DUST: Dual Union of Spatio-Temporal Subspaces for Monocular Multiple Object 3D Reconstruction

Antonio Agudo Francesc Moreno-Noguer
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), 08028, Barcelona, Spain

Abstract

We present an approach to reconstruct the 3D shape of multiple deforming objects from incomplete 2D trajectories acquired by a single camera. Additionally, we simultaneously provide spatial segmentation (i.e., we identify each of the objects in every frame) and temporal clustering (i.e., we split the sequence into primitive actions). This advances existing work, which only tackled the problem for one single object and non-occluded tracks. In order to handle several objects at a time from partial observations, we model point trajectories as a union of spatial and temporal subspaces, and optimize the parameters of both modalities, the non-observed point tracks and the 3D shape via augmented Lagrange multipliers. The algorithm is fully unsupervised and results in a formulation which does not need initialization. We thoroughly validate the method on challenging scenarios with several human subjects performing different activities which involve complex motions and close interaction. We show our approach achieves state-of-the-art 3D reconstruction results, while it also provides spatial and temporal segmentation.

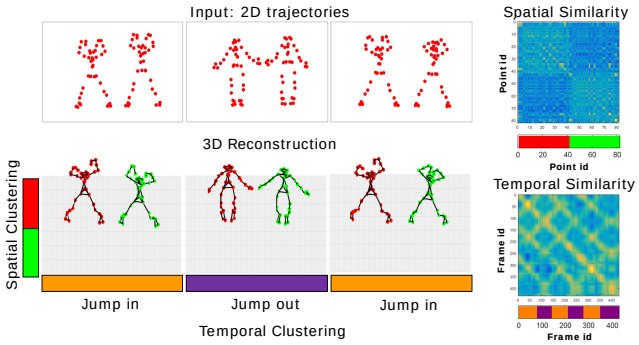

Figure 1. **Simultaneous 3D non-rigid reconstruction, spatial segmentation and temporal clustering from incomplete 2D point tracks. Top-left:** Example of input 2D point tracks (from the CMU MoCap dataset). For clarity we show a complete non-overlapped case, but our approach can handle discontinuous tracks and high degree of object overlapping. **Right:** Spatial and Temporal similarity matrices we retrieve. Each entry expresses the pairwise affinity between points or frames. Clusters are directly discovered by applying spectral clustering on these matrices. **Bottom-left:** 3D shape reconstruction together with the temporal and spatial clustering results. Spatial segmentation yields two objects, represented by red and green points. Temporal clusters clearly identify two motion primitives corresponding to the 'jump in' (orange) and 'jump out' (magenta) sub-actions.

1. Introduction

The problem of Non-Rigid Structure from Motion (NRSfM) involves simultaneously recovering deformable 3D shape and camera motion from monocular 2D point tracks. Since many different shape configurations may yield similar projections, NRSfM turns to be a highly ambiguous problem, which requires introducing prior information in order to be solved. Standard priors include the use of low-rank subspaces constraining the solution space of either the entire shape [2, 24, 37], the 3D point trajectories [6, 32] or the force patterns that induce the deformations [3].

All these previous approaches, though, use one single low-rank modality at a time. This prevents them from being applicable in situations that require models with high levels of expressiveness, such as for complex point trajectories or when dealing with multiple objects performing different types of deformations and motions. In this paper we tackle both these situations.

There exist previous works partially addressing these scenarios. For the rigid case, the shape of multiple moving objects can be retrieved by first segmenting the objects from the input 2D tracks and then applying a rigid SfM algorithm to each of them [30, 33, 39]. However, this strategy depends on the accuracy of the initial segmentation which, for the case of non-rigid and overlapping objects is likely to fail. Regarding the non-rigid case, there has been a recent attempt at reconstructing complex dynamics by modeling motion as a union of temporal subspaces [40]. This approach, however, has been applied to one single object, and relies on continuous and fully observed 2D point tracks.

In order to reconstruct multiple non-rigid objects with complex motions from partial 2D observations, we introduce a novel optimization framework that combines spatial

and temporal clustering in a unified manner. The two types of clustering are performed via affinity matrices, which are jointly learned, in conjunction with the 3D shape, using an Augmented Lagrange Multiplier (ALM) scheme. The approach is fully unsupervised and requires no initialization. We extensively evaluate the method on sequences with up to four subjects performing complex actions and interacting with each other. As shown in Fig. 1 the outcome of our algorithm is the spatial segmentation of each frame, which is likely to correspond to each of the subjects, a temporal clustering corresponding to motion primitives ('jump in' and 'jump out' for the example shown in the figure), plus the 3D reconstruction of each individual. We are not aware of any other approach solving the three problems simultaneously. Furthermore, as we will show in the results, the accuracy of the 3D reconstructions we obtain, improves that of state-of-the-art NRSfM methods by a considerable margin.

2. Related Work

We next review the most related work dealing with single- and multi-object reconstruction.

NRSfM for Single Object Reconstruction. The most standard approach to address the inherent ambiguity of the NRSfM problem is by assuming the underlying 3D shape is low rank. In order to estimate such low rank model, factorization-based approaches have been typically used [4, 10, 19, 31, 37]. Alternatively, other approaches impose the low-rank constraint by means of robust PCA-like formulations in which the rank of a matrix representing the shape is minimized. These type of methods either assume the data lies on a single low dimensional space [16, 18, 20] or in a union of temporal subspaces [40]. On top of these shape models, additional spatial [24] or temporal [1, 8, 26] smoothness constraints have also been considered. Low-rank models have been extended to the temporal domain, by fitting point trajectories to a series of predefined basis [6, 32, 38], to shape-and-temporal composite domains [21, 22, 35], and to the space of forces that induce the deformations [3].

All previous approaches, though, have been focused on retrieving the shape of single objects. Most of them, indeed, are not directly applicable to the multi-object scenario we contemplate in this paper, because they rely on a single linear subspace assumption that is not rich enough to model the variability occurring on scenarios with multiple objects performing different actions. Trajectory-based methods [6, 32, 38], can potentially tackle this type of scenarios because the low-rank is applied per point on the temporal domain. However, as we will show in the results section, a high sensitivity on to the dimension of the low-rank penalizes the accuracy of the reconstructions they provide. Furthermore, none of the previous methods is intended to provide full temporal and spatial segmentation of the sequence. Table 1 provides a qualitative comparison, in terms of available characteristics, of our approach and the most relevant NRSfM methods.

	[3, 37]	[21, 22]	[16, 20]	[18, 24, 25]	[40]	Ours
Rank required	–	–	✓	✓	✓	✓
Occlusion handling	✓	✓	–	✓	–	✓
Multiple objects	–	✓	–	–	✓	✓
Temporal clustering	–	–	–	–	✓	✓
Shape clustering	–	–	–	–	–	✓

Table 1. **Qualitative comparison of our approach with other NRSfM methods.** Our approach is the only one that simultaneously provides 3D reconstruction, shape segmentation and temporal clustering. Important characteristics are that it can also naturally handle complex scenarios with multiple interacting objects and incomplete 2D input tracks; and it does not need to adjust the rank of the basis. When this is required, it usually turns to be a very sensitive parameter for accuracy of the method. Note that [18] performs shape clustering directly from 2D, as an independent and separate task previous to the shape reconstruction.

Multi-Object Reconstruction. Most existing works in multi-object reconstruction from point tracks are applied to rigid objects, and follow a two-step pipeline. First the 2D motion tracks are segmented into several objects using a subspace clustering approach [17, 27]; and then rigid SfM techniques [36] are separately applied to each of the objects [15, 33, 39]. The technique in [30] is able to perform simultaneous segmentation and reconstruction [30], but it still is only applicable to rigid cases. One interesting exception is the recent work [34] which assumes the object to be represented as overlapping rigid parts, and simultaneously segments and reconstructs these parts using piecewise rigid models. However, while this approach provides dense (spatial) segmentation and reconstruction, suffers from the relative low expressiveness of the piecewise models, which limits the applicability to scenes with mild deformations.

Our Contributions. We depart from previous work in that our solution simultaneously provides 3D non-rigid reconstruction, spatial segmentation and temporal clustering. To the best of our knowledge, no previous approach has jointly addressed the three problems. Additionally, we can tackle sequences with complex motions and point track patterns with a high degree of overlapping, in a completely unsupervised manner. This outcome is the result of our technical contribution: a novel formulation of the problem that accounts for both temporal and spatial consistency of the point tracks while minimizing the rank of the solution. In the following we shall denote our approach as 'Dual Union of Spatio-Temporal subspaces' (DUST).

3. Revisiting NRSfM

We next revisit two NRSfM formulations that will be used later to model non-rigid shape as a union of spatial and temporal subspaces.

Let us consider a dynamic set of N 3D points observed along F frames. We denote by $\mathbf{x}_i^f = [x_i^f, y_i^f, z_i^f]^\top$ the 3D coordinates of the i-th point at frame f, and by $\mathbf{p}_i^f = [u_i^f, v_i^f]^\top$ the 2D orthographic projection of the same point in the image plane. To simplify subsequent formulation, the camera translation $\mathbf{t}^f = \sum_i \mathbf{p}_i^f / N$ is subtracted from the 2D projections, *i.e.*, we consider $\tilde{\mathbf{p}}_i^f = \mathbf{p}_i^f - \mathbf{t}^f$.

We can then build the following linear system mapping 3D-to-2D point coordinates:

$$\underbrace{\begin{bmatrix} \tilde{\mathbf{p}}_1^1 & \dots & \tilde{\mathbf{p}}_N^1 \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{p}}_1^F & \dots & \tilde{\mathbf{p}}_N^F \end{bmatrix}}_{\mathbf{P}} = \underbrace{\begin{bmatrix} \mathbf{R}^1 & & \\ & \ddots & \\ & & \mathbf{R}^F \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} \mathbf{x}_1^1 & \dots & \mathbf{x}_N^1 \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^F & \dots & \mathbf{x}_N^F \end{bmatrix}}_{\hat{\mathbf{X}}}$$

where $\mathbf{P}$ is a $2F \times N$ matrix storing the 2D point tracks arranged in columns, $\mathbf{G}$ is a $2F \times 3F$ block diagonal matrix, made of the F truncated 2×3 camera rotations $\mathbf{R}^f$, and $\hat{\mathbf{X}}$ is a $3F \times N$ matrix with the 3D positions of the points along the sequence, also arranged in columns. The NRSfM problem then entails retrieving the time-varying shape $\hat{\mathbf{X}}$ and camera motion $\mathbf{G}$ from 2D point tracks $\mathbf{P}$.

Early solutions based on the factorization method [10], constrained the matrix $\hat{\mathbf{X}}$ to be low-rank. For a given rank K it was shown that rank$(\hat{\mathbf{X}})$ was $\leq 3K$. Shape could then be estimated applying a rank $3K$ factorization over $\mathbf{P}$ followed by constraints ensuring rotation orthonormality [5]. However, despite their popularity, these methods are very sensitive to the value of the rank, which needs to be carefully chosen to obtain accurate results.

More recently, several approaches have enforced the low rank of the time-varying shape by applying nuclear norm minimization directly over the matrix encoding the 3D point positions [16, 18, 20]. Following [16, 23], we re-arrange the elements of $\hat{\mathbf{X}}$ into a new $3N \times F$ matrix $\mathbf{X}$ encoding the x, y and z coordinates in different rows:

$$\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_N^1 & y_1^1 & \dots & y_N^1 & z_1^1 & \dots & z_N^1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_1^F & \dots & x_N^F & y_1^F & \dots & y_N^F & z_1^F & \dots & z_N^F \end{bmatrix}^\top .$$

The interest of this matrix is that under a low-rank representation with K shape bases, it retains the rank K (in contrast to $\hat{\mathbf{X}}$, which was $3K$). Therefore, $\mathbf{X}$ naturally captures the fact that it is represented by a K-order linear model and avoids spurious degrees of freedom while allowing to learn redundancies between frames.

We shall use both $\mathbf{X}$ and $\hat{\mathbf{X}}$. In order to map one matrix to the other we define a function q such that $\hat{\mathbf{X}} = q(\mathbf{X}) = (\mathbf{I}_3 \otimes \mathbf{X}^\top)\mathbf{A}$, where $\mathbf{A}$ is a $9N \times N$ binary matrix, $\otimes$ is the Kronecker product operator and $\mathbf{I}_3$ the identity matrix. Similarly, we define the inverse mapping $\mathbf{X} = q^{-1}(\hat{\mathbf{X}}) = (\hat{\mathbf{X}}^\top \otimes \mathbf{I}_3)\mathbf{F}$, where $\mathbf{F}$ is a $9F \times F$ binary matrix.

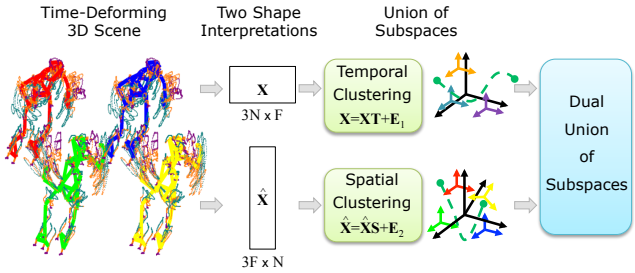

Figure 2. **Dual Union of Spatio-Temporal Subspaces.** Let us consider a scenario with four objects (initially unknown) that are non-rigidly moving and interacting. This 4D information can be encoded using two different representations given by matrices $\mathbf{X}$ and $\hat{\mathbf{X}}$ (see text). Post-multiplying these matrices by affinities $\mathbf{T}$ and $\mathbf{S}$, respectively, allows to simultaneously perform temporal and spatial clustering. Additionally, all these matrices are enforced to be low rank. This means that each temporal and spatial cluster (indicated by color vectors in the figure), is in turn represented by a union of subspaces (indicated by black vectors in the figure). Our Dual Union of Spatio-Temporal Subspaces model, combines the two typologies of subspaces.

4. Dual Union of Spatio-Temporal Subspaces

As we have just seen, the time-varying scene can be either represented by the matrices $\mathbf{X}$ or $\hat{\mathbf{X}}$. Even though the two matrices are made by exactly the same elements, each of it will be used for a different purpose. From one side, following [29, 40], we first define a temporal clustering over the shapes through a temporal affinity $F \times F$ matrix $\mathbf{T}$:

$$\mathbf{X} = \mathbf{X}\mathbf{T} + \mathbf{E}_t, \tag{1}$$

where $\mathbf{E}_t$ is a residual noise. The affinity matrix $\mathbf{T}$ measures the similarity between frames. As we shall see later, once this matrix is learned from data, spectral clustering algorithms [14] can be applied on it to discover and match different motion primitives within the sequence.

We additionally consider performing spatial segmentation by means of a so-called spatial affinity $N \times N$ matrix $\mathbf{S}$, which in this case, is applied on the matrix $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}\mathbf{S} + \mathbf{E}_s, \tag{2}$$

where $\mathbf{E}_s$ is a residual noise. The affinity matrix now encodes point similarity, and again, once it its learned, we can use spectral clustering on it to spatially segment the data and split it into different objects.

Equations (1) and (2) can be interpreted as a representation of the time-varying 3D points using a union of temporal and spatial subspaces, respectively. In the following section we will jointly apply the two types of representations, *i.e.*, we will merge two unions of subspaces, and hence the name of 'Dual' Union of Spatio-Temporal (DUST) subspaces we give to our approach. This idea is illustrated in Fig. 2.

5. 3D Reconstruction and Spatio-Temporal Clustering from 2D Trajectories

In this Section we combine the standard NRSfM projection equation described in Sect. 3, with Eqs. (1) and (2) enforcing temporal and spatial clustering, in order to simultaneously segment the 2D trajectories into different objects, provide their 3D reconstruction, and cluster their motion into a series of primitives. Note that [40] already presented an approach perform reconstruction and temporal grouping of one single object. Here we introduce the multi-object capability, and the possibility to handle occluded 2D tracks. As it will be shown shortly, this involves having to deal with a considerably more complex loss function and a more elaborate optimization strategy than that considered in [40].

5.1. Problem Formulation

Let $\bar{\mathbf{P}}$ be a possibly incomplete 2D measurement matrix, and $\mathbf{O}$ its corresponding $F \times N$ observation matrix with $\{1, 0\}$ entries indicating whether the two coordinates of a point in a specific frame are observed or not.

We can specifically formulate our problem as follows: given the partial 2D tracks $\bar{\mathbf{P}}$ and the observation matrix $\mathbf{O}$, we seek to estimate the temporal 3D location of all points $\hat{\mathbf{X}}$, the affinity matrices associated to the temporal $\mathbf{T}$ and spatial $\mathbf{S}$ clustering, the matrix $\mathbf{P}$ of complete 2D tracks and the matrix $\mathbf{G}$ of camera rotations. Let us denote by $\mathbf{\Theta} \equiv \{\mathbf{P}, \mathbf{G}, \mathbf{T}, \mathbf{S}, \mathbf{X}, \mathbf{E}_t, \mathbf{E}_s\}$ the set of all model parameters.

For estimating these parameters we propose a cost function that incorporates the spatio-temporal model described previously and enforces the matrices to lie in low-rank subspaces. As standard practice [16, 20], the nuclear norm is used as a convex approximation to the rank minimization. Our problem can therefore be written as follows:

$$\begin{aligned}\underset{\mathbf{\Theta}}{\arg\min}\ & \| (\mathbf{O} \otimes \mathbf{1}_2) \odot \left(\mathbf{P} - \bar{\mathbf{P}}\right) \|_F^2 + \beta\|\mathbf{P}\|_* + \phi\|\mathbf{T}\|_* \\ & + \phi\|\mathbf{S}\|_* + \gamma\|\mathbf{X}\|_* + \lambda_t\|\mathbf{E}_t\|_1 + \lambda_s\|\mathbf{E}_s\|_1 \quad (3)\end{aligned}$$

$$\begin{aligned}\text{subject to}\quad & \mathbf{P} = \mathbf{G}\hat{\mathbf{X}} \\ & \mathbf{I}_{2F} = \mathbf{G}\mathbf{G}^\top \\ & \mathbf{X} = \mathbf{X}\mathbf{T} + \mathbf{E}_t \\ & \hat{\mathbf{X}} = \hat{\mathbf{X}}\mathbf{S} + \mathbf{E}_s\end{aligned}$$

where $\odot$ represents the Hadamard product and $\mathbf{1}$ is a vector of ones. $\|\cdot\|_*$ is the nuclear norm, $\|\cdot\|_1$ is the convex approximation to sparse error and $\|\cdot\|_F$ indicates the Frobenius norm. $\{\beta, \phi, \gamma, \lambda_t, \lambda_s\}$ are penalty term parameters

We devise an approximated three-step strategy to minimize this cost function. First, we complete the partially observed measurement matrix $\mathbf{P}$. Then, we estimate the camera rotations matrix $\mathbf{G}$. And finally, we simultaneously solve for the shape $\mathbf{X}$ and clustering parameters $\mathbf{T}$ and $\mathbf{S}$. We next describe each of these steps.

5.2. Completing Missing Entries

To complete the unobserved tracks identified as zeros within the observation matrix $\mathbf{O}$, we separately optimize $\mathbf{P}$ taking the first two terms of Eq. (3):

$$\min_{\mathbf{P}} \quad \| (\mathbf{O} \otimes \mathbf{1}_2) \odot \left(\mathbf{P} - \bar{\mathbf{P}}\right) \|_F^2 + \beta\|\mathbf{P}\|_* \quad . \quad (4)$$

As it was shown in [7, 11, 12], this type of low-rank minimizations with the nuclear norm acting as a regularizer can be optimized with a bilinear factorization $\mathbf{P} = \mathbf{U}\mathbf{V}^\top$ and applying ALM [9]. By doing this, we obtain the following augmented Lagrangian function:

$$\begin{aligned}\underset{\mathbf{P},\mathbf{U},\mathbf{V}}{\arg\min}\ & \| (\mathbf{O} \otimes \mathbf{1}_2) \odot \left(\mathbf{P} - \bar{\mathbf{P}}\right) \|_F^2 + \frac{\beta}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \\ & + \langle \mathbf{L}, \mathbf{P} - \mathbf{U}\mathbf{V}^\top \rangle + \frac{\alpha}{2}\|\mathbf{P} - \mathbf{U}\mathbf{V}^\top\|_F^2 \ , \quad (5)\end{aligned}$$

where $\mathbf{L}$ is the $2F \times N$ Lagrange multiplier and $\alpha > 0$ a penalty parameter. We solve this optimization following the algorithm described in [12] (Alg. #1 in this paper).

5.3. Estimating Camera Rotation

Given the full matrix of point tracks $\mathbf{P}$, the camera rotation matrices $\mathbf{R}$, *i.e.*, the matrix $\mathbf{G}$, can be estimated independently from the rest of model parameters by using the projection and the orthonormality constraints. There are several alternatives and approximations for doing so, *e.g.*, strategies that enforce smooth trajectories [21, 22], methods based on trace-norm minimization that assume the rank of the subspace a priori [16, 35] or techniques based on Procrustes analysis [24]. Of course, for easier scenarios, $\mathbf{G}$ could also be recovered using a few background rigid points and then applying rigid factorization [32].

In any event, since the focus of the paper is on the accuracy of the 3D reconstruction and the ability to perform temporal and spatial segmentation, we will assume the matrix $\mathbf{G}$ has been computed by any of these previous methods. Furthermore, when comparing with state-of-the-art, we will use the same rotation matrices for all methods.

5.4. Joint Clustering and 3D Reconstruction

In order to jointly recover 3D shape and the spatio-temporal clustering, we again resort to the ALM method. Assuming $\mathbf{P}$ and $\mathbf{G}$ are already known, the minimization we need to perform is:

$$\underset{\mathbf{T},\mathbf{S},\mathbf{X},\mathbf{E}_t,\mathbf{E}_s}{\arg\min}\ \phi\|\mathbf{T}\|_* + \phi\|\mathbf{S}\|_* + \gamma\|\mathbf{X}\|_* + \lambda_t\|\mathbf{E}_t\|_1 + \lambda_s\|\mathbf{E}_s\|_1$$

$$\begin{aligned}\text{subject to}\quad & \mathbf{P} = \mathbf{G}\hat{\mathbf{X}} \\ & \mathbf{X} = \mathbf{X}\mathbf{T} + \mathbf{E}_t \\ & \hat{\mathbf{X}} = \hat{\mathbf{X}}\mathbf{S} + \mathbf{E}_s\end{aligned}$$

Since the parameters $\{\phi, \gamma, \lambda_t, \lambda_s\}$ can be scaled w.r.t. one of them, in the following, without loss of generality, we fix $\phi = 1$. The Lagrangian function in this is:

$$\begin{aligned}\underset{\boldsymbol{\Theta}_{s,t,x}}{\arg\min} \quad & \|\mathbf{J}\|_* + \|\mathbf{K}\|_* + \gamma\|\mathbf{X}\|_* + \lambda_t\|\mathbf{E}_t\|_1 + \lambda_s\|\mathbf{E}_s\|_1 \\ & + \langle \mathbf{L}_1, \mathbf{X} - \mathbf{X}\mathbf{T} - \mathbf{E}_t \rangle + \frac{\alpha}{2}\|\mathbf{X} - \mathbf{X}\mathbf{T} - \mathbf{E}_t\|_F^2 \\ & + \langle \mathbf{L}_2, \mathbf{P} - \mathbf{G}\mathbf{D} \rangle + \frac{\alpha}{2}\|\mathbf{P} - \mathbf{G}\mathbf{D}\|_F^2 \\ & + \langle \mathbf{L}_3, \mathbf{D} - \mathbf{D}\mathbf{S} - \mathbf{E}_s \rangle + \frac{\alpha}{2}\|\mathbf{D} - \mathbf{D}\mathbf{S} - \mathbf{E}_s\|_F^2 \\ & + \langle \mathbf{L}_4, q(\mathbf{X}) - \mathbf{D} \rangle + \frac{\alpha}{2}\|q(\mathbf{X}) - \mathbf{D}\|_F^2 \\ & + \langle \mathbf{L}_5, \mathbf{T} - \mathbf{J} \rangle + \frac{\alpha}{2}\|\mathbf{T} - \mathbf{J}\|_F^2 \\ & + \langle \mathbf{L}_6, \mathbf{S} - \mathbf{K} \rangle + \frac{\alpha}{2}\|\mathbf{S} - \mathbf{K}\|_F^2 \end{aligned} \tag{6}$$

where $\boldsymbol{\Theta}_{s,t,x} \equiv \{\mathbf{J}, \mathbf{T}, \mathbf{K}, \mathbf{S}, \mathbf{X}, \mathbf{D}, \mathbf{E}_s, \mathbf{E}_t\}$ are the spatio-temporal clustering and shape parameters, including three support matrices we have introduced corresponding to $\mathbf{D} \equiv q(\mathbf{X})$, $\mathbf{J} \equiv \mathbf{T}$ and $\mathbf{K} \equiv \mathbf{S}$. Additionally, $\mathbf{L}_1 \in \mathbb{R}^{3N \times F}$, $\mathbf{L}_2 \in \mathbb{R}^{2F \times N}$, $\{\mathbf{L}_3, \mathbf{L}_4\} \in \mathbb{R}^{3F \times N}$, $\mathbf{L}_5 \in \mathbb{R}^{F \times F}$ and $\mathbf{L}_6 \in \mathbb{R}^{N \times N}$ are the Lagrange multipliers; and $\alpha > 0$ is a penalty coefficient to improve convergence.

This minimization is highly under-constrained, but it can be carried out efficiently by solving each subproblem separately and in closed form, while keeping fixed the rests of variables. Algorithm 1 succinctly explains the details. The expressions for estimating $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{D}$ (steps 5, 7 and 10) are obtained by computing the derivatives of Eq. (6) in $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{D}$ and equating to zero. For $\mathbf{J}$, $\mathbf{K}$ and $\mathbf{X}$ matrices (steps 4, 6 and 8), we apply a Singular Value Thresholding minimization [13] with a 'shrinkage operator' $S_{\frac{*}{\alpha}}(x) = \max(0, x - \frac{*}{\alpha})$ where $* = \{1, \gamma\}$. The optimization of matrices $\mathbf{E}_t$ and $\mathbf{E}_s$ (steps 12 and 13) can be done in closed form by the element-wise shrinkage operator $S_{\frac{*}{\alpha}}(x) = \max(0, x - \frac{*}{\alpha})$ where $* = \{\lambda_s, \lambda_t\}$ [28]. After each iteration, the Lagrange multipliers are updated according to standard rules as shown in lines 14-19.

5.5. Spatial and Temporal Clustering

Once the affinity matrices $\mathbf{T}$ and $\mathbf{S}$ are estimated, we run the spectral clustering algorithm proposed in [14] to discover the actual clusters. Figure 1 shows an example of two matrices we obtain, where each entry (i, j) indicates the degree of similarity between the i-th and j-th frame (for the case of $\mathbf{T}$), or between the i-th and j-th data point (for the case of $\mathbf{S}$). The bar right below the affinity matrices represents the clusters discovered after applying [14]. The granularity of the segmentation can be controlled through a threshold on the eigenvalues internally computed by [14].

input : Possibly incomplete 2D trajectories $\bar{\mathbf{P}}$ and parameters $\{\lambda_t, \lambda_s, \gamma\}$ and $\{\alpha, \rho, \epsilon\}$
output: 3D reconstruction $\mathbf{D}$, camera rotation $\mathbf{G}$, spatial $\mathbf{S}$ and temporal $\mathbf{T}$ clustering

/* Complete 2D Traject., Eq. (5) */
1 **if** $\mathbf{P} \neq \bar{\mathbf{P}}$ **then** $\mathbf{P} = \min \| (\mathbf{O} \otimes \mathbf{1}_2) \odot (\mathbf{P} - \bar{\mathbf{P}}) \|_F^2 + \beta\|\mathbf{P}\|_*$
2 **else** $\mathbf{P} \equiv \bar{\mathbf{P}}$

/* Camera Rotation $\mathbf{G}$, Sect. 5.3 */

/* ALM optimization of Eq. (6) */
3 **while** *not converged* **do**
/* Update Model Parameters */
4 $\mathbf{J} = \min \frac{1}{\alpha}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{T} + \frac{\mathbf{L}_5}{\alpha})\|_F^2$
5 $\mathbf{T} = (\mathbf{X}^\top\mathbf{X} + \mathbf{I}_F)^{-1}(\mathbf{X}^\top(\mathbf{X} - \mathbf{E}_t) + \mathbf{J} + \frac{\mathbf{X}^\top\mathbf{L}_1 - \mathbf{L}_5}{\alpha})$
6 $\mathbf{K} = \min \frac{1}{\alpha}\|\mathbf{K}\|_* + \frac{1}{2}\|\mathbf{K} - (\mathbf{S} + \frac{\mathbf{L}_6}{\alpha})\|_F^2$
7 $\mathbf{S} = (\mathbf{D}^\top\mathbf{D} + \mathbf{I}_N)^{-1}(\mathbf{D}^\top(\mathbf{D} - \mathbf{E}_s) + \mathbf{K} + \frac{\mathbf{D}^\top\mathbf{L}_3 - \mathbf{L}_6}{\alpha})$
8 $\mathbf{X} = \min \frac{\gamma}{\alpha}\|\mathbf{X}\|_* + \frac{1}{2}\|\mathbf{X} - \big((\mathbf{E}_t - \frac{\mathbf{L}_1}{\alpha})(\mathbf{I}_F - \mathbf{T})^\top + q^{-1}(\mathbf{D} - \frac{\mathbf{L}_4}{\alpha})\big)\big((\mathbf{I}_F - \mathbf{T})(\mathbf{I}_F - \mathbf{T})^\top + \mathbf{I}_F\big)^{-1}\|_F^2$
9 $\mathbf{C} = \mathbf{G}^\top(\mathbf{P} + \frac{\mathbf{L}_2}{\alpha}) + (\mathbf{E}_s - \frac{\mathbf{L}_3}{\alpha})(\mathbf{I}_N - \mathbf{S}^\top) + \frac{\mathbf{L}_4}{\alpha} + q(\mathbf{X})$
10 $\text{vec}(\mathbf{D}) = \big(\mathbf{I}_N \otimes (\mathbf{G}^\top\mathbf{G} + \mathbf{I}_B) + \mathbf{B}^\top \otimes \mathbf{I}_B\big)^{-1}\text{vec}(\mathbf{C})$
11 $\mathbf{D} = \text{mat}(\text{vec}(\mathbf{D}))$
12 $\mathbf{E}_t = \min \frac{\lambda_t}{\alpha}\|\mathbf{E}_t\|_1 + \frac{1}{2}\|\mathbf{E}_t - (\mathbf{X} - \mathbf{X}\mathbf{T} + \frac{\mathbf{L}_1}{\alpha})\|_F^2$
13 $\mathbf{E}_s = \min \frac{\lambda_s}{\alpha}\|\mathbf{E}_s\|_1 + \frac{1}{2}\|\mathbf{E}_s - (\mathbf{D} - \mathbf{D}\mathbf{S} + \frac{\mathbf{L}_3}{\alpha})\|_F^2$
/* Update Lagrange Multipliers */
14 $\mathbf{L}_1 = \mathbf{L}_1 + \alpha(\mathbf{X} - \mathbf{X}\mathbf{T} - \mathbf{E}_t)$
15 $\mathbf{L}_2 = \mathbf{L}_2 + \alpha(\mathbf{P} - \mathbf{G}\mathbf{D})$
16 $\mathbf{L}_3 = \mathbf{L}_3 + \alpha(\mathbf{D} - \mathbf{D}\mathbf{S} - \mathbf{E}_s)$
17 $\mathbf{L}_4 = \mathbf{L}_4 + \alpha(q(\mathbf{X}) - \mathbf{D})$
18 $\mathbf{L}_5 = \mathbf{L}_5 + \alpha(\mathbf{T} - \mathbf{J})$
19 $\mathbf{L}_6 = \mathbf{L}_6 + \alpha(\mathbf{S} - \mathbf{K})$
/* Update penalty weights */
20 $\alpha = \min(\rho\alpha, 10^{12})$
/* Check Convergence */
21 $\|\mathbf{X} - \mathbf{X}\mathbf{T} - \mathbf{E}_t\|_\infty < \epsilon$
22 $\|\mathbf{P} - \mathbf{G}\mathbf{D}\|_\infty < \epsilon$
23 $\|\mathbf{D} - \mathbf{D}\mathbf{S} - \mathbf{E}_s\|_\infty < \epsilon$
24 $\|q(\mathbf{X}) - \mathbf{D}\|_\infty < \epsilon$
25 $\|\mathbf{T} - \mathbf{J}\|_\infty < \epsilon$
26 $\|\mathbf{S} - \mathbf{K}\|_\infty < \epsilon$
27 **end**
28 Notation: $\text{vec}(\cdot)$ and $\text{mat}(\cdot)$ are vectorization and matrization operators. $\mathbf{B} = (\mathbf{I}_N - \mathbf{S})(\mathbf{I}_N - \mathbf{S}^\top)$, $B = 3T$

Algorithm 1: Algorithm for optimizing Eq. (3).

6. Experimental Evaluation

We evaluate the proposed approach on the CMU MoCap Dataset. We consider several scenarios with two or more subjects interacting and performing complex motions (see videos in the supplemental material). Since 2D projections are not directly available on this dataset, we generate them

Method / Data / Metric:	CSF [21] e_X	KSTA [22] e_X	BMM [16] e_X	EM-PND [24] e_X	TUS [40] e_X	GBNR [18] e_X	CNR [25] e_X	Ours (DUST) 0% missing data e_X	e_S[%]	e_T[%]	sparse/structured e_X	e_X
Jump	0.053	0.071	0.078	0.065	0.054	0.070	0.074	**0.045**	0.0(2)	5.8(3)	0.047	0.062
Pull	0.123	0.128	0.146	**0.113**	0.116	0.138	0.183	0.118	0.0(2)	8.3(4)	0.120	0.121
Soldiers	0.104	0.106	0.080	0.342	0.073	0.076	0.091	**0.049**	1.2(2)	5.0(2)	0.050	0.067
Stares Down	0.036	0.022	0.050	**0.013**	0.032	0.048	0.038	0.016	0.0(2)	0.0(2)	0.018	0.024
Stumbles	**0.094**	0.102	0.124	0.099	0.112	0.119	0.119	0.096	0.0(2)	1.3(2)	0.098	0.111
Squats	0.047	0.041	0.040	0.055	0.016	0.036	0.023	**0.015**	4.8(2)	0.8(2)	0.018	0.021
Synchronized	0.141	0.145	0.152	0.145	0.091	0.147	0.112	**0.083**	0.0(2)	1.2(2)	0.085	0.086
Violence	0.072	0.073	0.090	0.150	0.081	0.085	0.135	**0.060**	0.0(2)	1.1(3)	0.062	0.076
Zombie	0.070	0.067	0.062	0.076	0.056	0.061	0.087	**0.043**	0.0(2)	9.3(3)	0.044	0.066
Average error:	0.082	0.084	0.091	0.117	0.070	0.087	0.096	**0.058**	0.6	3.6	0.060	0.070
Relative error:	1.41	1.44	1.56	2.01	1.21	1.50	1.65	**1.00**	-	-	1.04	1.21

Table 2. **Evaluation on CMU sequences with two subjects.** The table reports the 3D reconstruction error e_X for the following NRSfM baselines considering 2D tracks without missing data: CSF [21], KSTA [22], SPM [16], EM-PND [24], TUS [40], GBNR [18] and CNR [25]; and ours. For our approach, we also show the clustering errors e_S and e_T, where we include the number of spatial and temporal clusters in brackets. The two right-most columns show the reconstruction accuracy under random and structured patterns of missing data.

by synthesizing point tracks acquired by an orthographic camera that follows a circular trajectory around the scene, at an angular speed of $0.66\pi/sec$. In average, the sequences we consider below are 1,000 frames long, and the number of points per frame is either 82 (when considering two subjects) or 164 (four subjects).

For all experiments, we provide two types of validations: the 3D reconstruction accuracy that we compare to other NRSfM methods, and the results of the spatial and temporal subspace clustering, which is compared to a ground truth.

Regarding the reconstruction error, we report the normalized mean 3D error e_X, used before in [6, 16, 21]:

$$e_X = \frac{1}{\sigma F N}\sum_{f=1}^{F}\sum_{n=1}^{N} e_n^f, \quad \sigma = \frac{1}{3F}\sum_{f=1}^{F}(\sigma_x^f + \sigma_y^f + \sigma_z^f),$$

where e_n^f is the 3D error for the n-th point at frame f. σ_x^f, σ_y^f and σ_z^f are the error standard deviations at frame f.

We compare the reconstruction accuracy of our approach, denoted DUST, against seven NRSfM baselines: the trajectory-space methods CSF [21] and KSTA [22]; the block matrix approach BMM [16], the probabilistic-normal-distribution method EM-PND [24], the temporal union of subspaces TUS [40], the grouping-based NRSfM of GBNR [18] and the consensus NRSfM of CNR [25]. For CSF [21] and KSTA [22], we manually set the rank of the subspace to the value yielding the best results. For TUS [40], we use our own implementation as its source code is not publicly available. Our method does not require tuning the subspace rank parameter. Note that all methods decouple the problems of camera rotation estimation and shape reconstruction. In order to focus our analysis on the 3D shape reconstruction capacity, we will provide the same ground truth matrix $\mathbf{G}$ of camera rotations to all methods. The results when the camera motion is estimated are reported in the supplementary material.

For the assessment of the subspace clustering accuracy, we compare our results with a ground truth clustering obtained as follows: First, the 'ground truth' similarity matrices $\mathbf{S}^{GT}$ and $\mathbf{T}^{GT}$ are computed by applying the low-rank representation proposed in [29] over the matrices $\hat{\mathbf{X}}$ and $\mathbf{X}$ with the true 3D point positions. We then perform spectral clustering [14] over $\mathbf{S}^{GT}$ and $\mathbf{T}^{GT}$ to retrieve $\mathcal{S}^{GT}$ and $\mathcal{T}^{GT}$, which are N- and $F-$ dimensional vectors where each entry is an integer representing the ground truth cluster index. If we denote by $\mathcal{S}$ and $\mathcal{T}$ the corresponding cluster indexes obtained from the similarity matrices estimated by our approach, we define the following clustering errors:

$$e_S = \frac{100}{N}\sum_{i=1}^{N}\mathbb{I}(\mathcal{S}_i \neq \mathcal{S}_i^{GT}), \quad e_T = \frac{100}{F}\sum_{f=1}^{F}\mathbb{I}(\mathcal{T}_f \neq \mathcal{T}_f^{GT}),$$

where $\mathbb{I}(a)$ is the indicator function, *i.e.*, $\mathbb{I}(a) = 1$ if a is true, and 0 otherwise. In practice, for the results we report below, we run [14] for different levels of granularity and keep the result that minimizes e_S and e_T.

6.1. Sequences with Two Subjects

We select nine sequences of the CMU dataset with two subjects performing different activities and motion patterns. Namely, we consider *23_16 (Synchronized)*: subjects alternating synchronized jumping jacks; *19_05 (Pull)*: a subject pulls the other by the elbow; *22_20 (Violence)*: a subject picks up high stool and threatens to strike the other; *20_08 (Zombie)*: subjects follow a zombie march; *20_06 (Soldiers)*: subjects follow a soldiers march; *23_19 (Stares Down)*: a subject stares down the other and leans with hands on high stool; *22_12 (Stumbles)*: a subject stumbles into the other; *23_15 (Jump)*: subjects alternating jumping jacks; and *23_14 (Squats)*: subjects alternating squats.

Table 2 summarizes the reconstruction errors for all methods and the subspace clustering accuracy of ours. Note that DUST consistently outperforms state-of-the-art in terms of 3D reconstruction, reducing the 3D error of other

Method / Data / Metric:	CSF [21] e_X	KSTA [22] e_X	BMM [16] e_X	EM-PND [24] e_X	TUS [40] e_X	GBNR [18] e_X	CNR [25] e_X	Ours (DUST) 0% missing data e_X	e_S[%]	e_T[%]	sparse/structured e_X	e_X
Blind4	0.047	**0.040**	0.079	0.079	0.059	0.074	0.137	0.045	0.0(4)	0.3(2)	0.045	0.052
Chicken4	0.030	0.034	0.027	0.022	0.017	0.021	0.022	**0.015**	0.0(4)	0.2(3)	0.018	0.022
Greet4	0.048	**0.041**	0.078	0.069	0.072	0.077	0.085	0.051	0.0(4)	2.0(3)	0.052	0.053
Shelters4	0.055	0.053	0.087	0.053	0.037	0.085	0.069	**0.034**	0.0(3)	3.2(2)	0.034	0.045
Soda4	0.011	0.011	0.009	0.010	0.009	0.011	0.016	**0.007**	0.0(4)	1.0(2)	0.008	0.011
Synchronized4	0.093	0.077	0.056	0.042	0.046	0.049	0.078	**0.041**	0.0(4)	1.2(2)	0.043	0.045
Zombie4	0.055	0.067	0.047	0.051	0.043	0.046	0.061	**0.033**	0.0(4)	8.9(3)	0.034	0.034
Average error:	0.048	0.046	0.055	0.046	0.040	0.052	0.067	**0.032**	0.0	2.4	0.033	0.037
Relative error:	1.50	1.43	1.69	1.44	1.26	1.62	2.09	**1.00**	-	-	1.03	1.17

Table 3. **Quantitative comparison on human interaction with multiple subjects.** See caption of Table 2.

methods by large margins between the 21% and 101%. Additionally, DUST also performs shape and temporal clustering. The quality of these segmentations is also very good. In particular, the number of spatial clusters we retrieve in all experiments is two, and all points are correctly assigned to the specific subject. The number of temporal clusters we estimate is between 2 and 4, and the exact temporal split (*i.e.*, the moment when one sub-action switches to another one) is very close (if not equal) to that of the ground truth. Indeed, most temporal clusters match real motion primitives (*e.g.*, 'jump in' and 'jump out' in Fig. 1).

Figure 3 shows a qualitative comparison of the similarity matrices we estimate and those of the ground truth, which are directly computed from clean 3D data. Despite the matrices provided by our approach are noisier, we can clearly identify the same patterns as in the ground truth. The spectral algorithm we use [14] can easily handle this and yields the correct number of clusters in almost all experiments. In Fig. 4, we show several frames of the 3D reconstruction results for the 'Violence' sequence.

Robustness to Occlusions. We explicitly test the robustness to occlusions of our approach by artificially removing entries of the observation matrix $\bar{\mathbf{P}}$. We consider two cases: 1) sparse occlusion patterns generated by randomly removing 40% of the input data, and 2) structured noise, by removing rectangular regions (one per object) from the data matrix. For this case, the amount of non-observed data is approximately 15%. The results of these experiments are shown in the right-most columns of Table 2. The completion algorithm described in Sect. 5.2 does a pretty good job hypothesizing the missing observations, especially for the random scattered occlusions, and the final reconstruction is nearly unaffected by these artifacts. The accuracy of the clustering is almost identical to that for the artifacts-free case.

6.2. Sequences with Four Subjects

We also considered a more complex case with four subjects. Since the CMU dataset only includes sequences with one or two subjects, we combined several of them to generate seven new sequences including four subjects, namely: *Synchronized4*: subjects alternating synchronized jumping jacks; *Zombie4*: subjects follow a zombie march; *Chicken4*: subjects perform a non-synchronized chicken dance; *Greet4*: subjects walk and shake hands; *Blind4*: four subjects blind man's bluff; *Soda4*: two subjects pass soda to the other two and all of them drink; and *Shelters4*: two subjects individually shelter the other two. Again an orthographic camera moving slowly around the scene is considered. In these examples, the degree of superposition in the image plane is so extreme, that the task of performing the spatial segmentation becomes very difficult. Indeed, in some of the sequences two of the subjects are so intimately connected, that they can be interpreted as one single object.

The results are summarized in Table 3. Again, our approach improves other NRSfM approaches in reconstruction accuracy by a large margin. It is worth pointing out the good performance of KSTA [22] for the sequences in which the subjects perform larger trajectories ('Blind4' and 'Greet4'). Regarding the segmentation accuracy, note that for the 'Shelters4' we obtain a better segmentation accuracy by choosing a spatial granularity of 3 instead of 4. This is because for this specific sequence two of the objects are always together. We show some similarity matrices and reconstruction examples in Figs. 3 and 4.

Finally, we tested the robustness of the method to scattered and structured occlusions, and as shown in the rightmost columns of Table 3, our approach again demonstrates great resilience.

7. Conclusion

In this paper we have proposed a novel solution to the NRSfM paradigm that allows exploring a problem which had not been tackled before: given a possibly incomplete monocular sequence of 2D tracks, estimating 3D non-rigid shape while also providing temporal clustering of the data into deformation-primitives, and spatial segmentation into multiple objects. For this purpose, we have presented a new low-rank model to represent the shape as a dual combination of spatial and temporal subspaces. We formulate an optimization problem based on this representation which we solve by means of augmented Lagrange machinery. We

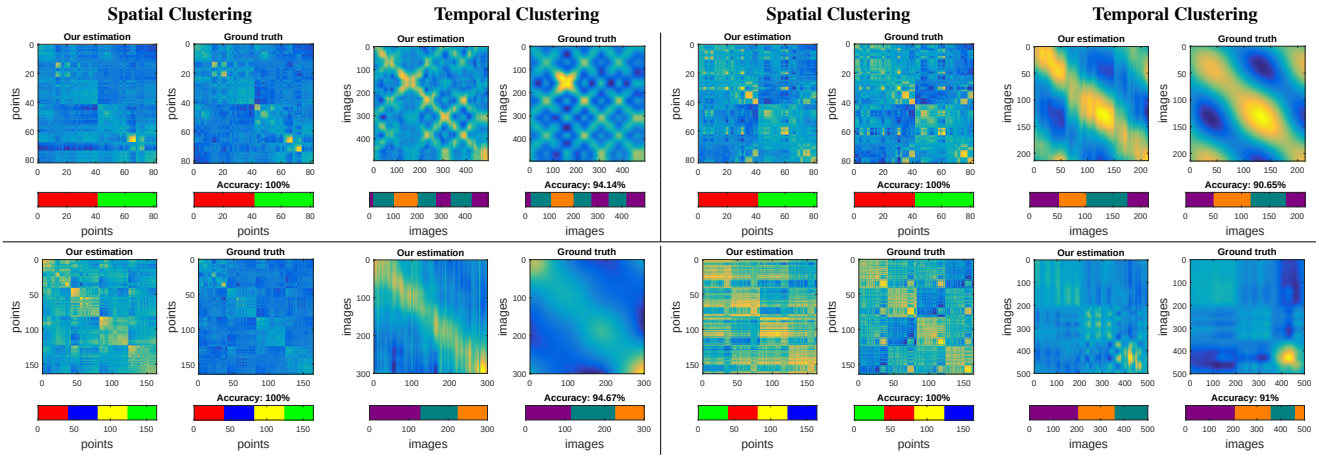

Figure 3. **Spatial and temporal clustering on CMU sequences.** We compare the spatial **S** and temporal **T** clustering matrices obtained with our approach with the ground truth ones. Below each matrix we plot a bar with the results of the spectral clustering. **Top:** *Jump* and *Zombie* sequences with two subjects and three primitives. **Bottom:** *Blind4* and *Chicken4* sequences with four subjects and three primitives.

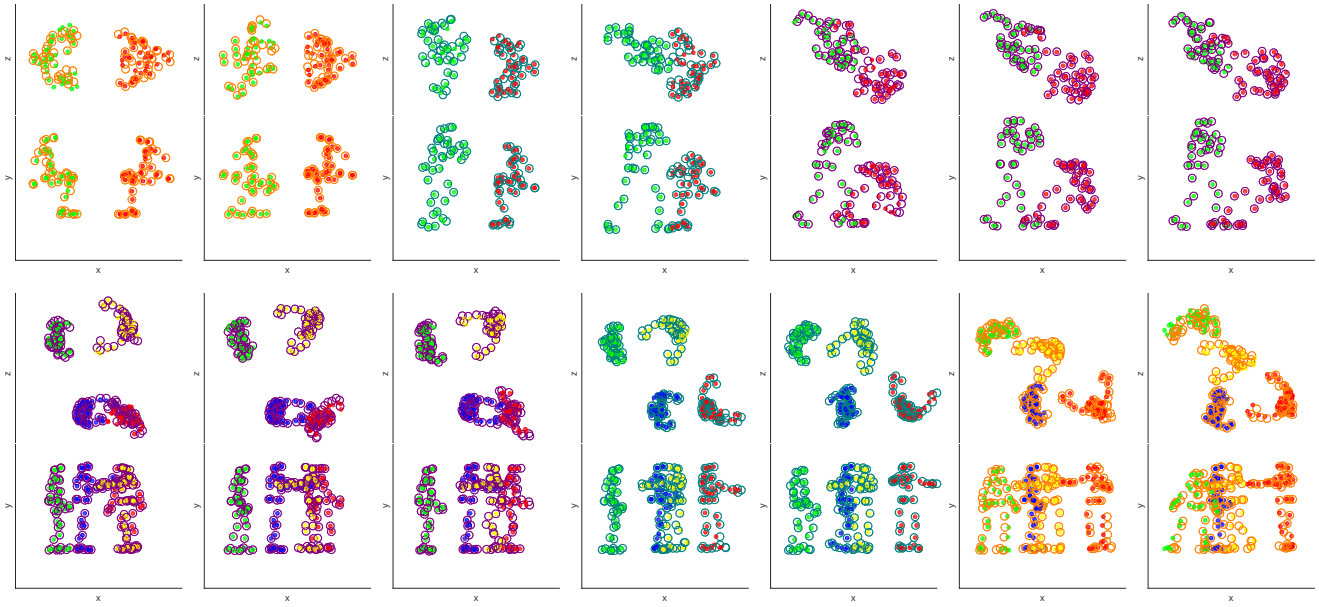

Figure 4. **3D reconstruction and spatio-temporal segmentation on multi-subject sequences**. Results for the 'Violence' (top) and 'Blind4' (bottom) sequences. For each scene we plot several frames, seen from two viewpoints (x-y and x-z). Colored dots represent the 3D position and spatial cluster index estimated by our approach. Note that the two subjects (top) and the four subjects (bottom) are clearly identified. No single point is assigned to a wrong subject. Empty circles indicate the ground truth 3D position. The color of these circles encodes to which temporal prior does the frame belong. Observe that in both sequences we identify three temporal priors. For the sequence on the top ('Violence'), the priors have a clear physical meaning: 'two subjects sitting down', 'one subject standing up an threatening the second one', 'one subject attacks the other that falls down'. The physical interpretation of the temporal priors for the four-subject sequence on the bottom is not that clear, but they seems to encode the type of subject interactions.

have thoroughly evaluated the approach on challenging sequences involving up to four interacting persons performing complex motion patterns. We show that besides providing correct spatio-temporal segmentation, our approach does also reconstruct the 3D human poses more accurately than current state-of-the-art NRSfM methods. In the future, we aim at using this research as a first step to perform complete reconstruction and recognition of human activities.

Acknowledgment: This work has been partially supported by the Spanish Ministry of Science and Innovation under project RobInstruct TIN2014-58178-R, by the ERA-net CHISTERA project I-DRESS PCIN-2015-147 and by the EU project LOGIMATIC H2020-Galileo-2015-1-687534.

References

[1] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Modal space: A physics-based model for sequential estimation of time-varying shape from monocular video. *JMIV*, 57(1):75–98, 2017.

[2] A. Agudo, J. M. M. Montiel, B. Calvo, and F. Moreno-Noguer. Mode-shape interpretation: Re-thinking modal space for recovering deformable shapes. In *WACV*, 2016.

[3] A. Agudo and F. Moreno-Noguer. Learning shape, motion and elastic models in force space. In *ICCV*, 2015.

[4] A. Agudo and F. Moreno-Noguer. Global model with local interpretation for dynamic shape reconstruction. In *WACV*, 2017.

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

[6] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Non-rigid structure from motion in trajectory space. In *NIPS*, 2008.

[7] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. *Technical report HAL-00345747*, 2008.

[8] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *CVPR*, 2008.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *FTML*, 3(1):1–122, 2011.

[10] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

[11] S. Burer and R. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Math. Prog.*, 103(3):427–444, 2005.

[12] R. Cabral, F. de la Torre, J. P. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, 2013.

[13] J. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM JO*, 20(4):1956–1982, 2010.

[14] W. Y. Chen, Y. Song, H. Bai, C. Lin, and E. Chang. Parallel spectral clustering in distributed systems. *TPAMI*, 33(3):568–586, 2010.

[15] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In *ICCV*, 1995.

[16] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure from motion factorization. In *CVPR*, 2012.

[17] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR*, 2009.

[18] K. Fragkiadaki, M. Salas, P. Arbeláez, and J. Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In *NIPS*, 2014.

[19] Y. Gao and A. L. Yuille. Symmetric non-rigid structure from motion for category-specific object structure estimation. In *ECCV*, 2016.

[20] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, 2013.

[21] P. F. U. Gotardo and A. M. Martinez. Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *TPAMI*, 33(10):2051–2065, 2011.

[22] P. F. U. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *ICCV*, 2011.

[23] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *ICCV*, 1999.

[24] M. Lee, J. Cho, C. H. Choi, and S. Oh. Procrustean normal distribution for non-rigid structure from motion. In *CVPR*, 2013.

[25] M. Lee, J. Cho, and S. Oh. Consensus of non-rigid reconstructions. In *CVPR*, 2016.

[26] M. Lee, C. H. Choi, and S. Oh. A procrustean markov process for non-rigid structure recovery. In *CVPR*, 2014.

[27] Z. Li, J. Guo, L. Cheong, and Z. Zhou. Perspective motion segmentation via collaborative clustering. In *ICCV*, 2013.

[28] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215*, 2009.

[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[30] K. Ozden, K. Schindler, and L. Van Gool. Multibody structure-from-motion in practice. *TPAMI*, 32(6):1134–1141, 2010.

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

[32] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3D reconstruction of a moving point from a series of 2D projections. In *ECCV*, 2010.

[33] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete or corrupted trajectories. *TPAMI*, 32(10):1832–1845, 2010.

[34] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3D reconstruction of dynamic scenes. In *ECCV*, 2014.

[35] T. Simon, J. Valmadre, I. Matthews, and Y. Sheikh. Separable spatiotemporal priors for convex reconstruction of time-varying 3D point clouds. In *ECCV*, 2014.

[36] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2):137–154, 1992.

[37] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors. *TPAMI*, 30(5):878–892, 2008.

[38] J. Valmadre and S. Lucey. General trajectory prior for non-rigid reconstruction. In *CVPR*, 2012.

[39] L. Zappella, A. Del Bue, X. Llado, and J. Salvi. Joint estimation of segmentation and structure from motion. *CVIU*, 117(2):113–129, 2013.

[40] Y. Zhu, D. Huang, F. de la Torre, and S. Lucey. Complex non-rigid motion 3D reconstruction by union of subspaces. In *CVPR*, 2014.