varying 3D point clouds. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 204–219.
13. Zhu, Y.; Lucey, S. Convolutional sparse coding for trajectory reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 529–540.
14. Lee, M.; Cho, J.; Oh, S. Consensus of non-rigid reconstructions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
15. Burenius, M.; Sullivan, J.; Carlsson, S. 3D pictorial structures for multiple view articulated pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3618–3625.
16. Oh, H.; Cha, G.; Oh, S. Samba: A Real-Time Motion Capture System Using Wireless Camera Sensor Networks. *Sensors* **2014**, *14*, 5516–5535.
17. Tsai, M.H.; Chen, K.H.; l-Chen, L. Real-time Upper Body Pose Estimation from Depth Images. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015.
18. Michel, D.; Panagiotakis, C.; Argyros, A.A. Tracking the articulated motion of the human body with two RGBD cameras. *Mach. Vis. Appl.* **2015**, *26*, 41–54.
19. Rádlová, R.; Bouwmans, T.; Vachon, B. Models Used by vision—Based motion capture. In Proceedings of the Computer Graphics and Artificial Intelligence (3IA), Limoges, France, 23–24 May 2006.
20. Sigal, L. Human pose estimation. In *Computer Vision: A Reference Guide*; Springer US: New York, NY, USA, 2014; pp. 362–370.
21. Andriluka, M.; Sigal, L. Human context: Modeling human-human interactions for monocular 3D pose estimation. In Proceedings of the Articulated Motion and Deformable Objects, Mallorca, Spain, 11–13 July 2012.
22. Andriluka, M.; Roth, S.; Schiele, B. Monocular 3D pose estimation and tracking by detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 623–630.
23. Yao, A.; Gall, J.; Gool, L.V.; Urtasun, R. Learning probabilistic non-linear latent variable models for tracking complex activities. In Proceedings of the Advances in Neural Information Processing System, Granada, Spain, 12–17 December 2011; pp. 1359–1367.
24. Taylor, G.W.; Sigal, L.; Fleet, D.J.; Hinton, G.E. Dynamical binary latent variable models for 3D human pose tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 631–638.
25. Tian, Y.; Sigal, L.; De la Torre, F.; Jia, Y. Canonical locality preserving latent variable model for discriminative pose inference. *Image Vis. Comput.* **2013**, *31*, 223–230.
26. Zhu, H.; Yu, Y.; Zhou, Y.; Du, S. Dynamic Human Body Modeling Using a Single RGB Camera. *Sensors* **2016**, *16*, 402.
27. Ek, C.H.; Torr, P.H.; Lawrence, N.D. Gaussian process latent variable models for human pose estimation. In Proceedings of the Machine Learning for Multimodal Interaction, Utrecht, The Netherlands, 8–10 September 2008; pp. 132–143.
28. Tekin, B.; Rozantsev, A.; Lepetit, V.; Fua, P. Direct prediction of 3D body poses from motion compensated sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

29. Elhayek, A.; de Aguiar, E.; Jain, A.; Tompson, J.; Pishchulin, L.; Andriluka, M.; Bregler, C.; Shiele, B.; Theobalt, C. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 3810–3818.
30. Gong, W.; Zhang, X.; Gonzàlez, J.; Sobral, A.; Bouwmans, T.; Tu, C.; Zahzah, E.-H. Human Pose Estimation from Monocular Images: A Comprehensive Survey. *Sensors* **2016**, *16*, 1966.
31. Zhu, Y.; Huang, D.; De La Torre, F.; Lucey, S. Complex non-rigid motion 3D reconstruction by union of subspaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 Jun 2014; pp. 1542–1549.
32. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
33. Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Trans Pattern Anal. Mach. Intell.* **2013**, *35*, 171–184.
34. Xiao, S.; Tan, M.; Xu, D.; Dong, Z.Y. Robust kernel low-rank representation. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 2268–2281.
35. Cai, J.; Candès, E.J.; Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **2010**, *20*, 1956–1982.
36. Moré, J.J. The Levenberg-Marquardt Algorithm: Implementation and Theory. *Numerical Analysis, Lecture Notes in Mathematics* **1977**, *630*, 105–116.
37. Lin, Z.; Chen, M.; Wu, L.; Ma, Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv* **2009**, arXiv:1009.5055.
38. Aa, N.; Luo, X.; Giezeman, G.; Tan, R.; Veltkamp, R. Utrecht multi-person motion (umpm) benchmark: A multiperson dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In Proceedings of the International Conference on Computer Vision Workshop HICV, Barcelona, Spain, 6–13 November 2011.