Spatial-Temporal Union of Subspaces for Multi-body Non-rigid Structure-from-Motion

Suryansh Kumar[1], Yuchao Dai[1], Hongdong Li[1,2]

[1]Research School of Engineering, The Australian National University
[2] Australian Centre for Robotic Vision.
{suryansh.kumar, yuchao.dai, hongdong.li}@anu.edu.au

Abstract. Non-rigid structure-from-motion (NRSfM) has so far been mostly studied for recovering 3D structure of a single non-rigid/deforming object. To handle the real world challenging multiple deforming objects scenarios, existing methods either pre-segment different objects in the scene or treat multiple non-rigid objects as a whole to obtain the 3D non-rigid reconstruction. However, these methods fail to exploit the inherent structure in the problem as the solution of segmentation and the solution of reconstruction could not benefit each other. In this paper, we propose a unified framework to jointly segment and reconstruct multiple non-rigid objects. To compactly represent complex multi-body non-rigid scenes, we propose to exploit the structure of the scenes along both temporal direction and spatial direction, thus achieving a spatio-temporal representation. Specifically, we represent the 3D non-rigid deformations as lying in a union of subspaces along the temporal direction and represent the 3D trajectories as lying in the union of subspaces along the spatial direction. This spatio-temporal representation not only provides competitive 3D reconstruction but also outputs robust segmentation of multiple non-rigid objects. The resultant optimization problem is solved efficiently using the Alternating Direction Method of Multipliers (ADMM). Extensive experimental results on both synthetic and real multi-body NRSfM datasets demonstrate the superior performance of our proposed framework compared with the state-of-the-art methods [1].

Keywords: Structure from Motion (SfM), Subspace Clustering, Alternating Direction Method of Multipliers (ADMM), Deformable Objects.

1 Introduction

Aiming at recovering the camera motion and non-rigid structure simultaneously from 2D images emanating from monocular cameras, non-rigid structure from motion (NRSfM) is central to many computer vision applications and has received considerable attention in recent years. This classical problem is highly under-constrained. Although existing approaches in NRSfM [6] [8] [24] [14] [4]

[1] This work was completed and submitted to ACCV on 27th May 2016 for review. "The author version of the paper has been accepted by Pattern Recognition".

have presented promising results but most of these methods assume that, there is only one object undergoing non-rigid deformation in the scene. However, real world non-rigid scenes are much more complex: for example multiple persons performing different activities, soccer players in the playground, salsa dance and etc. All these real world examples constitute multi-body non-rigid deformation, which could not be explained well with the single non-rigid object assumption. Therefore, it is quite natural to extend single-body NRSfM to multi-body NRSfM where the task would be to jointly reconstruct and segment multiple 3D deforming objects over-time.

In solving the problem of multi-body NRSfM, a natural and direct two-stage process is to reconstruct non-rigid multi-body structure by applying state-of-the-art non-rigid reconstruction methods[9][18] [29] and then segment distinct objects using subspace clustering methods such as Sparse Subspace Clustering (SSC) [12] or other clustering algorithms or vice-versa. However, by adopting such pipelines the inherent structure of the problem has never been exploited, i.e non-rigid motion segmentation provides critical information to constrain 3D reconstruction while 3D non-rigid reconstruction could also constrain the corresponding motion segmentation problem. Furthermore, since the non-rigid shape deformation actually occurs in 3D space, it is more intuitive to perform segmentation of objects in 3D space rather than on projected 2D image space.

Additionally, it is always convenient–both computationally and numerically to solve a given task using a unified approach than solving it in a sequential way. Therefore, in this paper, we propose a framework to simultaneously reconstruct and cluster multiple non-rigid shapes by exploiting the spatio-temporal correlation in data. By such approach we can explain the dynamics of non-rigid shape in a more intuitive way. Explicitly, we represent multi-body NRSfM as union of subspace both in 3D trajectory space (spatially) and 3D shape space (temporally). We use the fact that each 3D trajectory can be expressed with other trajectory only if the trajectory is from the same subspace (spatial clustering) [17], and each individual activity can be expressed with activity belonging to the same subspace (temporal clustering) [29]. A visual illustration of the spatio-temporal subspace concept is presented in Fig. 1. Concretely, spatial clustering tries to reconstruct a trajectory using affine combination of other trajectories from the same deforming object, while temporal clustering tries to explain the shape of deforming objects using affine combination of other shapes at different frame instance belonging to similar activity.

By exploiting the spatio-temporal clustering structure, our approach is able to learn the affinity matrices which naturally encode subspace information. From the affinity matrices, direct inference about number of deformable objects, different activities and membership of each sample to achieve reconstruction can be easily made. Furthermore, we exploit the fact that the connectivity between subspaces must be tight if it belongs to the same subspace and loose if belongs to different subspaces. Therefore, we propose to use a mixture of ℓ_1 norm and ℓ_2 norm regularization (also known as the Elastic Net [31]), which helps in controlling the sparsity of the affinity matrices.

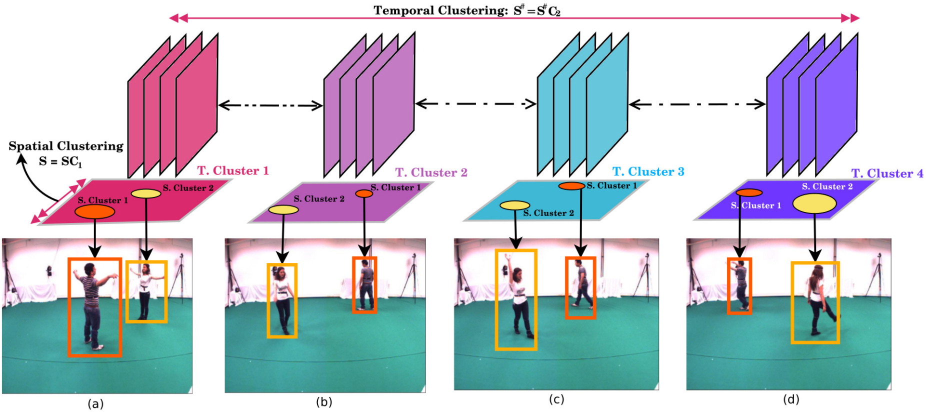


Fig. 1. Illustration of the two clustering constraints used in our framework. We observe that, when different objects are undergoing complex non-rigid motion, the temporal clustering helps in improving the 3D reconstruction by clustering different activities over-time such as stretch, walking, jumping and etc. The spatial clustering helps in explaining the segmentation of distinct structures over images. Frames with similar activities are shown in the same colors and different subjects undergoing deformations are shown in box. Here, **T. Cluster** refers to the Temporal cluster and **S. Cluster** refers to the Spatial Cluster. This flow diagram demonstrates that subjects performing different activities over-time lie in distinct temporal subspace and spatial subspace, subsequently different 3D trajectories spanned by different structures lies in distinct subspace. The example images are collected from the UMPM dataset [1]. (Best viewed on screen in color)

Contributions:

1. We propose a joint segmentation and reconstruction framework to the challenging task of complex multi-body NRSfM by exploiting the inherent spatio-temporal union of subspace constraint.
2. We propose to efficiently solve the resultant non-convex optimization problem based on the Alternating Direction Method of Multipliers (ADMM) method [5].
3. Extensive experimental results on both synthetic and real multi-body NRSfM datasets demonstrate the superior performance of our proposed framework.

2 Related Works

Multi-body structure from motion (SfM) is an important problem in computer vision. To work out this problem for rigid motion is a direct extension to elegant multi-view geometry techniques [13][20]. However, solution to multi-body NRSfM is not straightforward, due to the difficulty in modeling complex non-rigid variations. Recent state-of-the-art in NRSfM reconstruction [9] has shown promising results while Zhu et al. [29] proposed that such an approach may fail while modeling long-term complex non-rigid motions. The work quote that Dai et al. [8] work is "highly dependent on the complexity of the motion" [29]. Hence, to overcome this difficulty they suggested to represent long-term non-rigid motion as union of subspace rather than a single subspace. Subsequently, Cho et al. [7] used probabilistic variations to model complex shape.

Despite the above accomplishments, NRSfM is still far behind its rigid counterpart. This gap is principally due to difficulty in modeling real world non-rigid deformation. If the deformation is irregular or arbitrary then to explain the 3D structure is nearly impossible. Nevertheless, many real world deformation can be constrained; as a result Bergler [6] introduced NRSfM which is considered a seminal work in NRSfM. In the work, Bergler demonstrated that non-rigid deformation can be represented by a linear combination of a set of shape basis. Following the work, several researchers tried to model NRSfM by utilizing additional constraints [25], [27], [21]. In 2008, Akhter et al. [4] presented a dual approach by modeling 3D trajectories. In 2009, Akhter et al. [3] proved that even there is an ambiguity in shape bases or trajectory bases, non-rigid shapes can still be solved uniquely without any ambiguity. In 2012, Dai et al. [8] proposed a "prior-free" method to recover camera motion and 3D non-rigid deformation by exploiting low rank constraint only. Besides shape basis model and trajectory basis model, the shape-trajectory approach [16] combines two models and formulates the problems as revealing trajectory of the shape basis coefficients. Besides linear combination model, Lee et al. [18] proposed a Procrustean Normal Distribution (PND) model, where 3D shapes are aligned and fit into a normal distribution. Simon et al. [23] exploited the Kronecker pattern in the shape-trajectory (spati-temporal) priors. Zhu and Lucey [30] applied the convolutional sparse coding technique to NRSFM using point trajectories. However, the method requires to learn an over-complete basis of 3D trajectories, prior to performing 3D reconstruction.

Recently, Russell et al. [22] proposed to simultaneously segment a complex dynamic scene containing a mixture of multiple objects into constituent objects and reconstruct a 3D model of the scene by formulating the problem as hierarchical graph-cut based segmentation, where the whole scene is decomposed into background and foreground objects with complex motion of non-rigid or articulated objects are modeled as a set of overlapping rigid parts.

Our method varies from the aforementioned works in the following aspects: 1) We provide a novel framework to joint segmentation and reconstruction for multiple non-rigid deformation problem; 2) We propose a simple, yet efficient and elegant optimization routine and its solution based on ADMM; 3) Our method can be applied to both sparse and dense scenarios (up to the order of ten-thousand feature tracks).

A part of this work has been published in 3DV 2016 [17], which addressed multi-body NRSfM by using the spatial constraint only. The work of [17] can be viewed as a special case of the present work.

3 Formulation

Under our formulation, we intend to reconstruct 3D non-rigid shapes such that they satisfy both the spatio-temporal union of affine subspace constraint and the non-rigid shape constraints (low rank and spatial coherency). Let $\mathtt{W} \in \mathbb{R}^{2\mathtt{F}\times\mathtt{P}}$ represent the *measurement matrix*, with $\mathtt{F}$ the number of frames and $\mathtt{P}$ the number of feature points. We use the *orthographic camera* model and eliminate the translation component of camera motions as suggested in [6].

$$\mathtt{W} = \mathtt{RS}, \tag{1}$$

where $\mathtt{R} = \text{blkdiag}(\mathtt{R}_1, \cdots, \mathtt{R}_\mathtt{F}) \in \mathbb{R}^{2\mathtt{F}\times 3\mathtt{F}}$ denotes the camera rotation matrix and $\mathtt{S}$ represents the 3D shapes of deforming objects over entire frames. This classical representation for NRSfM problem [6] aims at recovering both the *camera motion* $\mathtt{R}$ and the non-rigid 3D shapes $\mathtt{S} \in \mathbb{R}^{3\mathtt{F}\times\mathtt{P}}$ from the 2D *measurement matrix* $\mathtt{W} \in \mathbb{R}^{2\mathtt{F}\times\mathtt{P}}$ such that $\mathtt{W} = \mathtt{RS}$. Following the same representation to cater 2D-3D relation, we use $\|\mathtt{W} - \mathtt{RS}\|_\mathtt{F}^2$ to infer the re-projection error.

3.1 Representing multiple non-rigid deformations in trajectory space

To represent multiple non-rigid objects using a single linear trajectory space does not provide compact representation of 3D trajectories [29]. When there are multiple non-rigid objects, each object can be characterized as lying in an affine subspace. Therefore, the 3D trajectories lie in a union of affine subspaces, which can equivalently be formulated in terms of self-expressiveness i.e,

$$\mathtt{S} = \mathtt{SC}_1, \text{diag}(\mathtt{C}_1) = \mathtt{0}, \mathtt{1}^\mathtt{T}\mathtt{C}_1 = \mathtt{1}^\mathtt{T}. \tag{2}$$

where $\mathtt{S} \in \mathbb{R}^{3\mathtt{F}\times\mathtt{P}}, \mathtt{C}_1 \in \mathbb{R}^{\mathtt{P}\times\mathtt{P}}$. To get rid of the trivial solution of $\mathtt{S} = \mathtt{S}$ or $\mathtt{C}_1 = \mathtt{I}$, we explicitly enforce the diagonal constraint as $\text{diag}(\mathtt{C}_1) = \mathtt{0}$. As we represent each non-rigid object as lying in an affine subspace, we further enforce the affine constraint $\mathtt{1}^\mathtt{T}\mathtt{C}_1 = 1^\mathtt{T}$. Besides the above constraint, we also want to enforce a constraint that if the trajectories belong to the same deforming object then it must be tightly connected or loosely connected the otherwise. To cater this idea of inter-class and intra-class trajectories clustering, we use the elastic net formulation [28] to compromise between connectedness and sparsity. Combining all the constraints together, we reach the following optimization:

$$\begin{aligned} &\underset{\mathtt{C}_1}{\text{minimize}}\ \lambda_1\|\mathtt{C}_1\|_1 + \frac{(1-\lambda_1)}{2}\|\mathtt{C}_1\|_\mathtt{F}^2 \\ &\textit{subject to:} \\ &\mathtt{S} = \mathtt{SC}_1, \text{diag}(\mathtt{C}_1) = \mathtt{0}, \mathtt{1}^\mathtt{T}\mathtt{C}_1 = \mathtt{1}^\mathtt{T}, \lambda_1 \in [0, 1]. \end{aligned} \tag{3}$$

A visual illustration of this idea in trajectory space for a single trajectory is provided in Fig. 2. Here, $\|.\|_1$ and $\|.\|_\mathtt{F}$ denote the ℓ_1-norm and the Frobenius norm respectively.

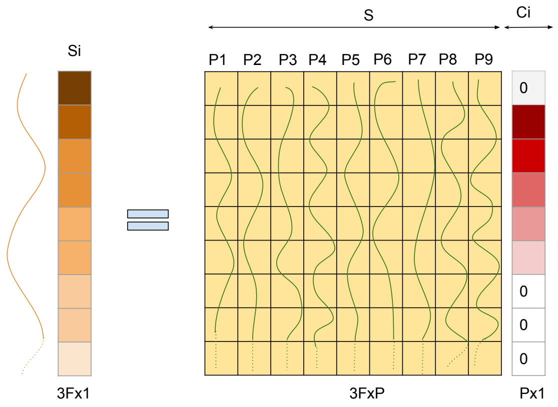
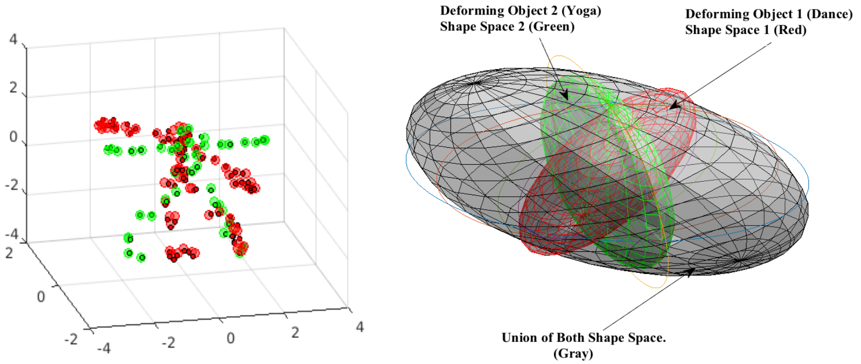


Fig. 2. Visual illustration of the affine subspace constraint $\mathtt{S_i} = \mathtt{SC_i}$ in trajectory space. Each column of $\mathtt{S}$ is a trajectory of a 3D point (shown in green). This visualization states that a trajectory $\mathtt{S_i}$ can be reconstructed using affine combination of few other trajectories. *Note* : This pictorial representation is provided for better understanding and is only for illustration purpose. (Best viewed in color)



(a) (b)

Fig. 3. Visual representation of union of subspace in shape space. (a) Two different subjects are performing Dance (Red) and Yoga (Green) respectively. (b) Equivalent representation of both activities in shape space for a single frame with green ellipsoid showing the shape space for Yoga activity and red ellipsoid showing the Dance activity. It can be observed that the space spanned by different shapes performing different activities span a distinct subspace. Gray color ellipsoid shows the union of both subspaces. (Best viewed in color)

3.2 Representing multiple non-rigid deformations in shape space

An example complex non-rigid motion is shown in Figure 1, where the subjects are performing different activities at different time instances. Such distinct motion adheres to different local subspace and complete non-rigid motion lies in union of shape subspace. As mentioned in [29] such assumption leads to superior 3D reconstruction. To incorporate this concept in our formulation that different activities lie in union of affine subspaces, we express the 3D shapes in terms of self-expressiveness of frames along temporal direction.

$$\mathtt{S}^{\sharp} = \mathtt{S}^{\sharp}\mathtt{C}_2, \mathrm{diag}(\mathtt{C}_2) = \mathtt{0}, \mathtt{1}^{\mathtt{T}}\mathtt{C}_2 = \mathtt{1}^{\mathtt{T}}. \tag{4}$$

where $\mathtt{S}^{\sharp} \in \mathbb{R}^{3\mathtt{P}\times\mathtt{F}}$ is the reshuffled version of $\mathtt{S}$ representing the per-frame 3D shape as a column vector, $\mathtt{C}_2 \in \mathbb{R}^{\mathtt{F}\times\mathtt{F}}$. A visual intuition of this idea in shape space for single frame is provided in Fig. 3.

For temporal clustering, we also use the elastic net as regularization parameters due to similar reason mentioned in Section 3.1 for $\mathtt{C}_2$, thereby formulating the following optimization:

$$\begin{aligned} &\underset{\mathtt{C}_2}{\mathrm{minimize}}\ \lambda_3\|\mathtt{C}_2\|_1 + \frac{(1-\lambda_3)}{2}\|\mathtt{C}_2\|_{\mathtt{F}}^2 \\ &\textit{subject to:} \\ &\mathtt{S}^{\sharp} = \mathtt{S}^{\sharp}\mathtt{C}_2, \mathrm{diag}(\mathtt{C}_2) = \mathtt{0}, \mathtt{1}^{\mathtt{T}}\mathtt{C}_2 = \mathtt{1}^{\mathtt{T}}, \lambda_3 \in [0,1]. \end{aligned} \tag{5}$$

3.3 Enforcing the global shape constraint

In seeking a compact representation for multi-body non-rigid objects, we penalize the number of independent non-rigid shapes. Similar to [8] and [14], we penalize the nuclear norm of the reshuffled shape matrix $\mathtt{S}^{\sharp} \in \mathtt{R}^{3\mathtt{P}\times\mathtt{F}}$, this is because the nuclear norm is known as the convex envelope of the rank function. In this way, the global shape constraint is expressed as:

$$\|\mathtt{S}^{\sharp}\|_*, \tag{6}$$

where $\|\|_*$ denotes the nuclear norm of the matrix, ie, sum of singular values.

3.4 Joint Reconstruction and Segmentation Formulation

Putting all the above constraints (spatio-temporal union of subspace constraint and global shape constraint) together, we reach a multi-body non-rigid reconstruction and segmentation formulation:

$$\begin{aligned}
&\underset{\mathtt{S},\mathtt{C}_1,\mathtt{C}_2}{\text{minimize}}\ \frac{1}{2}\|\mathtt{W}-\mathtt{R}\mathtt{S}\|_{\mathtt{F}}^2+\lambda_1\|\mathtt{C}_1\|_1+\frac{1-\lambda_1}{2}\|\mathtt{C}_1\|_{\mathtt{F}}^2+\lambda_2\|\mathtt{S}^\sharp\|_*+\lambda_3\|\mathtt{C}_2\|_1+\frac{1-\lambda_3}{2}\|\mathtt{C}_2\|_{\mathtt{F}}^2\\
&\textit{subject to:}\\
&\mathtt{S}=\mathtt{S}\mathtt{C}_1,\mathtt{S}^\sharp=\mathtt{S}^\sharp\mathtt{C}_2,\\
&\mathtt{1}^{\mathtt{T}}\mathtt{C}_1=\mathtt{1}^{\mathtt{T}},\mathtt{1}^{\mathtt{T}}\mathtt{C}_2=\mathtt{1}^{\mathtt{T}},\\
&\text{diag}(\mathtt{C}_1)=\mathtt{0},\text{diag}(\mathtt{C}_2)=\mathtt{0},\\
&\lambda_1,\lambda_3\in[0,1].
\end{aligned}\tag{7}$$

where $\mathtt{S}^\sharp\in\mathbb{R}^{\mathtt{3P}\times\mathtt{F}}$, $\mathtt{C}_1\in\mathbb{R}^{\mathtt{P}\times\mathtt{P}}$, and $\mathtt{C}_2\in\mathbb{R}^{\mathtt{F}\times\mathtt{F}}$. $\lambda_1,\lambda_2,\lambda_3$ are the trade-off parameters.

4 Solution

To solve the proposed optimization we introduce decoupling variables in Eq. 7, which leads to the following formulation:

$$\begin{aligned}
&\underset{\mathtt{S},\mathtt{J},\mathtt{E}_1,\mathtt{E}_2,\mathtt{C}_1,\mathtt{C}_2,\mathtt{S}^\sharp}{\text{minimize}}\ \frac{1}{2}\|\mathtt{W}-\mathtt{R}\mathtt{S}\|_{\mathtt{F}}^2+\lambda_1\|\mathtt{E}_1\|_1+\frac{1-\lambda_1}{2}\|\mathtt{E}_1\|_{\mathtt{F}}^2+\lambda_2\|\mathtt{J}\|_*+\lambda_3\|\mathtt{E}_2\|_1+\frac{1-\lambda_3}{2}\|\mathtt{E}_2\|_{\mathtt{F}}^2\\
&\textit{subject to:}\\
&\mathtt{S}^\sharp=\mathtt{g}(\mathtt{S}),\mathtt{S}^\sharp=\mathtt{J},\\
&\mathtt{S}=\mathtt{S}\mathtt{C}_1,\mathtt{S}^\sharp=\mathtt{S}^\sharp\mathtt{C}_2,\\
&\mathtt{1}^{\mathtt{T}}\mathtt{C}_1=\mathtt{1}^{\mathtt{T}},\mathtt{1}^{\mathtt{T}}\mathtt{C}_2=\mathtt{1}^{\mathtt{T}},\\
&\text{diag}(\mathtt{C}_1)=\mathtt{0},\text{diag}(\mathtt{C}_2)=\mathtt{0},\\
&\mathtt{C}_1=\mathtt{E}_1,\mathtt{C}_2=\mathtt{E}_2,\\
&\lambda_1,\lambda_3\in[0,1].
\end{aligned}\tag{8}$$

The auxiliary variables $\mathtt{E}_1,\mathtt{E}_2,\mathtt{J}$ are introduced to simplify the derivation. $\mathtt{g}(.):\mathtt{S}_{\mathtt{3F}\times\mathtt{P}}\rightarrow\mathtt{S}^\sharp_{\mathtt{3P}\times\mathtt{F}}$ denotes the linear mapping from $\mathtt{S}\in\mathbb{R}^{\mathtt{3F}\times\mathtt{P}}$ to its reshuffled version $\mathtt{S}^\sharp\in\mathbb{R}^{\mathtt{3P}\times\mathtt{F}}$. Specifically, $\text{S}=\begin{bmatrix}\mathtt{X}_{11}\ \mathtt{X}_{12}\ \mathtt{X}_{13}\ \ldots\ \mathtt{X}_{1\mathtt{P}}\\ \mathtt{Y}_{11}\ \mathtt{Y}_{12}\ \mathtt{Y}_{13}\ \ldots\ \mathtt{Y}_{1\mathtt{P}}\\ \mathtt{Z}_{11}\ \mathtt{Z}_{12}\ \mathtt{Z}_{13}\ \ldots\ \mathtt{Z}_{1\mathtt{P}}\\ \ldots\ldots\ldots\ldots\ldots\ldots\\ \mathtt{X}_{\mathtt{F}1}\ \mathtt{X}_{\mathtt{F}2}\ \mathtt{X}_{\mathtt{F}3}\ \ldots\ \mathtt{X}_{\mathtt{FP}}\\ \mathtt{Y}_{\mathtt{F}1}\ \mathtt{Y}_{\mathtt{F}2}\ \mathtt{Y}_{\mathtt{F}3}\ \ldots\ \mathtt{Y}_{\mathtt{FP}}\\ \mathtt{Z}_{\mathtt{F}1}\ \mathtt{Z}_{\mathtt{F}2}\ \mathtt{Z}_{\mathtt{F}3}\ \ldots\ \mathtt{Z}_{\mathtt{FP}}\end{bmatrix}$ and $\mathtt{S}^\sharp=\begin{bmatrix}\mathtt{X}_{11}\ldots\mathtt{X}_{1\mathtt{P}}\ \mathtt{Y}_{11}\ldots\mathtt{Y}_{1\mathtt{P}}\ \mathtt{Z}_{11}\ldots\mathtt{Z}_{1\mathtt{P}}\\ \mathtt{X}_{21}\ldots\mathtt{X}_{2\mathtt{P}}\ \mathtt{Y}_{21}\ldots\mathtt{Y}_{2\mathtt{P}}\ \mathtt{Z}_{21}\ldots\mathtt{Z}_{2\mathtt{P}}\\ \ldots\qquad\ldots\qquad\ldots\\ \mathtt{X}_{\mathtt{F}1}\ldots\mathtt{X}_{\mathtt{FP}}\ \mathtt{Y}_{\mathtt{F}1}\ldots\mathtt{Y}_{\mathtt{FP}}\ \mathtt{Z}_{\mathtt{F}1}\ldots\mathtt{Z}_{\mathtt{FP}}\end{bmatrix}^{\mathtt{T}}$. The first term in the above optimization is meant for penalizing re-projection error under *orthographic* projection. Under single-body NRSFM configuration, 3D shape $\mathtt{S}$ can be well characterized as lying in a single low dimensional linear subspace. However, when there are multiple non-rigid objects, each non-rigid object could be characterized as lying in an affine subspace. To represent this idea mathematically in shape and trajectory space respectively, we introduce $\mathtt{E}_1$ and $\mathtt{E}_2$.

In addition to this, to reveal the intrinsic structure of multi-body non-rigid structure-from-motion (NRSfM), we seek for the sparsest solution both in trajectory and shape space. Consequently, we enforce the ℓ_1 norm for $\mathtt{E}_1$ and $\mathtt{E}_2$. However, high sparsity may lead to misclassification of samples or trajectories. Therefore, to maintain the balance between sparsity and connectedness, we incorporate the elastic net for both $\mathtt{E}_1$ and $\mathtt{E}_2$. Lastly, we enforce a global shape constraint ($\|\mathtt{J}\|_*$) for compact representation of multi-body non-rigid objects by penalizing the rank of the entire non-rigid shape.

Due to the two bilinear terms $\mathtt{S} = \mathtt{S}\mathtt{C}_1$ and $\mathtt{S}^\sharp = \mathtt{S}^\sharp\mathtt{C}_2$, the overall optimization of Eq.-(8) is non-convex. We solve it via the alternating direction method of multipliers (ADMM), which has a proven effectiveness for many non-convex problems and is widely used in computer vision. ADMM works by decomposing the original optimization problem into several sub-problems, where each sub-problem can be solved efficiently. To this end, we seek to decompose Eq.-(8) into several sub-problems.

We introduce Lagrangian multipliers in the equation (8) and reach the Augmented Lagrangian formulation for Eq.-(8)

$$
\begin{aligned}
&\mathcal{L}(\mathtt{S},\mathtt{S}^\sharp,\mathtt{C}_1,\mathtt{C}_2,\mathtt{E}_1,\mathtt{E}_2,\mathtt{J},\{\mathtt{Y}_i\}_{\mathtt{i=1}}^{8}) = \frac{1}{2}\|\mathtt{W}-\mathtt{RS}\|_{\mathtt{F}}^2 + \lambda_1\|\mathtt{E}_1\|_1 + \gamma_1\|\mathtt{E}_1\|_{\mathtt{F}}^2 + \lambda_2\|\mathtt{J}\|_* + \\
&\lambda_3\|\mathtt{E}_2\|_1 + \gamma_3\|\mathtt{E}_2\|_{\mathtt{F}}^2 + <\mathtt{Y}_1,\mathtt{S}^\sharp-\mathtt{g}(\mathtt{S})> + \frac{\beta}{2}\|\mathtt{S}^\sharp-\mathtt{g}(\mathtt{S})\|_{\mathtt{F}}^2 + <\mathtt{Y}_2,\mathtt{S}-\mathtt{SC}_1> + \\
&\frac{\beta}{2}\|\mathtt{S}-\mathtt{SC}_1\|_{\mathtt{F}}^2 + <\mathtt{Y}_3,\mathtt{S}^\sharp-\mathtt{S}^\sharp\mathtt{C}_2> + \frac{\beta}{2}\|\mathtt{S}^\sharp-\mathtt{S}^\sharp\mathtt{C}_2\|_{\mathtt{F}}^2 + <\mathtt{Y}_4,\mathtt{1}^{\mathtt{T}}\mathtt{C}_1-\mathtt{1}^{\mathtt{T}}> + \\
&\frac{\beta}{2}\|\mathtt{1}^{\mathtt{T}}\mathtt{C}_1-\mathtt{1}^{\mathtt{T}}\|_{\mathtt{F}}^2 + <\mathtt{Y}_5,\mathtt{1}^{\mathtt{T}}\mathtt{C}_2-\mathtt{1}^{\mathtt{T}}> + \frac{\beta}{2}\|\mathtt{1}^{\mathtt{T}}\mathtt{C}_2-\mathtt{1}^{\mathtt{T}}\|_{\mathtt{F}}^2 + <\mathtt{Y}_6,\mathtt{C}_1-\mathtt{E}_1> + \\
&\frac{\beta}{2}\|\mathtt{C}_1-\mathtt{E}_1\|_{\mathtt{F}}^2 + <\mathtt{Y}_7,\mathtt{C}_2-\mathtt{E}_2> + \frac{\beta}{2}\|\mathtt{C}_2-\mathtt{E}_2\|_{\mathtt{F}}^2 + <\mathtt{Y}_8,\mathtt{S}^\sharp-\mathtt{J}> + \frac{\beta}{2}\|\mathtt{S}^\sharp-\mathtt{J}\|_{\mathtt{F}}^2,
\end{aligned}
\tag{9}
$$

where we define $\gamma_1 = \frac{1-\lambda_1}{2}$ and $\gamma_3 = \frac{1-\lambda_3}{2}$. $\mathtt{Y}_i, i = 1,\cdots,8$ are the Lagrange multipliers. β is the penalty parameter, where we use the same parameter for each augmented Lagrange term to simplify the derivation and parameter setting. The symbol $< .,. >$ represents the Frobenius inner product of two matrices, i.e, the trace of the product of two matrices. For example, given two matrices $\mathtt{A},\mathtt{B} \in \mathbb{R}^{\mathtt{m}\times\mathtt{n}}$, the Frobenius inner product is calculated as $< \mathtt{A},\mathtt{B} >=\mathrm{Tr}(\mathtt{A}^{\mathtt{T}}\mathtt{B})$.

The ADMM works by minimizing Eq. (9) with respect to one variable while fixing the others. During each iteration, we update each variable and the Lagrange multipliers in sequel. The detailed derivation for the solution is presented in the Appendix.

Solution for S: The closed form solution for $\mathtt{S}$ can be derived by taking derivative of (9) w.r.t to $\mathtt{S}$ and equating to zero.

$$\frac{1}{\beta}(\mathtt{R}^{\mathtt{T}}\mathtt{R}+\beta\mathtt{I})\mathtt{S}+\mathtt{S}(\mathtt{I}-\mathtt{C}_1)(\mathtt{I}-\mathtt{C}_1^{\mathtt{T}})=\frac{1}{\beta}\mathtt{R}^{\mathtt{T}}\mathtt{W}+(\mathtt{g}^{-1}(\mathtt{S}^{\sharp})+\frac{\mathtt{g}^{-1}(\mathtt{Y}_1)}{\beta}-\frac{\mathtt{Y}_2}{\beta}(\mathtt{I}-\mathtt{C}_1^{\mathtt{T}})). \tag{10}$$

Solution for S$^\sharp$: The closed form solution for $S^\sharp$ can be derived by taking derivative of (9) w.r.t $\mathtt{S}^\sharp$ and equating to zero.

$$\mathtt{S}^{\sharp}(2\mathtt{I}+(\mathtt{I}-\mathtt{C}_2)(\mathtt{I}-\mathtt{C}_2^{\mathtt{T}}))=(\mathtt{g}(\mathtt{S})-\frac{\mathtt{Y}_1}{\beta})+(\mathtt{J}-\frac{\mathtt{Y}_8}{\beta})-\frac{\mathtt{Y}_3}{\beta}(\mathtt{I}-\mathtt{C}_2^{\mathtt{T}}). \tag{11}$$

Solution for $\mathtt{C}_1$: The closed form solution for $\mathtt{C}_1$ can be derived as

$$(\mathtt{S}^{\mathtt{T}}\mathtt{S}+\mathtt{1}\mathtt{1}^{\mathtt{T}}+\mathtt{I})\mathtt{C}_1=\mathtt{S}^{\mathtt{T}}(\mathtt{S}+\frac{\mathtt{Y}_2}{\beta})+\mathtt{1}(\mathtt{1}^{\mathtt{T}}-\frac{\mathtt{Y}_4}{\beta})+(\mathtt{E}_1-\frac{\mathtt{Y}_6}{\beta}). \tag{12}$$

$$\mathtt{C}_1 := \mathtt{C}_1-\text{diag}(\mathtt{C}_1), \tag{13}$$

Solution for $\mathtt{C}_2$: The closed form solution for $\mathtt{C}_2$ can be derived as

$$((\mathtt{S}^{\sharp})^{\mathtt{T}}\mathtt{S}^{\sharp}+\mathtt{1}\mathtt{1}^{\mathtt{T}}+\mathtt{I})\mathtt{C}_2=(\mathtt{S}^{\sharp})^{T}(S^{\sharp}+\frac{\mathtt{Y}_3}{\beta})+\mathtt{1}(\mathtt{1}^{\mathtt{T}}-\frac{\mathtt{Y}_5}{\beta})+(\mathtt{E}_2-\frac{\mathtt{Y}_7}{\beta}). \tag{14}$$

$$\mathtt{C}_2 := \mathtt{C}_2-\text{diag}(\mathtt{C}_2), \tag{15}$$

Solution for $\mathtt{J}$: The optimization of $\mathtt{J}$ given all the remaining variables can be expressed as:

$$\begin{aligned}\mathtt{J} &= \underset{\mathtt{J}}{\text{argmin}}\lambda_2\|\mathtt{J}\|_{*}+<\mathtt{Y}_8,\mathtt{S}^{\sharp}-\mathtt{J}>+\frac{\beta}{2}\|\mathtt{S}^{\sharp}-\mathtt{J}\|_{\mathtt{F}}^2.\\ &= \underset{\mathtt{J}}{\text{argmin}}\lambda_2\|\mathtt{J}\|_{*}+\frac{\beta}{2}\|\mathtt{J}-(\mathtt{S}^{\sharp}+\frac{\mathtt{Y}_8}{\beta})\|_{\mathtt{F}}^2.\end{aligned} \tag{16}$$

A closed-form solution exists for this sub-problem. Let's define the soft-thresholding operation as $\mathcal{S}_{\tau}[x]=\text{sign}(x)\max(|x|-\tau,0)$, the optimal $\mathtt{J}$ can be obtained as:

$$\mathtt{J}=\mathtt{U}\mathcal{S}_{\frac{\lambda_2}{\beta}}(\Sigma)\mathtt{V}, \tag{17}$$

where $[\mathtt{U},\,\Sigma,\,\mathtt{V}]=\text{SVD}(\mathtt{S}^{\sharp}+\frac{\mathtt{Y}_8}{\beta})$.

Solution for $\mathtt{E}_1$: The closed-form solution for $\mathtt{E}_1$ can be obtained similarly:

$$\mathtt{E}_1=\mathcal{S}_{\frac{\lambda_1}{\gamma_1+\frac{\beta}{2}}}\left(\frac{\beta}{2\gamma_1+\beta}(\mathtt{C}_1+\frac{\mathtt{Y}_6}{\beta})\right). \tag{18}$$

Solution for $\mathtt{E}_2$ The derivation for the solution of $\mathtt{E}_2$ is similar to $\mathtt{E}_1$.

$$\mathtt{E}_2{}^{*}=\mathcal{S}_{\frac{\lambda_3}{\gamma_3+\frac{\beta}{2}}}\left(\frac{\beta}{2\gamma_3+\beta}(\mathtt{C}_2+\frac{\mathtt{Y}_7}{\beta})\right). \tag{19}$$

Algorithm 1 Multi-body non-rigid 3D reconstruction and segmentation using ADMM

```
Require:
  2D feature track matrix $W$, camera motion $R$, $\lambda_1$, $\lambda_2$, $\lambda_3$, $\rho > 1$, $\beta_m$, $\epsilon$;
Initialize: $S^{(0)}$, $S^{\sharp(0)}$, $C_1^{(0)}$, $E_1^{(0)}$, $C_2^{(0)}$, $E_2^{(0)}$, $\{\mathbf{Y}_i^{(0)}\}_{i=1}^{8} = \mathbf{0}$, $\beta^{(0)} = 1e^{-3}$;
  while not converged do
    1. Update $(S, S^{\sharp}, E_1, E_2, C_1, C_2)$ by Eq. (10), Eq. (11), Eq. (18), Eq. (19), Eq. (13)
       and Eq. (15); The new value for each variable is updated over iteration, which was
       initialized for the first iteration.
    2. Update $\{\mathbf{Y}_i\}_{i=1}^{8}$ and $\beta$ by Eq. (20)-Eq. (24);
    3. Check the convergence conditions $\|S^{\sharp} - g(S)\|_\infty \leq \epsilon$, $\|S - SC_1\|_\infty \leq \epsilon$, $\|S^{\sharp} - S^{\sharp}C_2\|_\infty \leq \epsilon$,
       $\|1^T C_1 - 1^T\|_\infty \leq \epsilon$, $\|1^T C_2 - 1^T\|_\infty \leq \epsilon$ and $\|C_1 - E_1\|_\infty \leq \epsilon$, $\|C_2 - E_2\|_\infty \leq \epsilon$;
       $\|S^{\sharp} - J\|_\infty \leq \epsilon$;
  end while
Ensure: $C_1$, $C_2$, $E_1$, $E_2$, $S, S^{\sharp}$.
  Form an affinity matrix $A_1 = |C_1| + |C_1^T|$, then apply spectral clustering [19] to $A_1$ to
  achieve non-rigid motion segmentation.
```

Detailed derivations to each sub-problems solution are provided in the supplementary material. Finally, the Lagrange multipliers $\{Y_i\}_{i=1}^{8}$ and β are updated as:

$$Y_1 = Y_1 + \beta(S^{\sharp} - g(S)), Y_2 = Y_2 + \beta(S - SC_1), \tag{20}$$

$$Y_3 = Y_3 + \beta(S^{\sharp} - S^{\sharp}C_2), Y_4 = Y_4 + \beta(1^T C_1 - 1^T) \tag{21}$$

$$Y_5 = Y_5 + \beta(1^T C_2 - 1^T), Y_6 = Y_6 + \beta(C_1 - E_1), \tag{22}$$

$$Y_7 = Y_7 + \beta(C_2 - E_2), Y_8 = Y_8 + \beta(S^{\sharp} - J). \tag{23}$$

$$\beta = \min(\beta_m, \beta\rho). \tag{24}$$

Initialization: Since the proposed problem is non-convex, proper initialization is required for fast convergence. In this work, we obtained rotation using [8] and initialized the S matrix as pinv(R)* W. β_0, β_m, ρ were kept as 10^{-3}, 10^{3}, and 1.1 respectively. The complete implementation is provided in Algorithm 1.

5 Experiments and Results

We performed extensive experiments on benchmark data-sets that are freely available. We tested our approach on both real data and synthetic data under sparse and semi-dense scenarios. Denote S^{est} as the estimated 3D structure and S^{GT} as the ground-truth structure, we use the following error metrics to evaluate the performance of the approach:

(i) Relative error in multi-body non-rigid 3D reconstruction

$$e_{3D} = \frac{1}{F}\sum_{f=1}^{F} \|S_f^{est} - S_f^{GT}\|_F / \|S_f^{GT}\|_F, \tag{25}$$

(ii) Error in multi-body non-rigid motion segmentation,

$$e_{MS} = \frac{\text{Total number of incorrectly segmented trajectories}}{\text{Total number of trajectories}}. \tag{26}$$

5.1 Experiment 1: Performance on sparse dataset

Since our approach simultaneously reconstructs and segments multi-body non-rigid motions. Thus, we conducted the first experiment to verify the advantage of our method compared with alternative two stage approaches. To this end, we devise the following experimental setup, namely first segmenting the 2D tracks and then reconstructing each body with single body non-rigid structure-from-motion algorithm and vice-versa. Specifically, the two baseline setups are:

1) Baseline method 1: Single body non-rigid structure-from-motion (State-of-the-art “block-matrix method” [8] was used) followed by subspace clustering of the 3D trajectories (SSC [11] was used), denoted as “BMM+SSC(3D)”.
2) Baseline method 2: Subspace clustering of the 2D feature tracks (2D trajectories) followed by single body non-rigid structure-from-motion for each cluster of 2D feature tracks, denoted as “SSC(2D)+BMM”.

In Table 1, we provide the experimental comparisons between our method and the two baseline methods in dealing with multi-body non-rigid structure-from-motion task.

Datasets	BMM+SSC(3D)		SSC(2D)+BMM		Our Method	
	e_{3D}	e_{MS}	e_{3D}	e_{MS}	e_{3D}	e_{MS}
Dance + Yoga	0.045	0.034	0.058	0.026	**0.045**	0.00
Drink + Walking	0.074	0.0	0.085	0.0	**0.073**	0.00
Shark + Stretch	0.024	0.401	0.098	0.394	**0.021**	0.00
Walking + Yoga	0.070	0.0	0.090	0.0	**0.066**	0.00
Face + Pickup	0.032	0.098	**0.023**	0.098	0.027	0.00
Face + Yoga	**0.017**	0.012	0.033	0.012	0.021	0.00
Shark + Yoga	0.035	0.416	0.105	0.409	**0.033**	0.00
Stretch + Yoga	0.039	0.0	0.055	0.0	**0.036**	0.00

Table 1. Performance comparison between our method and the two stage methods i.e first cluster and then reconstruct or vice-versa, where 3D reconstruction error (e_{3D}) and non-rigid motion segmentation error (e_{MS}) are used as error metrics. The statistics clearly shows the superior performance of our method in both 3D reconstruction and motion segmentation compared with the two stage methods.

Comments: In all of these sequences, our method achieves perfect motion segmentation and better non-rigid 3D reconstruction in most of the sequences compared with the two-stage approaches–statistical value for the same sequences can be inferred from Table 1. Furthermore, a visual comparison is presented in

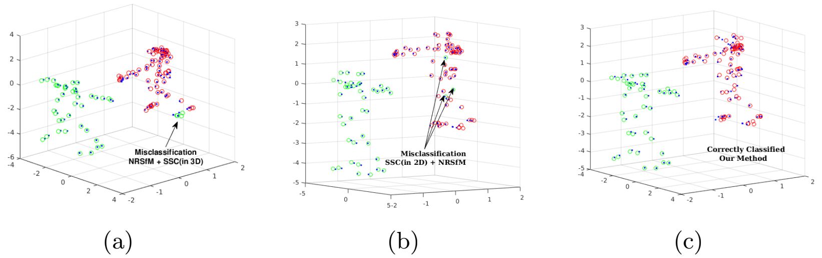


Fig. 4. An illustration of the efficacy of our approach. The plot shows the results on the "Dance + Yoga" sequence. (a) Result obtained by applying BMM method [8] to get 3D reconstruction and then using SSC [11] to segment 3D points. (b) Result obtained by applying SSC [11] to 2D feature tracks and then using BMM [8] to each cluster to get 3D reconstruction. (c) Result from our simultaneous reconstruction and segmentation framework. (Best viewed on screen in color)

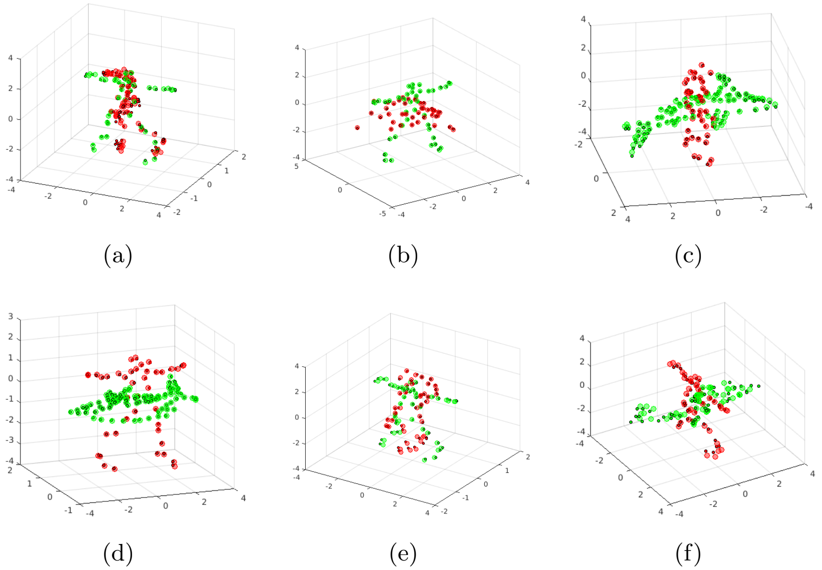
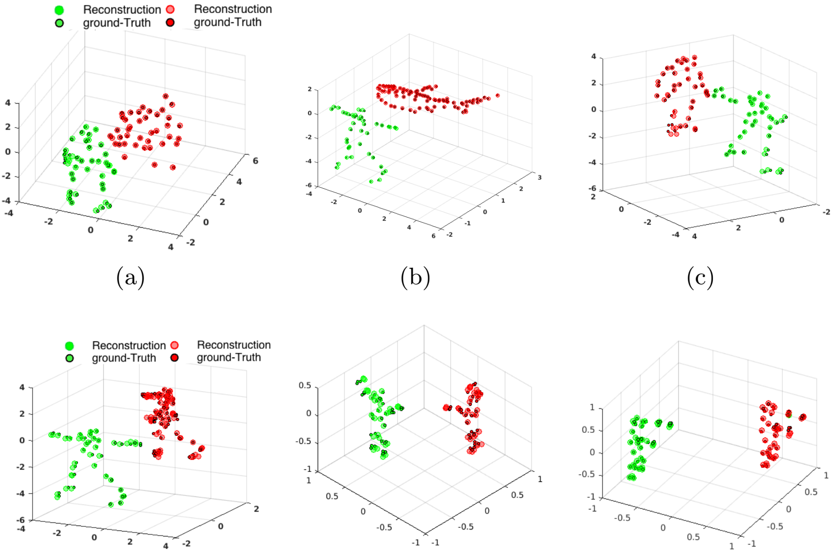


Fig. 5. 3D reconstruction and segmentation of different complex multi-body non-rigid motion sequences, where different objects intersect with each other. a) Dance-Yoga Sequence b) Face-Yoga Sequence c) Shark-Stretch Sequence d) Shark-Yoga Sequence e) Stretch-Yoga Sequence f) Walking-Yoga. Different colors indicate different clusters with dark small circles in the respective segments shows the ground-truth 3D points. (Best viewed in color)

Fig. 4, which illustrates that with the proposed framework we can procure correct features belonging to each object than the two-stage approaches.



(a) (b) (c)

(d) (e) (f)

Fig. 6. 3D reconstruction and segmentation of different multi-body non-rigid motion sequences a) Face-Pickup Sequence b) Shark-Yoga Sequence c) Stretch-Yoga Sequence d) Dance-Yoga Sequence e) p3_ball_1 f) p4_meet_12. The non-rigid motion sequences are generated from the CMU MoCap dataset [4], Torresani et al. [26] dataset and the UMPM dataset [1]. Different colors indicate different clusters with dark small circles in the respective segments shows the ground-truth 3D points. (Best viewed in color)

To further test the segmentation of different deforming objects performing different activities, we simulated two synthetic experimental settings. In the first setting, we combined non-rigid objects such that they are well separated in 3D space while in the next setting the objects are intersecting with each other in 3D space. We obtained perfect segmentation results for both settings. Fig. 5 and Fig. 6 show the qualitative segmentation and reconstruction results for the corresponding experiment. Quantitative performance comparison of segmentation with SSC [12] on synthetic sequence is presented in Table 1 .

Performance comparison of reconstruction error with state-of-the-art methods on synthetic dataset We compare the performance of our approach with other state-of-the-art non-rigid reconstruction methods on same data-set under similar settings. Synthetic data-set that are used for evaluating reconstruction error of multi-body non-rigid deformations are created by combining different objects from the CMU Mocap [4] and Torresani et al. dataset [26]. We compare our approach with state-of-the-art non-rigid methods such as BMM [8], PND [18], Zhu et al. [29] and Kumar et al. [17]. Statistical results are provided in Fig. 7, which clearly indicates the improvement of our method in 3D reconstruction in contrast of other approaches.

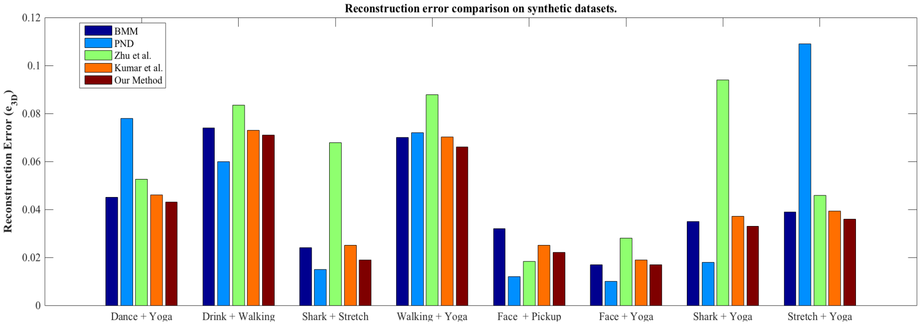


Fig. 7. Comparison of 3D reconstruction error with other competitive methods on synthetic datasets (CMU Mocap [4] and [25]). The comparison methods (BMM [8], PND [18], Zhu et al. [29], Kumar et al. [17]) present state-of-the-art approaches. Note: Code for Zhu et al. [29] work is not publicly available, the stats we provided here are taken from our own implementation. For exact numerical values, please refer to the supplementary material (Best viewed in color).

Comments: It can be observed from Fig. 7 that the reconstruction error obtained by our method in comparison to other state-of-the-art is either better or close to other competing approaches on all the datasets. We would like to mention that code for Zhu et al. [29] is not publicly available. Therefore, we used our own implementation of this algorithm for numerical comparison. MATLAB codes for other method such as BMM [8] and PND [18] are freely available.

5.2 Experiment 2: Performance on real image dataset UMPM [1].

UMPM : The Utrecht Multi-Person Motion (UMPM) dataset [1] is a benchmark dataset for multiple person interaction. It consists of synchronized videos with 644 × 484 resolution images. Each dataset consists of long-video sequence with multiple activities and different articulated motions. Although data are provided from four view point for each category, we only used one view point for evaluation. This dataset has been used in the past as a benchmark to evaluate multi-person motion capturing technique and many state-of-the-art techniques have used it to evaluate the performance of NRSfM methods [18], [10].

Performance comparison of 3D reconstruction error with state-of-the-art methods on real dataset UMPM [1] Following previous works over this topic, we also used the UMPM dataset for evaluation of our method in comparison to other competing methods. We evaluated our performance on five long video sequence, which are composed of complex non-rigid motion and extensive

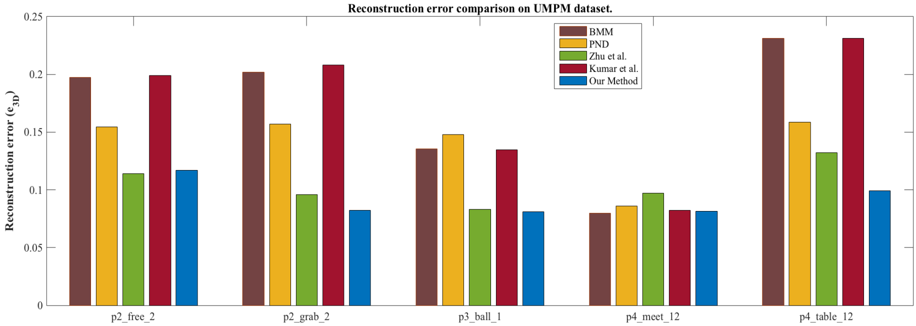


Fig. 8. Comparison of 3D reconstruction error with other competitive methods on real image data-set(UMPM [1]), which is composed of complex non-rigid deformation along with different activities over-time. The comparison methods (BMM [8], PND [18], Zhu et al. [29], Kumar et al. [17]) present state-of-the-art approaches. For exact numerical values, please refer to the supplementary material (Best viewed in color).

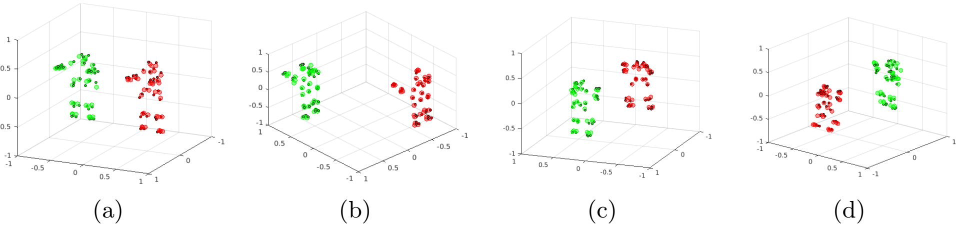

(a) (b) (c) (d)

Fig. 9. In (a), (b), (c), (d) larger and smaller circles shows the 3D reconstruction and ground-truth of p4_table_12, p4_meet_12, p2_grab_2, p2_free_2 data-set respectively. Different colors show the corresponding segmentation.(Best viewed in color)

variations of daily human actions with severe pose changes. Those sequences are p4_table_12, p4_meet_12, p2_grab_2, p2_free_2, and p3_ball_1.

Comments: The observations on real image experiments are very similar to the synthetic ones. In all the aforementioned data-sets, we obtained almost perfect segmentation along with reliable 3D reconstruction. Fig. 8 demonstrates the superior 3D reconstruction performance of our method in comparison to other methods. Furthermore, qualitative results obtained using our approach on the UMPM dataset can be inferred in Fig 9 and Fig. 10. Spatial and temporal affinity matrices obtained during the experiment on real sequence are analogous to synthetic sequence and therefore, similar inference can be drawn. The stats clearly indicate the superiority of our approach on 3D reconstruction, in addition it provides robust segmentation of multiple deformable objects.

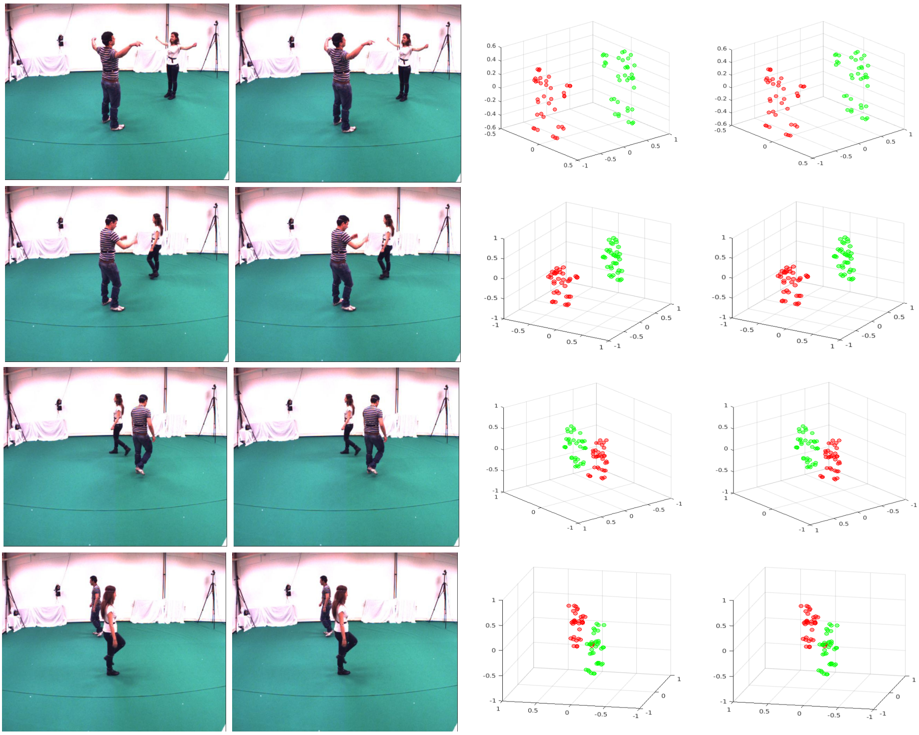


Fig. 10. 3D non-rigid reconstruction and segmentation results on p2_free_2 sequence of the UMPM dataset [2]. We obtained perfect segmentation and reliable 3D reconstruction over the entire video sequence which comprises of complex non-rigid deformation followed by different activities. (Best viewed in color)

5.3 Experiment 3: Performance on dense sequences

We also tested our method on freely available dense datasets [14]. Although our method is not scalable to millions of feature tracks, for completeness of our evaluation on bench-mark dataset that consists of human facial expressions, we tested our method on the uniformly sampled version of the original sequences. We performed experiments on benchmark NRSfM synthetic and real data-set sequence [14] introduced by Grag et al. This synthetic face sequence consists of four different datasets. Each sequence consists of different deformation and smooth camera rotations over time.

We sampled 3275 trajectories from each synthetic face sequence to verify the performance of our approach. 3D reconstruction errors obtained over these four face sequence are shown in Fig. 11. Furthermore, Fig. 12 caters the quality of reconstruction that is obtained using our method. In qualitative illustration (Fig. 12), the green dots show the reconstructed points whereas the red dots show the ground-truth 3D structure.

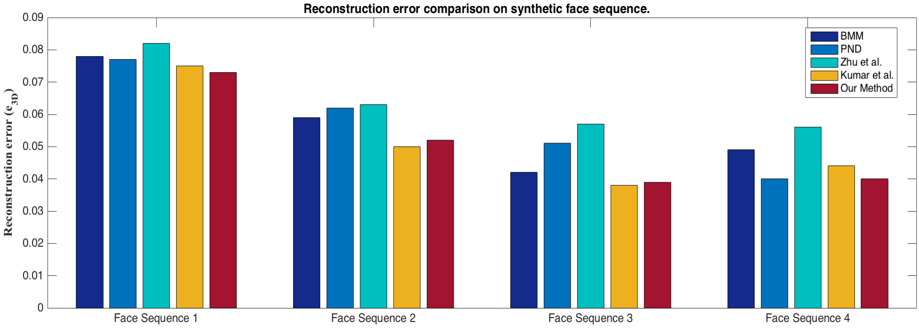


Fig. 11. Comparison of 3D reconstruction error with other competitive methods on synthetic dense face sequence ([14]) which is composed of non-rigid face deformation of different facial expression over-time. The comparison methods (BMM [8], PND [18], Zhu et al. [29], Kumar et al. [17]) represent the state-of-the-art approaches. This comparison is made over 3275 feature tracks which is taken by uniformly sampling the dense feature tracks. For exact numerical values, please refer to the supplementary material. (Best viewed in color).

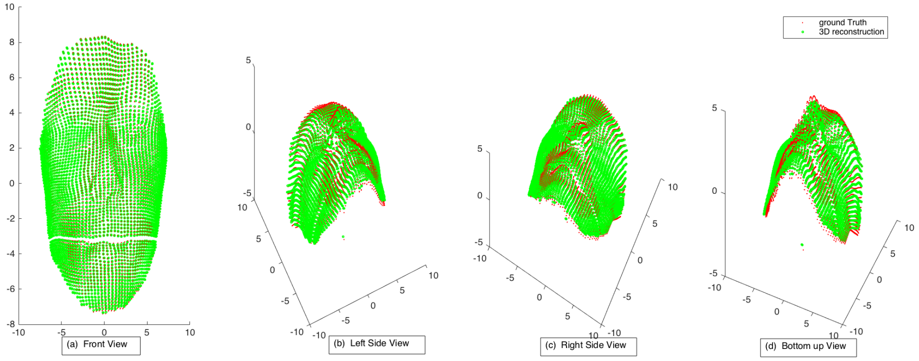
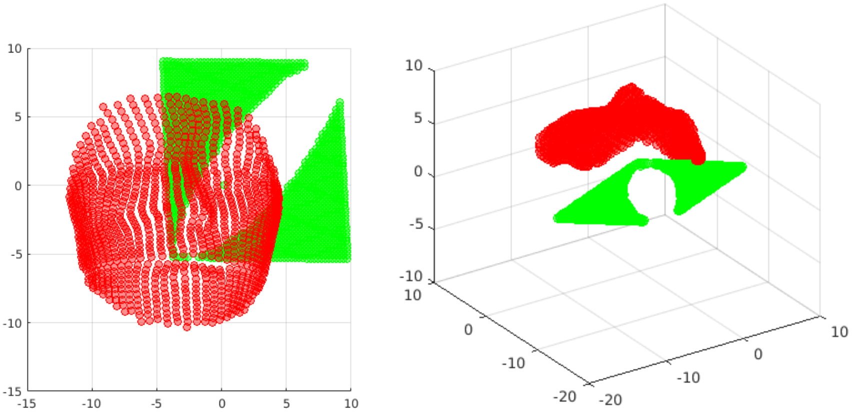


Fig. 12. Results on synthetic face sequence [14]. Red and green color show the ground-truth and reconstructed 3D structures respectively. (Best viewed in color)

Face with a background is very common in real world scenarios. To test segmentation and reconstruction in such cases, we combined synthetic face with an artificial background and projected it using an orthographic camera model. We provided projected the 2D feature tracks as input to our algorithm and obtained 3D shapes as shown in Fig. 13. Different colors represent distinct clusters that are recovered using our method.

Real face, back and heart sequence Garg et al. [14] provided three monocular videos composed of face, back and heart sequence respectively. These sequences capture the natural human deformation with considerable displacements from one frame to other. In the face sequence, the subject performs day-to-



(a) (b)

Fig. 13. (a), (b) show the front view and side view of the reconstruction and segmentation result obtained on "Face+Background" Sequence. This dataset was synthetically generated by combining synthetic face sequence [15] with background as mask. (Best viewed in color)

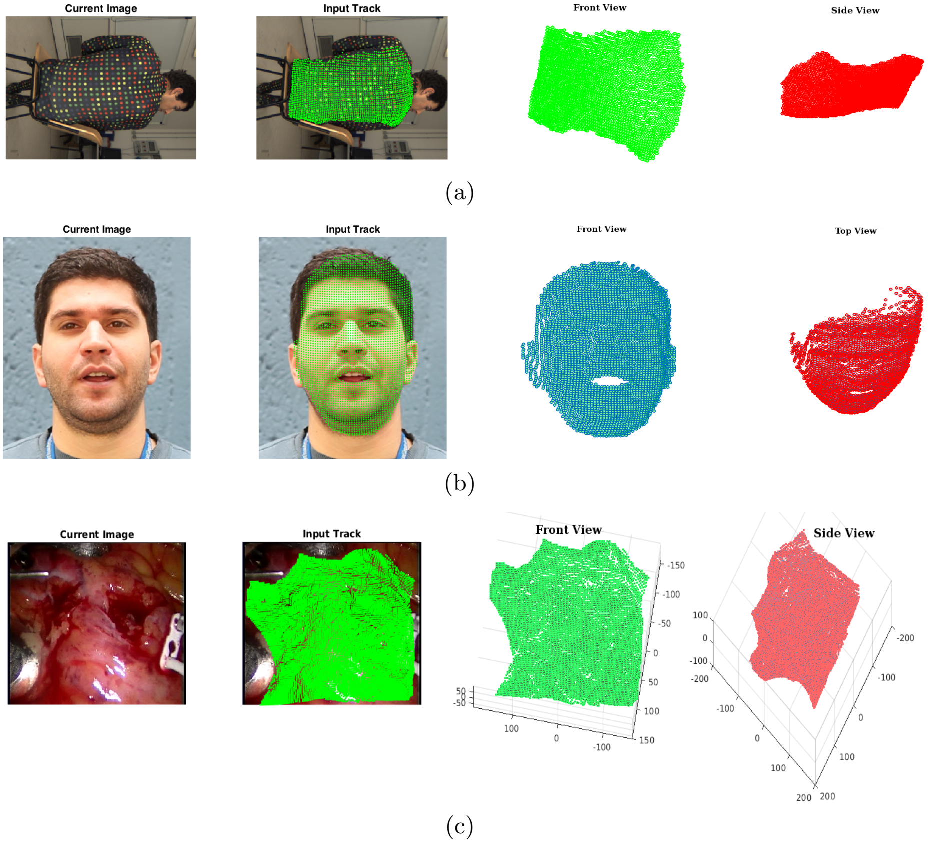
day facial expression whereas in the back sequence the person is stretching and shrinking his back wearing a textured t-shirt. Lastly, this dataset also provides a challenging monocular heartbeat sequence taken during bypass surgery. Quantitative evaluation over this dataset is not provided due to the absence of 3D ground-truth. However, qualitative results obtained are shown in Fig. 14(a), 14(b) and 14(c) respectively, which demonstrates the superior performance of our method in handling these real world challenging scenarios.

5.4 Experiment 4: Evaluation on more than two objects.

We also evaluated our method when three objects in the scene are performing complex motions over time. Adding shape clustering with trajectory clustering does not affect the segmentation, while improves reconstruction. A graphical illustration of such example and along with our obtained results in this case is shown in Fig. 15

5.5 Experiment 5: Convergence and analysis of the proposed optimization.

Since the proposed optimization is non-convex, we conducted experiments to study the convergence and timings of our approach. **Fig. 16** shows the a typical convergence curve of the proposed optimization on Shark+Yoga dataset. The optimization curve is provided only for better intuition of the algorithm. In our experiments similar convergence curves were obtained for other datasets as



(a)

(b)

(c)

Fig. 14. (a), (b), (c) shows the 3D reconstruction obtained on the Back, Face and Heart sequences respectively. Here, 2D trajectories are shown over the images to give more intuitive representation of the obtained structure. These results were obtained on uniformly sampled feature tracks. The number of feature points used for reconstruction of the Back, Face and Heart sequence are 2281, 3146 and 7546 respectively. (Best viewed in color)

well. In the figure different curves shows the primal residuals for each optimization terms over iteration. The current implementation takes around 5-7 minutes for thousand feature tracks to converge on commodity desktop with MATLAB R2015b on Ubuntu 14.04 and intel core i7 processor with 16GB RAM.

High values of λ_1 and λ_3 (say 0.6 or 0.7) during optimization may lead to higher segmentation error due to the highly sparse structure in $\mathtt{C}$ matrices. The benefit of elastic net is that it provides the flexibility of trade off between the sparsity and connectedness among different classes. Mathematically it means, with elastic net we have the freedom to adjust between ℓ_1 and ℓ_2 minimization of the same optimization variable, which is handy in controlling the sparsity of the matrix. Figure 18 shows the sparsity of $\mathtt{C}_1$ matrix with variation in λ_1 for

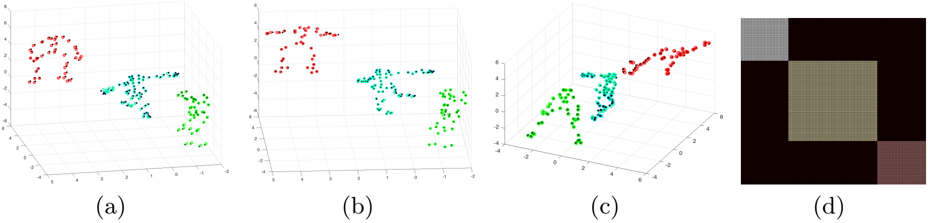


Fig. 15. (a)-(c) NRSfM with segmentation results for three objects on synthetic CMU MoCap dataset [4]. Our approach is able to reconstruct and segment each action such as stretch (red), dance (cyan) and yoga (green) faithfully with 3D reconstruction error of 0.0407. Here, different color corresponds to distinct deforming object, while dark and light color circles show ground-truth and reconstructed 3D coordinates respectively. (d) Affinity matrix obtained after spectral clustering [19]. (Best viewed in color)

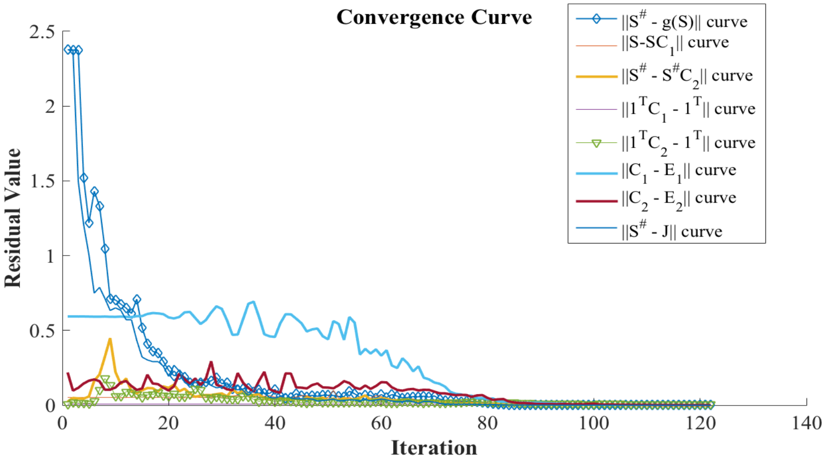


Fig. 16. Convergence curve of the proposed optimization. Each curve represents the residual value associated with each terms shown in legends over iteration. (Best viewed in color)

different sparse synthetic dataset where as Fig. 17(a) and 17(b) show the affinity matrix of $\mathtt{C}_1 \in \mathbb{R}^{\mathtt{P}\times\mathtt{P}}$ and $\mathtt{C}_2 \in \mathbb{R}^{\mathtt{F}\times\mathtt{F}}$ for the Dance with the Yoga sequence. The block-diagonal structure corresponding to both deforming objects is shown in Fig. 17(a). Clearly, the two objects span subspace that are independent of each other. In addition the obtained affinity matrix of $\mathtt{C}_1$ implies that the trajectories of each individual objects are self-expressive and thus each trajectory can be represented as a linear combination of all other trajectories. Similarly, Fig. 17(b) shows similar activity spans its own subspace and therefore, frames corresponding identical action can be clustered.

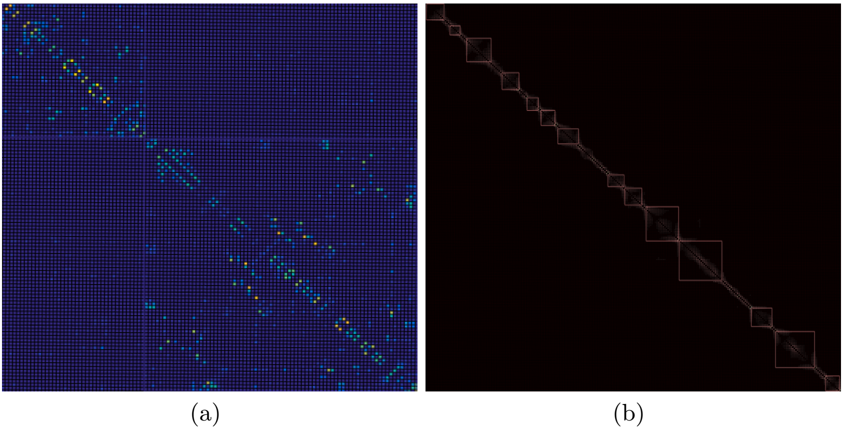


Fig. 17. (a) Affinity matrix obtained on the "Dance + Yoga" Sequence. Clearly, it shows two block diagonal structure, corresponding to the two objects, which is an interesting observation during our experiment. Thus, number of deforming objects can be directly inferred from the affinity matrix. (b) Affinity matrix obtained with temporal clustering, it shows similar activities are encapsulated in the same block structure or captured in local subspace. (Best viewed in color)

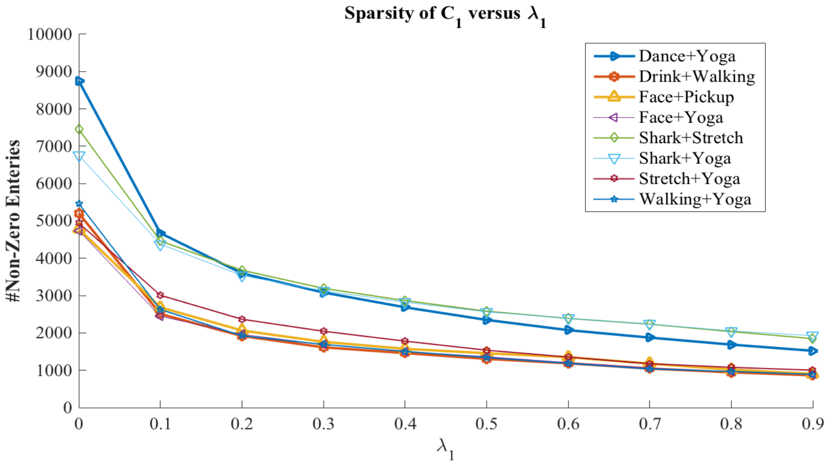


Fig. 18. Sparsity of $\mathtt{C}_1$ matrix vs λ_1 on different sparse data-set, it can be inferred that by using a proper value of λ_1 one can control the balance between sparsity and connectedness. Similar inference can be drawn for non-zero entries of $\mathtt{C}_2$ with variation in λ_3. (Best viewed in color)

6 Conclusion

In this paper, we proposed a novel framework to handle complex multi-body non-rigid structure from motion by exploiting spatio-temporal relation of deforming shapes, thus, providing a new way to compactly represent deformable shapes. Despite being a non-convex problem, we provided a solution to the resultant optimization using ADMM [5] which is effective, fast and easy to implement. Extensive experiments on both synthetic and real benchmark datasets demonstrate that the present approach outperforms state-of-the-art non-rigid reconstruction methods, by providing competitive 3D reconstruction and highly reliable segmentation. Even though methods such as [8], [21], [26], [17] can handle simple variations of non-rigid deformation well, our approach provides robust reconstruction for both simple and complex multi-body deformations. In future, we plan to investigate the scalability issue with the current implementation, thus extending the framework to deal with full resolution dense reconstruction tasks (hundreds of thousands of points).

7 Acknowledgment

This work was supported in part by Australian Research Council (ARC) grants (DE140100180, DP120103896, LP100100588, CE140100016), Australia ARC Centre of Excellence Program on Roboitic Vision, NICTA (Data61) and Natural Science Foundation of China (61420106007).

References

1. van der Aa, N., Luo, X., Giezeman, G., Tan, R., Veltkamp, R.: Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 1264–1269 (Nov 2011)
2. Van der Aa, N., Luo, X., Giezeman, G.J., Tan, R.T., Veltkamp, R.C.: Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 1264–1269. IEEE (2011)
3. Akhter, I., Sheikh, Y., Khan, S.: In defense of orthonormality constraints for non-rigid structure from motion. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 1534–1541 (2009)
4. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid structure from motion in trajectory space. In: Advances in Neural Information Processing Systems. pp. 41–48 (2008)
5. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3(1), 1–122 (2011)

6. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 690–696 (2000)
7. Cho, J., Lee, M., Oh, S.: Complex non-rigid 3d shape recovery using a procrustean normal distribution mixture model. International Journal of Computer Vision pp. 1–21 (2015)
8. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 2018–2025 (2012)
9. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. International Journal of Computer Vision 107(2), 101–122 (2014), http://dx.doi.org/10.1007/s11263-013-0684-2
10. Eldar, Y.C., Needell, D., Plan, Y.: Uniqueness conditions for low-rank matrix recovery. Applied and Computational Harmonic Analysis 33(2), 309–314 (2012)
11. Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11), 2765–2781 (2013)
12. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 2790–2797. IEEE (2009)
13. Fitzgibbon, A.W., Zisserman, A.: Multibody structure and motion: 3-d reconstruction of independently moving objects. In: Computer Vision-ECCV 2000, pp. 891–906. Springer (2000)
14. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 1272–1279 (2013)
15. Garg, R., Roussos, A., Agapito, L.: A variational approach to video registration with subspace constraints. International Journal of Computer Vision 104(3), 286–314 (2013)
16. Gotardo, P., Martinez, A.: Non-rigid structure from motion with complementary rank-3 spaces. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 3065–3072 (2011)
17. Kumar, S., Dai, Y., Li, H.: Multi-body non-rigid structure-from-motion. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 148–156. IEEE (2016)
18. Lee, M., Cho, J., Choi, C.H., Oh, S.: Procrustean normal distribution for non-rigid structure from motion. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 1280–1287 (2013)
19. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14, pp. 849–856. MIT Press (2002)
20. Ozden, K.E., Schindler, K., Van Gool, L.: Multibody structure-from-motion in practice. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(6), 1134–1141 (2010)
21. Paladini, M., Del Bue, A., Stosic, M., Dodig, M., Xavier, J., Agapito, L.: Factorization for non-rigid and articulated structure using metric projections. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 2898–2905 (2009)
22. Russell, C., Yu, R., Agapito, L.: Video pop-up: Monocular 3d reconstruction of dynamic scenes. In: European Conference on Computer Vision, pp. 583–598 (2014)
23. Simon, T., Valmadre, J., Matthews, I., Sheikh, Y.: Separable spatiotemporal priors for convex reconstruction of time-varying 3d point clouds. In: European Conference on Computer Vision, pp. 204–219 (2014)

24. Torresani, L., Hertzmann, A.: Automatic non-rigid 3D modeling from video. In: Proc. European Conf. Computer Vision. pp. 299–312 (2004)
25. Torresani, L., Hertzmann, A., Bregler, C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE Trans. Pattern Anal. Mach. Intell. 30(5), 878–892 (2008)
26. Torresani, L., Yang, D.B., Alexander, E.J., Bregler, C.: Tracking and modeling non-rigid objects with rank constraints. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 493–500 (2001)
27. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. Int'l J. Computer Vision 67(2), 233–246 (2006)
28. You, C., Li, C.G., Robinson, D.P., Vidal, R.: Oracle based active set algorithm for scalable elastic net subspace clustering. arXiv preprint arXiv:1605.02633 (2016)
29. Zhu, Y., Huang, D., De La Torre, F., Lucey, S.: Complex non-rigid motion 3d reconstruction by union of subspaces. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 1542–1549 (June 2014)
30. Zhu, Y., Lucey, S.: Convolutional sparse coding for trajectory reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3), 529–540 (March 2015)
31. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320 (2005)