Complex Non-Rigid Motion 3D Reconstruction by Union of Subspaces

Yingying Zhu
University of Queensland
zhuyingying2@gmail.com

Dong Huang
Carnegie Mellon University
dghuang@andrew. cmu.edu

Fernando De La Torre
Carnegie Mellon University
ftorre@cs.cmu.edu

Simon Lucey
Carnegie Mellon University
slucey@andrew.cmu.edu

Abstract

The task of estimating complex non-rigid 3D motion through a monocular camera is of increasing interest to the wider scientific community. Assuming one has the 2D point tracks of the non-rigid object in question, the vision community refers to this problem as Non-Rigid Structure from Motion (NRSfM). In this paper we make two contributions. First, we demonstrate empirically that the current state of the art approach to NRSfM (i.e. Dai et al. [5]) exhibits poor reconstruction performance on complex motion (i.e motions involving a sequence of primitive actions such as walk, sit and stand involving a human object). Second, we propose that this limitation can be circumvented by modeling complex motion as a union of subspaces. This does not naturally occur in Dai et al.'s approach which instead makes a less compact summation of subspaces assumption. Experiments on both synthetic and real videos illustrate the benefits of our approach for the complex nonrigid motion analysis.

1. Introduction

Recovering non-rigid 3D motion from 2D point tracks stemming from a monocular camera is a problem of considerable interest in numerous disciplines and applications throughout science and industry. This task is widely referred to as Non-Rigid Structure from Motion (NRSfM) in the vision community. Dai et al. [5], recently proposed a strategy for motion reconstruction that is now considered state of the art in the field. The approach has close links with approaches in the learning community, most notably Robust PCA (RPCA) [4], for recovering low-rank subspaces in the presence of noise.

In this paper we demonstrate empirically and characterize theoretically that the utility of Dai et al.'s [5] approach for reconstructing 3D motion is highly dependent on the complexity of the motion. We use the term "complex" qualitatively to describe motions with multiple "primitive" or "simple" actions. For example, in the case of a human object a motion sequence containing the subject raising their hand, walking, sitting then standing would be considered complex, whereas the individual actions themselves would be considered simple (see Figure 1).

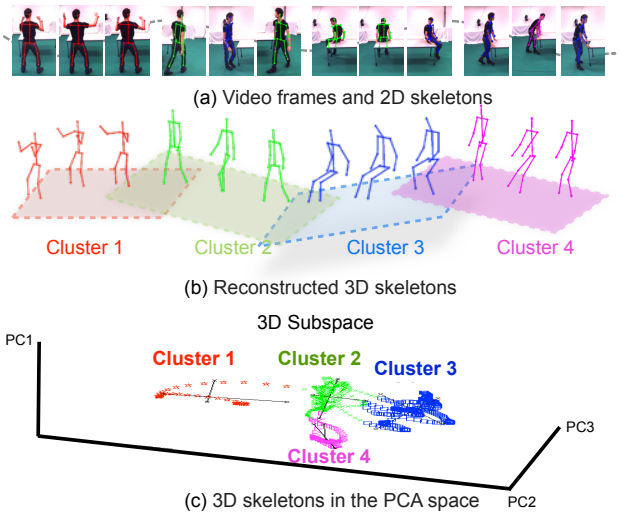


Figure 1. An example of complex non-rigid motion using a human body. (a) A video sequence from the UMPM dataset [1] in which a subject sequentially performs actions such as: raise hand (red), walk (green), sit (blue) and stand (magenta). 2D body joints tracked in the videos are connected to form 2D skeletons in each frame. (b) Reconstructed and clustered 3D skeletons using our method. Different color represents different clusters/subspaces obtained by our method. (c) Projection of the 3D-skeletons in the local subspaces spanned by the three largest principal components (PCs). Observe that the human poses stemming from different actions adhere to separate local subspaces/clusters and the overall complex nonrigid motion lies in a union of subspaces. (**Best viewed in color**).

An obvious strategy, is to first group the sequence into local clusters, and then apply Dai et al.'s approach to each

cluster. This approach is problematic from two perspectives. First, one would have to perform cluster on the 2D projected motion, rather than the preferred raw 3D motion making the approach extremely sensitive to camera motion and projection ambiguities. Second, one would need to also solve the non-trivial problem of how many local clusters are in the motion sequence.

It has been noted by Liu et al. [9], that when data is drawn from a union of subspaces RPCA [4] actually treats the data as being sampled from a single subspace. Since the sum can be much larger than the union the specifics of the individual subspaces are not well considered and so the reconstruction may be inaccurate. We argue in this paper that Dai et al.'s approach to NRSfM intrinsically makes this same assumption and suffers from the same inaccuracies when reconstructing complex motion. To address this problem we propose an approach to NRSfM aimed directly at motion sequences stemming from the union of subspaces. Our approach attempts to simultaneously learn the 3D reconstruction and the affinity matrix to cluster 3D nonrigid motion into a union of subspaces The affinity matrix is of great importance here as it naturally encodes the cluster/subspace membership of each sample in the motion sequence.. We demonstrate impressive reconstruction and results over a series of 2D projected human action sequences.

2. Nonrigid Structure from Motion

In this section, we briefly review the methods on Non-Rigid Structure from Motion (NRSfM), paying particularly attention to the NRSfM method of Dai et al. [5] considered by most in the vision community as state of art. Furthermore, we discuss the problems of Dai et al.'s approach on complex nonrigid motion reconstruction.

Factorization Methods: Factorization approaches, first proposed for recovering rigid structures by Tomasi and Kanade in [11], were extended to handle non-rigidity in the seminal work by Bregler et al. [3]. As the name suggests, factorization approaches to NRSfM assumes that the 3D non-rigid structure is of intrinsically low rank. A natural way of discovering this low rank is through the employment of either Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) on the projected 2D structure.

In reality the projection matrix is never low-rank due to measurement and sample error. Fortunately, there exists a suite of SVD/PCA style algorithms for discovering the "clean" low-rank matrix in the presence of noisy data. Candès et al. [4] recently proposed Robust PCA (RPCA) which assume that the "clean" data lies in a single low dimensional subspace. RPCA solves the following objective,

$$\begin{aligned} \arg\min_{\mathbf{D},\mathbf{E}} \quad & \|\mathbf{D}\|_* + \lambda\|\mathbf{E}\|_\ell \\ \text{s. t.} \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \end{aligned} \tag{1}$$

where $\mathbf{D}$ is the "cleaned" low-rank matrix that models a single low dimensional subspace, $\mathbf{E}$ is the residual noise and $\lambda > 0$ is the error penalty parameter. The use of the nuclear norm $\|\cdot\|_*$ on $\mathbf{D}$ is a convex approximation to rank, and $\|\mathbf{E}\|_\ell$ on $\mathbf{E}$ is the norm on the error where $\ell = 1$ and $\ell = 1, 2$ denotes the convex approximations to sparse or group sparse error respectively. An inherent advantage of the objective in Equation (1) is that it is convex and can be efficiently solved using Augmented Lagrangian Methods (ALMs) [2] allowing it to handle large scale problems in both learning and vision.

Dai et al.'s Approach: Dai et al. proposed an elegant "prior-free" solution to NRSfM [5] which adopted a single low dimensional subspace for modeling the nonrigid 3D shapes. The authors claim the approach to be prior-free in the sense that it does not make any prior assumption about the non-rigid structure (other than low rank)

$$\begin{aligned} \arg\min_{\mathbf{X},\mathbf{E}} \quad & \|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_\ell \\ s.t. \quad & \mathbf{W} = \mathbf{R}\mathbf{X}^{\#} + \mathbf{E} \ . \end{aligned} \tag{2}$$

Similar to RPCA, the objective in Equation (2) is convex and can be efficiently solved using ALMs. One thing in particular to note in Dai et al's approach is that the nuclear norm of $\mathbf{X}$ is being minimized, rather than $\mathbf{X}^{\#}$. This is done because the rank of $\mathbf{X}$ is bound by $\min(F, 3N)$ whereas the rank of $\mathbf{X}^{\#}$ is bound by $\min(3F, N)$. Minimizing the rank of $\mathbf{X}$ is preferable as it attempts to directly learn redundancies between frames. In Dai et al.'s original method the norm on $\mathbf{E}$ was assumed to be $\ell = 2$, although other error norms $\ell = 1$ or $\ell = 1, 2$ are possible as in RPCA. Similar to traditional SFM, the camera matrix $\mathbf{R}$ in this method is pre-computed from the 2D sequence given the rank of $\mathbf{X}$ a priori. In practice, we can alternatively track the points of the rigid background in the video, and use the standard Tomasi & Kanade rigid SFM to estimate the cameras. Dai et al.'s [5] work was extended by Lee et al. [7] to estimate a better rotation matrix $\mathbf{R}$ by a generalized Procrustes analysis. However, due to the least-square nature of Lee et al.s cost function, the reconstructed 3D structures are still modeled as a single mode Gaussian distribution, or geometrically speaking a single subspace. In other word, theres no fundamental difference between Lee et al. and Dai et al. on modeling the non-rigid motion. Lee et al.s experiments on single action sequences show, in some cases (not all cases), improvements over Dai et al.'s.

Complex Nonrigid Motion: The focus of our work in this paper is reconstructing complex nonrigid motion from projections on monocular camera image sequences. Complex nonrigid motion contains multiple primitives or simple actions, which are shown empirically tend to adhere to a union of subspaces in Fig. 1 . When data is drawn from a union

of K subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_K$, Liu et al. [9] has noted that the feasible region of the RPCA solution is a convex envelop over the underling multi-subspaces, which is a summation of multi-subspaces. In other words, RPCA only treats the data as being sampled from a single summation of subspaces,

$$\sum_{k=1}^{K} \mathcal{S}_k = \{\mathbf{x} : \mathbf{x} = \sum_{k=1}^{K} \mathbf{x}_k, \quad \mathbf{x}_i \in \mathcal{S}_i\} \ . \tag{3}$$

Since the sum $\sum_{k=1}^{K} \mathcal{S}_k$ can be much larger than the union

$$\cup_{k=1}^{K} \mathcal{S}_k = \{\mathbf{x} : \mathbf{x} \in \mathcal{S}_k, \quad \text{for some } 1 \leq k \leq K\}, \tag{4}$$

the specifics of the individual subspaces are not well considered and so the reconstruction may be inaccurate.

We argue that Dai et al.'s approach implicitly make a similar summation subspace assumption as RPCA and suffers from the same inaccuracy on reconstructing complex nonrigid motion. Particularly, Dai et al.'s method seeks the solutions on the boundary of the"loose" feasible region (the summation of individual subspaces) for modeling complex nonrigid motion. Therefore, it gives inaccurate 3D reconstruction as the combinations of the true 3D motion lying in individual subspaces.

3. Clustering Subspaces: 2D vs. 3D

An obvious strategy is to apply Dai et al.'s approach in conjunction with clustering 3D local subspaces. Specifically, one could apply a canonical clustering approach to the 2D projected motion then apply Dai et al.'s approach to each cluster. However, clustering nonrigid shapes into 3D local subspaces using 2D projections is problematic due to the inherent projection ambiguities. Further, if one wants to successfully categorize whether two nonrigid shape frames belong to the same 3D cluster/subspace, one first needs to remove differences between these frames according to a 3D rigid transform. This is trivial to do in 3D, but difficult to do in 2D (with the exception of in plane 2D translation, rotation and scale). We refer to this problem herein as **relative camera motion**. Nonrigid shape registration requires similar 3D nonrigid shapes to be as aligned as possible (zero relative rigid camera motion). Unfortunately, 2D nonrigid shape registration alone is unable to remove the 3D rigid relative camera motion.

Experimental Setup: This section will highlight empirically the intrinsic difference of 2D and 3D nonrigid shape clustering. We applied canonical subspace clustering [9] on a human motion sequence taken from the CMU MoCap Dataset [1]. The sequence we used (the 10^{th} sequence of subject 86) contains 3D points of the subject performing 4 different actions: walk, sit, stand and run (See Figure 2). We generated 2D projected sequences from the 3D points under an orthographic assumption $\mathbf{W} = \mathbf{R}\mathbf{X}^{\#}$, where $\mathbf{X}^{\#} \in \mathbb{R}^{3F \times N}$ is the 3D coordinates of N points for F frames, $\mathbf{R} \in \mathbb{R}^{2F \times 3F}$ is the block diagonal matrix of F 2×3 camera matrices, and $\mathbf{W} \in \mathbb{R}^{2F \times N}$ are the 2D projections. We reserve $\mathbf{X} \in \mathbb{R}^{3N \times F}$ as a reshaping of $\mathbf{X}^{\#}$ (where the $^{\#}$ has been omitted), the reasons for this distinction has been elucidated in Section 3. The 3D and 2D points are respectively aligned by 3D and 2D affine transformation with respect to the torso.

For clustering nonrigid shapes, we employed the Local Rank Representation (LRR) [9] method to estimate the affinity matrices for both 3D and 2D data (a full review of subspace clustering is outside the scope of this paper, readers are encouraged to inspect [9] for a full treatment). K-means clustering was then applied to the affinity matrices to obtain the final group/clusters [6].

To show the impact of camera motion on the 2D nonrigid shape clustering results, we synthetically generated two sequences of 2D projected human motion using a static and fast moving orthographic camera. The y-axis of the camera was pointed to the center of the moving subject ensuring that the camera was rotating around the subject.

Results: Clustering results can be seen in Fig. 2 with each cluster denoted by different colors. We compared the 2D cluster results of the static camera and moving camera to the 3D cluster results (which we consider here as our ground truth). The 2D clustering results show a substantial difference to 3D clustering in both the static and moving camera cases. In Fig. 2 (b), the 2D LLR clustering for the static camera depicts confusion between cluster 2 and cluster 4 possibly stemming from the projection ambiguity. In Fig. 2 (c) we see a larger departure from the 3D clustering result due to the additional camera motion. This experiment validates our assumption that the 2D clustering is not effective to build 3D local subspaces due to projection ambiguity and relative 3D rigid camera motion.

4. Our Approach

Reconstructing 3D complex nonrigid motion from 2D projections on image sequences requires three interdependent sub-tasks: (i) registration in 3D by removing 3D rigid relative camera motion; (ii) clustering 3D nonrigid motion into a union of local subspaces; (iii) reconstructing 3D nonrigid motion in each local subspaces. Fortunately, task (i) is naturally accounted for in the NRSfM formulation. As shown in Equation 2, camera rotation matrix $\mathbf{R}$ "absorbs" all the 3D rigid relative camera motion. In other words, NRSfM is able to align the 2D projections in 3D space automatically.The problem of simultaneously

[1] More details can be found at http://mocap.cs.cmu.edu/

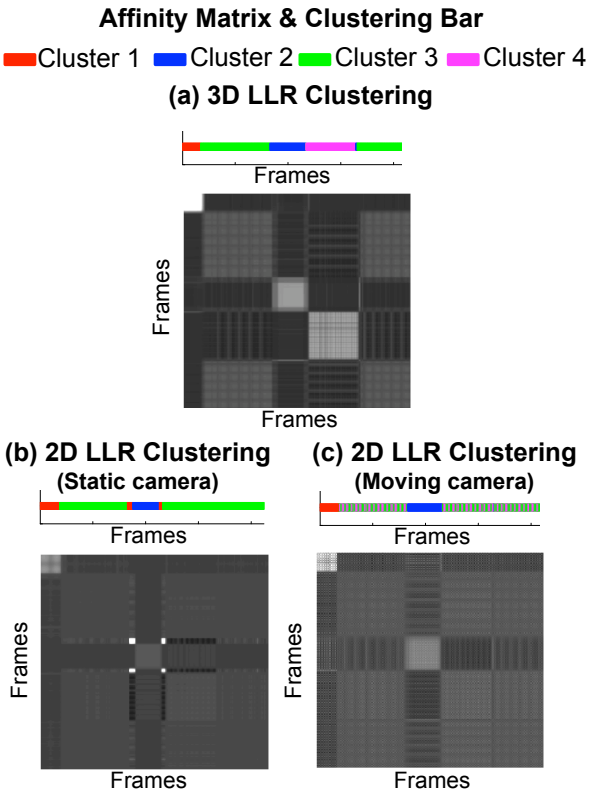


Figure 2. 3D LRR subspace clustering vs. 2D LRR subspace clustering on complex nonrigid motion. The 2D projections are generated by different camera motion. We show the 3D LRR subspace clustering and 2D LRR subspace clustering results from an stationary camera and fast moving camera. In case of stationary camera, there is no camera motion, the only difference is the projection ambiguity between 2D and 3D nonrigid shapes. In case of fast moving camera, the 2D subspace clustering is corrupted by camera motion. (**Best viewed in color**)

addressing tasks (ii) and task (iii) will now be described.

Drawing Inspiration from LRR: To recover the affinity matrix associated with a union of subspaces model in the presence of noise, Liu et al. proposed an alternative to RPCA which they referred to as Low Rank Representation (LLR) [9]. In this method the authors advocated solving for the affinity matrix $\mathbf{Z}$ directly,

$$\begin{aligned} \arg\min_{\mathbf{Z},\mathbf{E}} \quad & \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_\ell \\ s.t. \quad & \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E} \ . \end{aligned} \tag{5}$$

Affinity matrices are important when one wants to ascertain if a dataset belongs to a union of subspaces as it encodes the subspace/cluster membership of each sample in the ensemble. The power of this approach stems from the fact that it is directly estimating the affinity matrix $\mathbf{Z}$ from $\mathbf{X}$ in the presence of noise, as opposed to the canonical approach of first obtaining a cleaned version of $\mathbf{X}$ (through a procedure like RPCA) from which the affinity matrix is then found indirectly.

NRSfM in Union of Subspaces : Inspired by LRR, we propose a NRSfM strategy for simultaneously solving for the 3D structure $\mathbf{X}$ and the affinity matrix $\mathbf{Z}$

$$\begin{aligned} \arg\min_{\mathbf{X},\mathbf{Z},\mathbf{E}} \quad & \|\mathbf{Z}\|_* + \gamma\|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_\ell \\ s.t. \quad & \mathbf{X} = \mathbf{X}\mathbf{Z} \\ & \mathbf{W} = \mathbf{R}\mathbf{X}^\# + \mathbf{E}, \end{aligned} \tag{6}$$

where γ and λ are penalty parameters for $\|\mathbf{X}\|_*$ and $\|\mathbf{E}\|_\ell$ respectively. Equation (6) has two constraints: (i) $\mathbf{X} = \mathbf{X}\mathbf{Z}$ (which we shall refer to as the subspace clustering constraint), and (ii) $\mathbf{W} = \mathbf{R}\mathbf{X}^\# + \mathbf{E}$ (which we shall refer to as the NRSfM constraint). The subspace clustering constraint automatically enforces the union of subspaces structure of $\mathbf{X}$ with a low-rank coefficients matrix $\mathbf{Z}$, while the NRSfM constraint performs 3D reconstruction and registration from the 2D projections $\mathbf{W}$ to the 3D points $\mathbf{X}$. We are simultaneously reconstructing and clustering the 3D complex nonrigid motion $\mathbf{X}$ in a union of subspaces through the low rank coefficients/affinity matrix $\mathbf{Z}$.

Solving Equation 6: Equation (6) can be efficiently optimized using Augmented Lagrangian Methods (ALMs) [2]. In equation 6, matrices $\mathbf{X}$ and $\mathbf{X}^\#$ contain the same elements but with different organization. We introduced matrix $\mathbf{H} = \mathbf{X}$ to connect $\mathbf{X}^\#$ and $\mathbf{X}$, and ($\mathbf{H} = \mathbf{X}$) subsequently is used as the 3rd constraint appended to equation 6. The complete cost function becomes

$$\begin{aligned} \min_{\mathbf{X},\mathbf{Z},\mathbf{E},\mathbf{H}} \mathcal{L} \quad = \quad & \|\mathbf{Z}\|_* + \gamma\|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_1 \\ & + < \mathbf{\Gamma}_1, \mathbf{X} - \mathbf{X}\mathbf{Z} > + \frac{\mu_1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2 \\ & + < \mathbf{\Gamma}_2, \mathbf{W} - \mathbf{R}\mathbf{H}^\# - \mathbf{E} > \\ & + \frac{\mu_2}{2}\|\mathbf{W} - \mathbf{R}\mathbf{H}^\# - \mathbf{E}\|_F^2 \\ & + < \mathbf{\Gamma}_3, \mathbf{X} - \mathbf{H} > + \frac{\mu_3}{2}\|\mathbf{X} - \mathbf{H}\|_F^2. \end{aligned} \tag{7}$$

where $\mathbf{\Gamma}_1 \in \mathbb{R}^{3N\times F}, \mathbf{\Gamma}_2 \in \mathbb{R}^{3F\times N}$ are the Lagrange multiplier matrices and μ_1, μ_2 are the penalty parameters,$\mathbf{X}^\# \in \mathbb{R}^{3F\times N}$ is a reshape matrix of $\mathbf{X} \in Re^{3N\times F}$.

The variables $\{\mathbf{X}, \mathbf{Z}, \mathbf{E}, \mathbf{H}\}$ are solved by the following subproblems:

$$\mathbf{X}^{k+1} = \arg\min_{\mathbf{X}} \mathcal{L}(\mathbf{X}^k, \mathbf{Z}, \mathbf{E}, \mathbf{H}), \tag{8}$$

$$\mathbf{Z}^{k+1} = \arg\min_{\mathbf{Z}} \mathcal{L}(\mathbf{X}, \mathbf{Z}^k, \mathbf{E}, \mathbf{H}), \tag{9}$$

$$\mathbf{E}^{k+1} = \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{X}, \mathbf{Z}, \mathbf{E}^k, \mathbf{H}), \tag{10}$$

$$\mathbf{H}^{k+1} = \arg\min_{\mathbf{H}} \mathcal{L}(\mathbf{X}, \mathbf{Z}, \mathbf{E}, \mathbf{H}^k) \tag{11}$$

where k is the index of iterations. The reader can refer to the supplementary material for the detailed algorithm.

5. Experiments

We evaluated our method for complex nonrigid motion using: (i) synthetic camera 2D projections generated from the 3D CMU Motion Capture (MoCap) dataset [2], and (ii) real-world 2D projections stemming from 2D point-tracks of videos in the Utrecht Multi-Person Motion (UMPM) benchmark [1], where the 3D ground-truth joints location is available through the mounted motion sensors.

On both datasets, we performed two sets of experiments: (i) *subspace clustering*: assessing the ability of our approach to build the union of 3D subspaces, this is the key indicator of the accuracy of the estimated affinity matrix; (ii) *3D reconstruction*: qualitatively and quantitatively evaluation of our approach against Dai et al.' summation of subspaces approach on the ground truth.

Subspace Clustering: Subspace clustering splits the frames in a sequence into several subsets where the nonrigid shapes in each subset form a local subspace. Testing the clustering accuracy is the simplest way to test if our approach produces the union of subspace as the 3D ground truth.

To evaluate the robustness to camera motion, we first performed a synthetic experiment using the 10^{th} sequence of subject 86 from the CMU MoCap dataset. This is the longest sequence that simultaneously includes several different actions: walking, running, jumping and sitting and the continuous transition among the actions. We generated several 2D projections of the true 3D joint sequences using synthetic camera sequences under different moving speed. The synthetic camera is an orthographic camera with the y-axis points to the center of the moving object. The angle speed (the relative projection direction with respect to the subject) of the camera varies from $0.1°$/frame to $8.0°$/frame.

We compared our approach against the LRR subspace clustering results on the synthesized 2D sequences, and on the ground truth 3D clustering. In each case, the clustering results are obtained from the affinity matrix $\mathbf{Z} \in \mathbb{R}^{n \times n}$, where the element $z_{ij} = 0$ if the i^{th} and the j^{th} samples are not belong to the same subspace, and if $z_{ij} > 0$, the same subspace. The success method for clustering the motion subspace should produces the similar $\mathbf{Z}$ to the LLR on the true 3D shapes. To measure the clustering accuracy, we perform K-means clustering on the affinity matrix $\mathbf{Z}$ obtained by both LLR on 2D projections and our method on 2D projections to get the final cluster. We followed [8] to estimate the subspace number K (cluster number) using the soft-thresholding approach. By tuning K, each cluster/subspace

[2] http://mocap.cs.cmu.edu

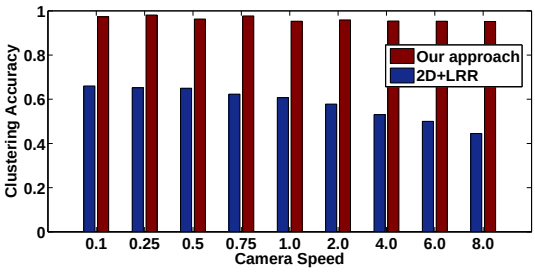


Figure 3. Frame-based clustering accuracies on the same sequence (the 10^{th} sequence of subject 86 in CMU Mocap database) under different camera angle speeds:$0.1°$/frame - $8.0°$/frame. The blue bars are direct clustering on the projected 2D data. The red bars are results of our method, which is robust to camera motion as the direct 3D clustering(**Best viewed in color**).

contain different granularity of the actions. However, the effect of K on the 3D reconstruction accuracy is not noticeable. The clustering accuracy is computed as the ratio of correct clustered nonrigid shapes (use the 3D clustering results as ground truth). The higher the better subspace clustering performance.

Fig. 3 shows the variation of clustering accuracy for 2D LRR clustering and our method versus the camera motion speeds. Observe that in Fig. 3, the clustering accuracy of the 2D clustering (blue bars in Fig. 3) not only is lower than the our method on clustering (blue bars in Fig. 3), but also decreases when the camera speed increases. In other word, the 2D subspace clustering is very sensitive to camera motion, more involved camera motion leads to less accurate 2D subspace clustering. Our method, by building the 3D clusters on the reconstructed 3D motion, (the red bars in Fig. 3) produces similar clustering results as the direct 3D LLR subspace clustering based solely on the 2D projected sequence.

The above evaluation is consistent with the results on real 2D sequences from video. As a simple qualitative illustration, we take the "p1_grab_3" video from the UMPM Benchmark Dataset. This video shows a person performs 3 actions around a table: walk, grab and stand. The data contains both the 2D joints from the video and and 3D joints sequence from the motion sensors. Similar to the synthetic experiment above, in Fig. 4, we compared the affinity matrix learned by our approach with the LLR subspace clustering on both 2D points (denoted by "2D Clustering"), and 3D joints (denoted by "3D Clustering"). Fig. 4 (a) shows the affinity matrix and the temporal relation of the frames by "3D Clustering" (ground truth). The "2D Clustering" (b) produce poor cluster assignment of the frames in comparison to "3D Clustering". Our approach (c) achieves similar results to the 3D subspace clustering, but only employs the 2D point projections during clustering.

For quantitative explantation, we used all complex

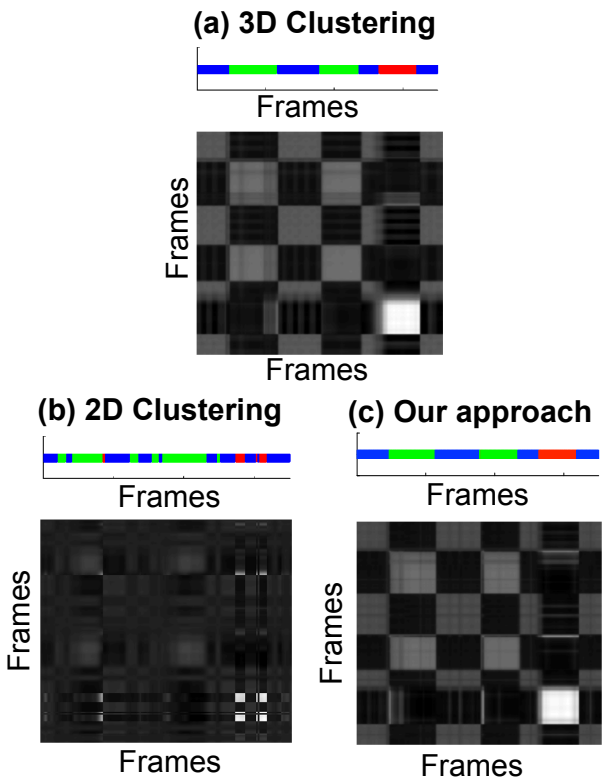


Figure 4. Comparison of human poses clustering bar and affinity matrix on the human action sequence"p1_grab_3" from the UMPM database [1]. This sequence includes 3 actions: walk (blue), grab (green), stand (red). (a) shows the 3D subspace clustering bar and affinity matrix. (b) and (c) show the clustering bar and affinity matrix by applying LLR subspace clustering and our method to 2D data. Our approach (c) get similar results as the 3D subspace clustering (a) but only using 2D points tracked in the videos. **(Best viewed in color)**

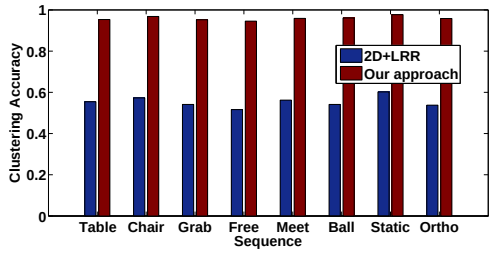


Figure 5. Clustering accuracy on the selected complex nonrigid motion sequences from UMPM database (using 3D clustering results as ground truth) [1]. The blue bars are direct clustering on the projected 2D data. The red bars are results of our method, which is similar to the direct 3D clustering(**Best viewed in color**).

nonrigid motion sequences from the UMPM benchmark dataset. Those sequences include 8 sequences ("p1_table_2", "p1_grab_3", "p1_chair_2", "p1_staticsyn_1", "p4_free_11", "p1_orthosyn_1", "p3_ball_1", "p4_meet_2"), which cover a large variety of daily human actions with strong changes in human poses. Fig 5 shows the quantitative results on clustering accuracy (3D LRR subspace clustering as ground truth) for all sequences. As the qualitative evaluation above, we use 3D LRR subspace clustering results as the ground truth to compute the clustering accuracy. For all sequences in Fig 5, the 2D LRR subspace clustering results (blue bars) is strongly different to the 3D LRR subspace clustering results due to the projection ambiguities and relative 3D rigid camera motion. Our approach (red bars) achieves similar clustering performance to 3D LLR subspace clustering for clustering nonrigid motion into local 3D subspaces solely on 2D

Remarks: An interesting observation from our approach shows that, the 3D shapes reconstructed and clustered in each subspace generally form meaningful action primitives. In Fig. 4 (c), the bar segments labeled as blue are walk, the green segments are grab, the red segment is stand. The impressive thing to note here, is that this unsupervised semantic segmentation is coming from 2D projected motion rather than raw 3D motion.

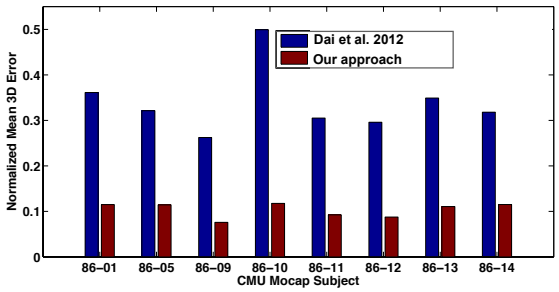
3D Reconstruction: Reconstructing 3D complex nonrigid motion from 2D projections is the final goal of our work. We evaluate the performance of our method on 3D complex nonrigid motion reconstruction against the Dai et al.'s method on two sets of data from CMU Mocap and UMPM benchmark quantitatively.

We first select 8 complex nonrigid motion sequences performed by subject #86 in CMU Mocap data. Each of these sequences consists of different types of nonrigid motions. As either 2D videos or projections are not available in the CMU Mocap data, we generated several 2D sequences by synthesizing a slow moving orthographic camera. The synthetic camera circled the body at the the camera angle speed $0.75\pi/sec$ with the y-axis of the camera pointing at the center of the body. At each time instance, 2D points are produced by projecting the original 3D nonrigid motion sequence to the synthetic camera.

This is a slow camera motion case. We adopted this slow camera motion for two reasons: Firstly, in most real world applications, the camera are slowly and smoothly moving; Secondly, 3D motion reconstruction from 2D projection generated by slow moving camera is very challenging for the very limited view. Reviewers are encouraged to read the literature [10, 12] for more details on explanation of poor reconstruction in slow moving camera case.

Fig. 6 shows the normalized 3D reconstruction error of Dai et al.'s method and our method 6 on the synthetic 2D data. As expected, our method performs inherently better than Dai et al.'s approach for those complex nonrigid motion sequences by employing a union of subspaces.

The above evaluation is also consistent with the results on real world 2D sequences. We test our method against Dai et al.'s approach on the 2D sequences in real world



(a)

Figure 6. Comparison on the normalized 3D reconstruction error by single subspace approach (Dai et al. 2012) and our proposed union of subspace approach on the synthetic 2D data from CMU Mocap data. We generated the 2D points by a slow moving orthographic camera, the 3D points are reconstructed by Dai et al.'s approach and our approach with assuming known the ground truth camera projection matrix. The 3D reconstruction error is normalized by the 3D ground truth data (**Best viewed in color**).

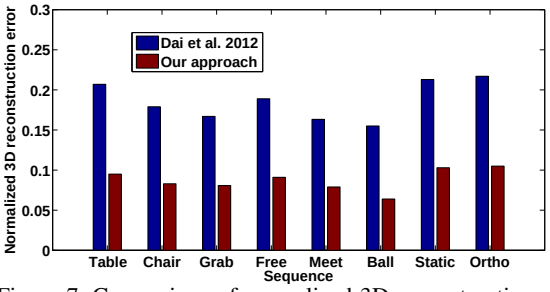


Figure 7. Comparison of normalized 3D reconstruction error between our approach (red bars) and Dai et al. [5] (blue bars) on complex nonrigid motion sequences which has several human actions and interactions with objects from the UMPM database [1].**(Best viewed in color)**

data: UMPM Benchmark. We use the same sequences as the subspace clustering experiments from the real world data set UMPM benchmark [1]. There are 8 sequences ("p1_table_2", "p1_grab_3", "p1_chair_2", "p1_staticsyn_1", "p4_free_11", "p1_orthosyn_1", "p3_ball_1", "p4_meet_2") selected to evaluate our method. Those sequences are considered as complex nonrigid motion for containing multiple daily human actions and human object interactions with strong changes in poses/shapes. The camera relative motion in those real world sequence are all slow and smooth, which is more challenging for NRSfM methods. Fig. 7 shows the comparison of our method against Dai et al.'s approach on reconstructing those selected complex nonrigid motion sequences. As expected, our method performs substantially better than Dai et al.'s approach for reconstructing those complex nonrigid motion sequences with slow camera motion.

To illustrate the performance of our method against Dai et al.'s method, we visualize the 3D reconstruction and subspace clustering results in Fig. 8("p1_chair_2" and "p1_table_2"). For both sequences, we visualized the skeletons in the sub-sequences (a) and (b) respectfully. The colors of the skeletons are correspond to the same cluster in the clustering bars. The ground-truth 3D skeletons are in gray, and respectively overlapped with the reconstructed 3D skeletons (cyan) using Dai et al. and our method. As shown in Fig. 8, "2D Clustering" is heavily affected by relative camera motion of the same human action. Our approach, however, not only reconstructs the most accurate 3D nonrigid motion but also achieves camera-view-invariant 3D subspace clustering similar to the subspace clustering on ground truth 3D data (See more video visualizations in the supplementary materials). It is also interesting to notice in Fig. 8 that different actions are clustered into local subspaces in 3D. These results could potentially be applied to unsupervised 3D action segmentation using 2D projections in the future work.

Discussion:

Similar to the previous work on NRSfM, our method is also limited by the reconstructability problem [10]. When the relatively motion between camera and human body is extremely small, Eq. 6 is an ill-posed problem. In those cases, robust 3D reconstruction requires more prior information, such as articulated length, connections, pre-learned shape basis or trajectory basis.

6. Conclusion

We proposed a novel strategy for simultaneously performing estimating the 3D structure, and subspace/cluster affinity matrix from an ensemble of 2D projections. We have motivated our approach based on the insight that complex nonrigid motion, such as sequences contains different types of human actions, can be modeled as a union of subspaces. We demonstrated the superiority of our approach in comparison to Dai et al.'s state of the art NRSfM approach for reconstructing complex nonrigid motion on both synthetic data and real-world data.

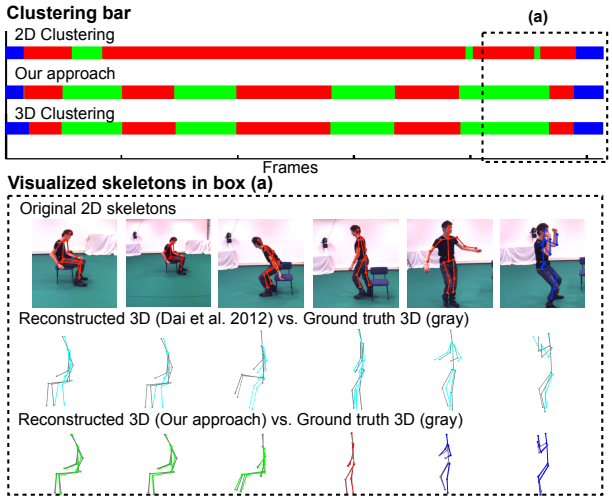
References

[1] N. Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp. Utrecht multi-person motion (umpm) benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *ICCV Workshop HICV*, 2011. 1, 5, 6, 7

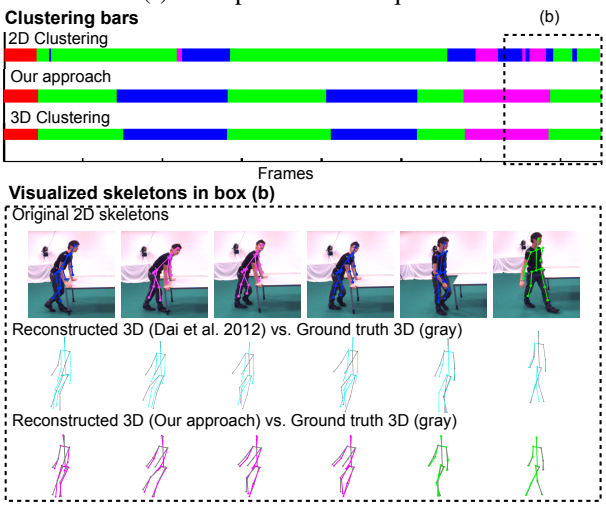
[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1), 2011. 2, 4

[3] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000. 2

[4] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. *J. ACM*, 58(3):1–37, 2011. 1, 2



(i) The "p1_chair_2" sequence



(ii) The "p1_table_2" sequence

Figure 8. Visualization of the 3D reconstruction and clustering results on two sequences: Sequence (i) shows one person interacts with the table (walk, sit, stand, raise hands). Sequence (ii) shows one person interacts with a chair (walk, raise hand, sit). The clustering bars are show for both two sequences. Different colors denote different clusters. For each of the two sequences, we show several 2D image sequences and reconstructed 3D skeletons in the selected box (a) and (b) in the bars. Different color of the skeletons are labeled by 2D clustering and our approach respectively. It is interesting to note that the different actions are clustered into different 3D subspaces by our method and 3D LRR subspace clustering. This result encourages to explore unsupervised 3D action segmentation on using 2D projected points in the future work. (**Best viewed in color**)

[5] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. In *CVPR*, 2012. 1, 2, 7

[6] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm: analysis and implementation. 24(7):881–892, 2002. 3

[7] M. Lee, J. Cho, C.-H. Choi, and S. Oh. Procrustean normal distribution for non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1280–1287, June 2013. 2

[8] G. Liu, Z. Lin, S. Yan, J. Sun, Y., and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. PAMI*, 35(1), 2013. 5

[9] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. PAMI*, 35(1):171–184, 2013. 2, 3, 4

[10] 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. 6, 7

[11] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9:137–154, November 1992. 2

[12] Y. Zhu, M. Cox, and S. Lucey. 3d motion reconstruction for real-world camera, 2012. 6