Convolutional Sparse Coding for Trajectory Reconstruction

Yingying Zhu, *Student Member, IEEE,* Simon Lucey, *Member, IEEE*

Abstract—Trajectory basis Non-Rigid Structure from Motion (NRSfM) refers to the process of reconstructing the 3D trajectory of each point of a non-rigid object from just their 2D projected trajectories. Reconstruction relies on two factors: (i) the condition of the composed camera & trajectory basis matrix, and (ii) whether the trajectory basis has enough degrees of freedom to model the 3D point trajectory. These two factors are inherently conflicting. Employing a trajectory basis with small capacity has the positive characteristic of reducing the likelihood of an ill-conditioned system (when composed with the camera) during reconstruction. However, this has the negative characteristic of increasing the likelihood that the basis will not be able to fully model the object's "true" 3D point trajectories. In this paper we draw upon a well known result centering around the Reduced Isometry Property (RIP) condition for sparse signal reconstruction. RIP allow us to relax the requirement that the full trajectory basis composed with the camera matrix must be well conditioned. Further, we propose a strategy for learning an over-complete basis using convolutional sparse coding from naturally occurring point trajectory corpora to increase the likelihood that the RIP condition holds for a broad class of point trajectories and camera motions. Finally, we propose an ℓ_1 inspired objective for trajectory reconstruction that is able to "adaptively" select the smallest sub-matrix from an over-complete trajectory basis that balances (i) and (ii). We present more practical 3D reconstruction results compared to current state of the art in trajectory basis NRSfM.

Index Terms—Nonrigid Structure From Motion, Convolutional Sparse Coding, ℓ_0 Norm, ℓ_1 Norm, Reconstructability.

1 Introduction

Non-rigid Structure from Motion (NRSfM) refers to the task of recovering the time varying 3D coordinates of each point on a deforming object from just their 2D projections. One prevalent approach for solving this problem is to seek the 3D reconstruction with the most compact linear shape basis that still satisfies the 2D projections [3]. This strategy works well for short simple sequences containing a single action/event. However, it has inherent problems when dealing with real-world complex motion sequences (e.g. a person walking, jumping, sitting and dancing all in the same sequence) [1]. In these instances the strategy of seeking the most compact linear shape is not sufficient to ensure an accurate reconstruction.

Recently, Akhter et al. [1] proposed a strategy to seek the 3D reconstruction with the most compact temporal basis. This approach, which we shall refer to herein as Trajectory Basis NRSfM, has two advantages over conventional NRSfM. First, it can handle long complex motion sequences since each point is being modeled independently. Second, since the trajectory basis is modeling the natural smoothness of most 3D motions occurring in the real-world it can be assumed to be object agnostic and known before reconstruction.

It has been pointed out, however, [16], [24] that trajectory basis NRSfM suffers from two strongly opposing requirements: (i) that the camera matrix composed with the trajectory basis matrix must be well-conditioned, and (ii) that the trajectory basis must have enough degrees of freedom to model the "true" 3D trajectory of the point.

Balancing the requirements of (i) and (ii) has turned out not to be easy. In early work, a discrete cosine transform (DCT) basis was advocated for trajectory reconstruction where the capacity of the trajectory basis was controlled by the choice of the highest order harmonic K. A drawback to this strategy, however, is the sensitivity of the reconstruction to the correct selection of K in order to balance (i) and (ii). Due to this sensitivity, a new K must be selected for each new 2D projected trajectory sequence that needs to be reconstructed. Recently, Valmadre & Lucey [24] have proposed a practical measure of reconstructability (motivated by earlier work by Park & Sheikh [15]) that allows one to provide an "adaptive" estimate of K given only the 2D projection trajectories. Specifically, this approach gives an upper bound on the largest K one can select that still allows for a well conditioned system for solving the 3D point trajectories. Unfortunately, for the common practical scenario of a slowly moving camera the K that allows for a well conditioned system is often too small to model the "true" 3D trajectory of the points.

1.1 Contributions

In this paper we explore the notion that, under certain circumstances, a unique solution to 3D point trajectories can be estimated even if the composed camera and trajectory basis matrix is ill-conditioned. We borrow upon

- *Y. Zhu is with the University of Queensland (UQ) and the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia.*
E-mail: zhuyingying2@gmail.com
- *S. Lucey is a Principal Research Scientist at the Commonwealth Scientific and Industrial Research Organization (CSIRO). He is also an adjunct professor at both the University of Queensland and the Queensland University of Technology.*
E-mail: simon.lucey@csiro.au

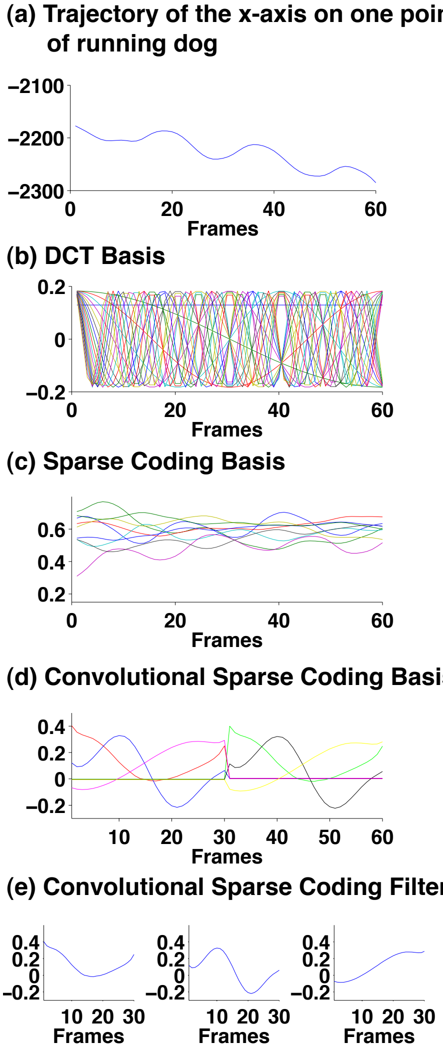


Fig. 1. The top row (a) depicts a single point x-coordinate trajectory of a 3D moving canine. The second row (b) depicts the set of fixed size DCT basis vectors that are required to reconstruct the trajectory. The third row (c) depicts the set of fixed size sparse coding basis vectors that are required to reconstruct the trajectory. The fourth row (d) depicts the set of basis vectors generated from the set of convolutional sparse filters. The final row (e) depicts the individual convolutional sparse filters. The sparse coding bases and convolutional sparse coding filters are learned independently on the CMU motion capture dataset including many different subjects of human motion (**Better viewed in color**).

the well understood Restricted Isometry Property (RIP) [7] [10] from sparse signal reconstruction literature. We propose a set of conditions based on the RIP for estimating a unique and exact reconstruction of a given 3D point trajectory. Specifically, if the trajectory basis coefficients are $K-$sparse and that all $2K$ sub-matrices within the composed camera and trajectory basis matrix are well conditioned then a unique solution to the trajectory can be found. We further demonstrate based on well known results in sparse signal reconstruction [7] [10] that a convex ℓ_1 objective can be employed to simultaneously estimate: (i) how $K-$sparse a given 3D trajectory is based solely on the 2D projection trajectory, and (ii) the non-zero trajectory basis coefficients for 3D reconstruction.

An advantage of the DCT trajectory basis in previous trajectory basis NRSfM work, is that it is simple to cater for varying length trajectories since it is based on a pre-defined mathematical form. Unfortunately, the DCT basis is not suitable for sparse trajectory reconstruction. To circumvent this limitation, we advocate the use of convolutional sparse coding to learn an over-complete trajectory basis from offline 3D trajectory observations (see Figure 1). This approach has two advantages. First, it is able to compactly represent a large space of commonly encountered 3D trajectories. Second, since we are learning convolutional filters they are easily generalizable to varying length trajectories. Finally, we demonstrate impressive reconstruction results compared to previous state of the art.

1.2 Related work

Factorization approaches, first proposed for recovering rigid 3D structure by Tomasi and Kanade in [19], were extended by Bregler et al. to handle non-rigidity in [3]. The motivation in Bregler et al.'s work was to seek the 3D reconstruction with the most compact (i.e. low rank) shape basis that satisfies the 2D point projections. Further, work by Torresani et al. adopted the low-rank constraint to assist in non-rigid tracking [22], incorporated temporal constraints by modeling the shape basis coefficients as a linear dynamical system [20] and modeled the distribution of non-rigid deformation by a hierarchical prior [21]. Bartoli et al. [2] used a temporal smoothness prior to reduce the sensitivity of their solution to the number of shape bases. Rabaud and Belongie [17] shifted away from the linear basis interpretation, proposing to learn a smooth manifold of shape configurations from video. They incorporated temporal regularization to prevent the camera and structure from changing excessively between frames.

A number of approaches that develop the use of a shape basis have been subsequently proposed, including [20], [21], [27]. A fundamental criticism, however, of all these approaches is the compactness of shape basis. For example, the linear shape basis of a "person walking" differs substantially from a "person dancing" or "person jumping", etc. For complex motion sequence which includes different actions: walk, dance, walk, etc, a linear shape basis lacks the ability to model these complicated nonrigid motions compactly.

To handle complicated nonrigid motions, Akhter et al. [1] proposed that the trajectory of each point could instead be restricted to a low-dimensional subspace typically a Discrete Cosine Transform (DCT). This approach, which is of central interest in this paper, is often referred to in literature as trajectory basis NRSfM. The key advantages of this approach over shape basis methods are: (i) one no longer requires a compact shape basis (allowing for the reconstruction of complex sequences), and (ii) the trajectory basis is typically object agnostic (allowing for a fixed trajectory basis) enabling a simpler reconstruction strategy that does not rely on a SVD style factorization. A drawback to Akhter et al.'s approach, was that the trajectory basis needed to have a pre-defined size. More recently, Chen et. al. proposed a strategy that employed an ℓ_1 norm to automatically select the active DCT basis size [8].

Gotardo and Martinez [13] recently combined shape and trajectory basis approaches, describing the shape basis coefficients with a DCT basis over time. In their subsequent work [12], they extended this approach to include non-linear shape models using kernels.

Park et al. [16] examined the limitations of trajectory basis NRSfM in solving for structure given known cameras. Specifically, Park et al. attempted to characterize theoretically what projected 2D trajectories can and cannot be successfully reconstructed. They referred to this theoretical measure as "reconstructability". In subsequent work, Valmadre and Lucey [24] refined this measure so that it considers the condition of the resulting system of equations. Zhu and Lucey [29] proposed to employ a ℓ_1 penalty to minimize the size of active trajectory basis and demonstrated that the reconstructability restriction decreased by using ℓ_1 penalty empirically.

2 Problem Formulation

One point $\mathbf{x}_t \in \mathbb{R}^3$ imaged as $\mathbf{w}_t \in \mathbb{R}^2$ by a pinhole camera $\mathbf{P}_t$ for all $t \in \{1, \ldots, F\}$,

$$\begin{bmatrix}\mathbf{w}_t \\ 1\end{bmatrix} \simeq \mathbf{P}_t \begin{bmatrix}\mathbf{x}_t \\ 1\end{bmatrix} \tag{1}$$

Partitioning the projection matrix,

$$\mathbf{P}_t = \begin{bmatrix}\mathbf{R}_t & \mathbf{d_t} \\ \mathbf{c_t}^T & b_t\end{bmatrix} \tag{2}$$

The projective equality in Equation 1 yields the under-determined 2×3 system of linear equations.

$$\mathbf{Q}_t \mathbf{x}_t = \mathbf{u}_t, \tag{3}$$

where $\mathbf{Q}_t = \mathbf{R}_t - \mathbf{w}_t \mathbf{c_t}^T$ and $\mathbf{u}_t = b_t \mathbf{w}_t - \mathbf{d_t}$. Each $\mathbf{Q}_t$ matrix has a 1D right nullspace corresponding to the ray connecting the camera center and the projection on the image plane. When $\mathbf{P}_t$ represents an affine camera,

$$\mathbf{P}_t = \begin{bmatrix}\mathbf{R}_t & \mathbf{d_t} \\ \mathbf{0} & 1\end{bmatrix} \quad \Rightarrow \quad \mathbf{R}_t \mathbf{x}_t = \mathbf{w}_t - \mathbf{d_t}. \tag{4}$$

3 Trajectory Reconstruction

Given a full subspace $\mathbf{\Phi} = [\phi_1, \ldots, \phi_F]$ where $\phi_f \in \mathbb{R}^F$, Park et al. [1], [16] constrained individual point trajectories to lie on a low dimensional subspace $\mathbf{\Phi}_{\mathcal{S}} = [\phi_{\mathcal{S}(1)}, \ldots, \phi_{\mathcal{S}(K)}]$ where $\mathcal{S}$ is a subset of $K < F$ indices that are most energy preserving such that,

$$\mathbf{X} \approx \mathbf{\Phi}_{\mathcal{S}} \boldsymbol{B}_{\mathcal{S}} \ . \tag{5}$$

Let $\mathbf{X} \in \mathbb{R}^{F \times 3}$ be a matrix of F concatenated 3D positions of a single continuously sampled point, and $\boldsymbol{B}_{\mathcal{S}} \in \mathbb{R}^{K \times 3}$ be the resultant energy preserving compact representation. For compactness, we shall herein represent Equation 5 in vectorized form,

$$\mathbf{x} \approx \mathbf{\Theta}_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}} \tag{6}$$

where $\mathbf{x} = \text{vec}(\mathbf{X})$, $\boldsymbol{\beta}_{\mathcal{S}} = \text{vec}(\boldsymbol{B}_{\mathcal{S}})$ and $\mathbf{\Theta}_{\mathcal{S}} = \mathbf{\Phi}_{\mathcal{S}} \otimes \mathbf{I}_3$ [1].

In practice one can only observe the 2D projection trajectory of the 3D point,

$$\mathbf{Qx} = \mathbf{u} \tag{7}$$

where

$$\mathbf{Q} = \begin{bmatrix}\mathbf{Q}_1 & & \\ & \ddots & \\ & & \mathbf{Q}_F\end{bmatrix}, \mathbf{u} = \begin{bmatrix}\mathbf{u}_1 \\ \vdots \\ \mathbf{u}_F\end{bmatrix}. \tag{8}$$

As noted by Akhter et al. if $3K \leq 2F$ (i.e. more observations than unknowns) one can attempt to find the least-squares estimate of $\boldsymbol{\beta}_{\mathcal{S}}$,

$$\tilde{\boldsymbol{\beta}}_{\mathcal{S}} = \arg\min_{\boldsymbol{\beta}_{\mathcal{S}}} \|\mathbf{Q}\mathbf{\Theta}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \mathbf{u}\|_2^2 \ . \tag{9}$$

Or alternatively,

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{\mathcal{S}} = \quad & \arg\min\nolimits_{\boldsymbol{\beta}_{\mathcal{S}}} \quad ||\boldsymbol{\beta}_{\mathcal{S}}||_2^2 \\ & \text{s.t.} \qquad \mathbf{u} = \mathbf{Q}\mathbf{\Theta}_{\mathcal{S}}\boldsymbol{\beta} \ .\end{aligned} \tag{10}$$

A least-squares estimate of the 3D trajectory $\tilde{\mathbf{x}}$ can then be obtained by applying Equation 6. An important realization, however, is the requirement that $3K \leq 2F$ is a necessary but not sufficient condition for accurate reconstruction.

3.1 Reconstructability Known Set

As implied in the previous section, one can have multiple solutions $\mathbf{x}'$ that satisfy the projection equation

$$\mathbf{u} = \mathbf{Qx}' \ . \tag{11}$$

One can represent this ambiguity in an alternate manner

$$\mathbf{x} = \mathbf{x}' + \mathbf{Q}_{\perp}\mathbf{z} \tag{12}$$

where $\mathbf{Q}_{\perp} \in \mathbb{R}^{F \times 3F}$ is a matrix whose columns are an orthonormal basis for null($\mathbf{Q}$) such that

$$\mathbf{QQ}_{\perp} = \mathbf{0}, \quad \mathbf{Q}_{\perp}^T\mathbf{Q}_{\perp} = \mathbf{I} \tag{13}$$

and $\mathbf{z} \in \mathbb{R}^F$. As proposed by Valmadre and Lucey [24] a least-squares estimate for $\mathbf{z}$ can be found using the trajectory basis $\mathbf{\Theta}_{\mathcal{S}}$ for any given $\mathbf{x}'$,

$$\tilde{\mathbf{z}}(\mathbf{x}') = \arg\min_{\mathbf{z}} ||\mathbf{x}' + \mathbf{Q}_{\perp}\mathbf{z}||_{\mathbf{M}_{\mathcal{S}}}^2 \text{ [2]} \tag{14}$$

where,$\mathbf{M}_{\mathcal{S}} = \mathbf{\Theta}_{\mathcal{S}\perp}\mathbf{\Theta}_{\mathcal{S}\perp}^T$ and $\mathbf{\Theta}_{\mathcal{S}\perp} \in \mathbb{R}^{F \times (F-K)}$ is the orthonormal basis for null($\mathbf{\Theta}_{\mathcal{S}}$). The least-squares estimate of $\mathbf{x}$, given the trajectory basis $\mathbf{\Theta}_{\mathcal{S}}$, for any given $\mathbf{x}'$ becomes

$$\tilde{\mathbf{x}}(\mathbf{x}') = \mathbf{x}' + \mathbf{Q}_{\perp}\mathbf{z}(\mathbf{x}') \ . \tag{15}$$

[1] We refer to $\mathbf{I}_3$ as an 3×3 identity matrix, and the operator $\otimes$ denotes a Kronecker product.

[2] We defined $||\mathbf{X}||_{\mathbf{M}}^2$ to represent $\mathbf{X}^T\mathbf{MX}$.

Considering the case where $\mathbf{x}'$ is the ground truth trajectory $\mathbf{x}$, we obtain an expression for the reconstruction error

$$||\mathbf{x} - \tilde{\mathbf{x}}(\mathbf{x})||_2^2 = ||(\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp)^{-1} \mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{x}||_2^2 \quad . \tag{16}$$

This facilitates the definition of an upper bound v on reconstruction error

$$v(\mathbf{x}, \mathbf{Q}, \mathbf{\Theta}, \mathcal{S}) = \underbrace{\text{cond}(\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp)}_{\text{gain } \gamma} \underbrace{\frac{||\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{x}||_2^2}{||\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp||_2^2}}_{\text{contradiction } \epsilon} \tag{17}$$

where

$$||\mathbf{x} - \tilde{\mathbf{x}}(\mathbf{x})||_2^2 \leq v(\mathbf{x}, \mathbf{Q}, \mathbf{\Theta}_\mathcal{S}) \quad . \tag{18}$$

In general terms, one can see from Equation 17 that if one has knowledge of the set $\mathcal{S}$ that can model the ground truth trajectory one only requires that the matrix $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ is well conditioned in order to obtain an exact reconstruction. Consequently, without knowledge of the optimal set no guarantees on the optimality of the solution can be made.

3.2 reconstructability Unknown Harmonic

An important assumption in trajectory basis NRSfM hitherto, has been a priori knowledge of what subset $\mathcal{S}$ of the basis $\mathbf{\Theta}$ is active. It is clear, however, from the previous section that a sub-optimal selection of this active set $\mathcal{S}$ can have dire consequences with respect to reconstructability. Choose a set $\mathcal{S}$ that is too expressive and one can obtain a poorly conditioned matrix $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ that reduces reconstructability. Conversely, choose a set that is too constrained and one affects how well one can represent the ground truth trajectory.

A popular choice for $\mathbf{\Theta}$ is an orthonormal harmonic basis, such as the discrete cosine transform (DCT) basis. For this type of basis, choosing the set $\mathcal{S}$ simplifies to choosing the integer scalar K referring to the highest harmonic such that $\mathcal{S} = \{1, \ldots, K\}$. Valmadre and Lucey [24] proposed a strategy of exhaustive search where one chose the largest K that satisfies a gain threshold γ based on Equation 17. The strategy here is that by choosing the largest possible K (contradiction) that ensures a well conditioned system (gain) one can effectively balance the opposing forces of contradiction and gain in Equation 17. As discussed in the previous section, no guarantees on the uniqueness of the solution from this strategy can be made (since we do not know the optimal set $\mathcal{S}$ a priori that can reconstruct the ground-truth trajectory). Empirical evidence [24] suggests, however, that this strategy performs well in practice.

3.3 Reconstructability Unknown Set

Inspecting Equation 17 it is clear one is not restricted to harmonic bases, like DCT, in order to obtain good reconstructability. In fact it is quite reasonable to entertain an over-complete basis $\mathbf{\Phi} \in \mathbb{R}^{F \times M}$ where $\mathbf{\Theta} = \mathbf{I}_3 \otimes \mathbf{\Phi} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{3M}]$ and $M > F$. A central motivation here is by removing orthonormality from the problem it may be possible to entertain more compact sets $\mathcal{S}$ with similar reconstruction properties. We shall herein denote $\mathbf{\Theta}_\mathcal{S} = [\boldsymbol{\theta}_{\mathcal{S}(1)}, \ldots, \boldsymbol{\theta}_{\mathcal{S}(K)}]$ as opposed to the earlier definition $\mathbf{\Theta}_\mathcal{S} = \mathbf{\Phi}_\mathcal{S} \otimes \mathbf{I}_3$. This was done to ensure greater flexibility when choosing a set $\mathcal{S}$. Further, K now reflects the cardinality of this more expressive set $\mathcal{S}$.

In principle, one could apply a similar strategy to the previous section, where one attempts to exhaustively search for the largest $\mathcal{S}$ that satisfies a gain threshold. There are two problems with this strategy. First, the sheer cost of the search. The search space is no longer over a single value K, but over all possible submatrices of $\mathbf{\Theta}$ a task which is NP-hard. Second, one is still not guaranteed an exact reconstruction unless one knows the optimal set $\mathcal{S}$.

3.3.1 *Uniqueness and Sparseness*

In this paper we propose an alternate strategy

$$\tilde{\boldsymbol{\beta}} = \arg \quad \min_{\boldsymbol{\beta}} \quad ||\boldsymbol{\beta}||_0 \qquad \text{s.t.} \quad \mathbf{u} = \boldsymbol{A}\boldsymbol{\beta} \quad . \tag{19}$$

where $\boldsymbol{A} = \mathbf{Q}\mathbf{\Theta}$. $||.||_0$ denotes the ℓ_0 "norm" which counts the number of non-sparse (i.e. non-zero) elements. Equation 19 can be interpreted as finding the most compact set $\mathcal{S}$ that satisfies $\mathbf{u} = \boldsymbol{A}\boldsymbol{\beta}$ (where the set $\mathcal{S}$ is defined as the indices of $\boldsymbol{\beta}$ that are non-sparse).

It is well understood in sparse signal reconstruction literature [7] that if $\boldsymbol{A}$ obeys the Restricted Isometry Property (RIP) and one knows the cardinality of $\mathcal{S}$ (i.e. the sparsity of $\boldsymbol{\beta}$) one can find an exact unique solution to $\boldsymbol{\beta}$.

Definition : For each integer $K = 1, 2, \ldots, 3M$ define the isometry constant σ_K of the matrix $\boldsymbol{A}$ as the smallest number such that

$$(1 - \sigma_K)\|\boldsymbol{\beta}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{\beta}\|_2^2 \leq (1 + \sigma_K)\|\boldsymbol{\beta}\|_2^2 \tag{20}$$

holds for all K-sparse vectors $||\boldsymbol{\beta}||_0 = K$. It has been proved [7] that if the isometry constant $\sigma_{2K} < 1$ the K-sparse solution of vector $\boldsymbol{\beta}$ is unique. In laymen terms RIP implies that if *all* $2F \times 2K$ sub-matrices of $\boldsymbol{A}$ are well conditioned then a unique solution for $\boldsymbol{\beta}$ can be found if it is K-sparse. A strength of this strategy is that one only needs to know the cardinality of the set $\mathcal{S}$, as opposed to the actual set itself in the canonical ℓ_2 approach, in order to obtain an exact reconstruction of the trajectory.

3.3.2 *Convex Relaxation*

A drawback to Equation 19 is that like Valmadre and Lucey's approach the computational cost of the solution is NP-hard. Fortunately, a convex relaxation on Equation 19 can be obtained with a ℓ_1 norm replacement. The employment of a ℓ_1 norm to encourage sparsity in estimation problems (e.g., LASSO regression, compressed sensing, etc.) is quite common across many areas of computer science and can be found efficiently using a variety of different packages. Readers are encouraged to inspect [6], [26] on the employment of ℓ_1 norms for encouraging sparseness

$$\tilde{\boldsymbol{\beta}} = \arg \quad \min_{\boldsymbol{\beta}} \quad ||\boldsymbol{\beta}||_1 \qquad \text{s.t.} \quad \mathbf{u} = \boldsymbol{A}\boldsymbol{\beta} \quad . \tag{21}$$

It can be shown [7] that if the restricted isometry constant satisfies

$$\sigma_{2K} < \sqrt{2} - 1 \tag{22}$$

the solution of the ℓ_1 norm is equal to ℓ_0 "norm" solution.

3.3.3 Quality of Reconstruction

At first glance, it seems like the ℓ_1 approach shares a similar drawback to the canonical ℓ_2 approach, in that it is difficult to know anything about the optimal set $\mathcal{S}$ (either the cardinality or the set itself) a priori. Fortunately, through the use of the ℓ_1 strategy we obtain: (i) an estimate of trajectory coefficients $\tilde{\boldsymbol{\beta}}$, and (ii) an estimate of the set $\tilde{\mathcal{S}}$ in polynomial time. Given that the equality constraint $\mathbf{u} = \boldsymbol{A}\boldsymbol{\beta}$ is satisfied, then all that is required to see if $\tilde{\boldsymbol{\beta}}$ is an exact reconstruction is to check that $\boldsymbol{A}$ satisfies the RIP for the cardinality of the estimated set.

Unfortunately, checking that $\boldsymbol{A}$ satisfies RIP is itself an NP-hard task. A common tractable metric for gauging the RIP of a matrix $\boldsymbol{A}$ is mutual coherence. It is defined, assuming the columns of $\boldsymbol{A}$ are normalized to unit ℓ_2 norm, in terms of the Gram matrix $\mathbf{G} = \boldsymbol{A}^T\boldsymbol{A}$. With $\mathbf{G}(k,j)$ denoting entries of this matrix, the mutual coherence is

$$\mu(\boldsymbol{A}) = \max_{1 \le k,j,M, k \ne j} |\mathbf{G}(k,j)| \tag{23}$$

The matrix $\boldsymbol{A}$ is deemed incoherent if $\mu(\boldsymbol{A})$ is small. In general the more incoherent the matrix, the more likely it is likely to satisfy the RIP. Results in [11] show that if there exists a representation $\mathbf{u} = \boldsymbol{A}\boldsymbol{\beta}$ with sparsity $K = ||\boldsymbol{\beta}||_0$, and K does not exceed a threshold $(1 + \mu^{-1})/2$ then this is the unique sparsest representation. Stricter bounds have been proposed by Candies et al. [5] the details of which, however, are outside the scope of this paper.

As discussed in [6] another advantage of the RIP is that if the true $\boldsymbol{\beta}$ is not K-sparse, but the estimate $\tilde{\boldsymbol{\beta}}$ is K-sparse then the quality of the reconstruction is as good as if one knew ahead of time the optimal set $\mathcal{S}$ of cardinality K for reconstruction. As a result, one can be assured of a graceful degradation even if $\boldsymbol{A}$ does not satisfy RIP. One can see empirical evidence of this in Section 5.

3.3.4 Noise

In practice a given 2D measurements $\mathbf{u}$ may be imperfect, so an assumption of noise must be made when attempting to use an ℓ_1 strategy

$$\begin{aligned} \tilde{\boldsymbol{\beta}} = \arg \quad & \min_{\boldsymbol{\beta}} \quad ||\boldsymbol{\beta}||_1 \\ & \text{s.t.} \quad ||\mathbf{u} - \boldsymbol{A}\boldsymbol{\beta}||_2^2 \le \epsilon \ . \end{aligned} \tag{24}$$

where ϵ bounds the amount of noise in the data. This problem is often referred to as LASSO, and like Equation 21 is convex and can be solved efficiently.

4 Learning the Sparse Trajectory Basis

In this section we present our strategy for learning the over-complete basis $\boldsymbol{\Phi}$. Specifically, we attempt to learn a $\boldsymbol{\Phi}$ that gives the "sparsest" representation of an ensemble of trajectories in order to realize the RIP ℓ_1 condition (Equation 22) across as many camera matrices as possible. Many strategies/approaches have been put forward previously in literature to solve this problem as its solution is applicable to broad class of applications in computer science (e.g., compressed sensing, classification, etc.). A popular strategy [14] is to solve the following objective

$$\begin{aligned} & \arg\min_{\boldsymbol{\Phi},\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{x}_n - \boldsymbol{\Phi}\boldsymbol{\beta}_n||_2^2 + \sum_{n=1}^{N} \lambda ||\boldsymbol{\beta}_n||_1 \\ & \text{s.t.} \quad ||\boldsymbol{\phi}_m||_2 \le 1 \quad m = 1, \ldots, M \end{aligned} \tag{25}$$

where λ is the ℓ_1 penalty and $\mathbf{x}_n \in \mathbb{R}^{F \times 1}$ is a 1D trajectory sampled from all axes x, y and z. $\boldsymbol{\beta}$ is a supervector of the sparse code vectors $\{\boldsymbol{\beta}_n\}_{n=1}^N$ and $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M]$ is the over-complete basis. N is the number of samples in the training set. An alternation strategy was proposed to minimize this objective [14] and exhibited good performance in our experiments.

4.1 Convolutional Sparse Trajectory Basis Learning

Sparse coding has two fundamental drawback however, as: (i) it assumes the ensemble of input trajectory vectors $\{\mathbf{x}_n\}_{n=1}^N$ are of a fixed length, and (ii) the resultant sparse coding basis may be highly redundant as it learns shifted versions of the same trajectory.

Convolutional spare coding offers a natural mechanism to overcome these drawbacks. Zeiler et al. [28] proposed a sparse coding objective taking convolution directly into account

$$\begin{aligned} \arg\min_{\mathbf{h}_l,\boldsymbol{\beta}} \quad & \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{x}_n - \sum_{l=1}^{L} \mathbf{h}_l * \boldsymbol{\beta}_{l,n}||_2^2 + \\ & \lambda \sum_{n=1}^{N}\sum_{l=1}^{L} ||\boldsymbol{\beta}_{l,n}||_1 \\ \text{subject to} \quad & ||\mathbf{h}_l||_2^2 \le 1 \text{ for } l = 1 \ldots L. \end{aligned} \tag{26}$$

Now $\boldsymbol{\beta}_{l,n}$ takes the role of a sparse feature map which, when convolved with a filter $\mathbf{h}_l$ and summed over all l, should approximate the n-th input signal $\mathbf{x}_n$. Unlike traditional sparse coding the estimated sparse filters $\{\mathbf{h}_l\}_{l=1}^L$ will be of a fixed spatial support and the input signal $\mathbf{x}_n$ and the sparse feature maps $\boldsymbol{\beta}_n = \{\boldsymbol{\beta}_{l,n}\}_{l=1}^L$ are of a different and usually much larger dimensionality. Another fortunate aspect of the objective in Equation 26 is that it can handle varying length training examples $\{\mathbf{x}_n\}_{n=1}^N$.

4.1.1 Convolutional Sparse Trajectory Basis

It is trivial to obtain an over-complete basis $\boldsymbol{\Phi}$ from a set of filters $\{\mathbf{h}_l\}_{l=1}^L$. Specifically, one can form the basis

$$\boldsymbol{\Phi}\boldsymbol{\beta} = \sum_{l=1}^{L} \mathbf{h}_l * \boldsymbol{\beta}_l \tag{27}$$

where $\boldsymbol{\Phi}$ is a concatenation of convolutional Toeplitz matrices corresponding to each sparse filter and $\boldsymbol{\beta}$ =

$[\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_L^T]^T$. To keep consistence with the notation in previous sections (Equation 21) $M = LF$ where $\boldsymbol{\Phi} \in \mathbb{R}^{F \times M}$, F is the number of samples in the trajectory and L is the number of filters learned through convolutional sparse coding.

In practice constructing the full matrix $\boldsymbol{\Phi}$ can be computationally taxing, especially for long signals. In these instances it is often more computationally tractable to solve the ℓ_1 trajectory reconstruction problem stated in Equation 21 in the following manner

$$\begin{aligned} \tilde{\boldsymbol{\beta}} = \quad & \arg\min_{\boldsymbol{\beta}} \quad ||\boldsymbol{\beta}||_1 \\ \text{s.t.} \qquad & \mathbf{u} = \mathbf{Q} \sum_{l=1}^{L} \mathbf{h}_l * \boldsymbol{\beta}_l \ . \end{aligned} \tag{28}$$

This objective can be solved efficiently through the use of an iterative Augmented Lagrangian Method (ALM) that allows one to efficiently solve for $\boldsymbol{\beta}$ in the Fourier domain. More details on this approach can be found in [4].

5 Experiments

For all our trajectory basis learning experiments we employed the widely used CMU Motion Capture dataset[3] as training dataset. This dataset contains 3D trajectory point information covering a large variety of human actions.

To evaluate the generalization properties of our learned bases, we evaluate their reconstruction performance on 3D sequences of a moving canine [4] and the "real-world" UMPM benchmark which has 2D tracked points of moving humans in video and calibrated 3D ground truth [25]. The CMU Motion Capture dataset was resampled from 120 fps (frames per second) to 30 fps to ensure consistency between training and testing sequences. A cross-validation procedure was used to find the best filter size for the convolutional sparse coding reconstructions.

To encourage generalization, we learned bases in a point and coordinate independent manner (i.e., a single basis was learnt across all points and across $x-$, $y-$ and $z-$ coordinate systems). We selected a subset of sequences which includes a large variety of daily human motions from CMU Motion Capture dataset as our training data. These sequences which spans diverse motions are chosen from subject 1, 2, 3, 5, 9, 11, 15, 17, 21 24, 27, 28, 33, 41, 60, 118, 106, 122, 124, 126 in CMU Motion Capture dataset. The number of learned convolutional sparse coded filters in our experiment is 750. We found that the learned convolutional sparse coded filters are able to reconstruct large variety of nonrigid motions, such as different human actions, human-object interactions and animals motions.

5.1 Relation to Low-Rank Shape Methods

The central assumption of shape basis NRSfM is that the nonrigid shape lies in a single low rank shape subspace or shape basis [9]. Specifically, we argue that for complex motion sequences, such as those found in the UMPM benchmark dataset, it does not always make sense to enforce a low-rank assumption on the shape space. These complex sequences contain multiple actions (e.g. raise hands, stand, walk, and sit). Since different actions are dominated by different shapes, different actions tend to lie in different local shape subspaces. Relying on low-rank shape space assumption for these types of sequences will result in poor reconstruction performance.

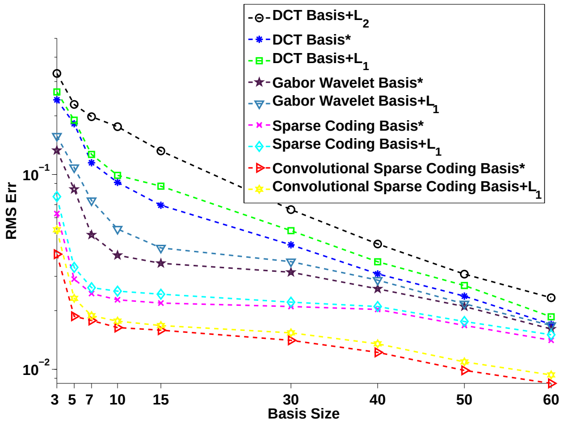


Fig. 2. Comparison of different trajectory basis in terms of normalized root mean square (RMS) error (Err) as a function of the number of active ($K(\mathcal{S})$) basis vectors. Reconstruction error was calculated on the actual 3D trajectories (not 2D projections). The sequence length is 150. Results show that a sparse coded basis can encode unseen 3D trajectory observations far more sparsely than conventional trajectory bases (DCT) (**Better viewed in color**).

This insight has been well argued in previous works on trajectory basis NRSfM [1]. Trajectory basis NRSfM methods reconstruct the trajectory of each nonrigid point in a shape independent way. The performance of trajectory basis methods is independent from the shape, although dependent on camera motion. The focus in this paper is on how to make trajectory basis NRSfM methods less sensitive to this camera motion while enjoying the useful shape independence assumption afforded by the approach. As a result we deemed a direct empirical comparison with shape basis NRSfM algorithms unnecessary as it distracts from the central focus of the paper.

5.2 Synthetic Experiment

We used 3 canine sequences which shows a dog running, jumping and turning around from the free Motion Capture dataset for our synthetic experiments. The canine sequences contain 36 points. We synthesized 2D points by projecting the original 3D motion sequence using a moving synthetic orthographic camera. The synthetic camera circled the nonrigid 3D object and was at all times pointing at the object's center.

5.2.1 Compressibility Comparison

In this experiment we wanted to evaluate how well various trajectory bases were able to compress 3D motion trajec-

[3] More details on the CMU Motion Capture can be found at http://mocap.cs.cmu.edu.

[4] More details on this canine sequence can be found at http://motioncapturedata.com/2009/05/animal-motion-capture-dog.html

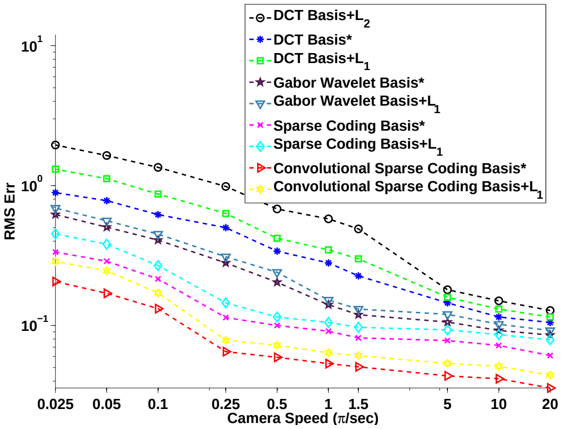


Fig. 3. Reconstruction (normalized RMS error) versus different camera angle speeds for the canine sequences. 2D projections were generated using a synthetic orthographic camera with the y-axis pointed to the center of the object with the angle of rotation speed being varied (**Better viewed in color**).

tories. Figure 2 depicts a comparison of the normalized root mean square (RMS) error (Err) [5] of 3D motion reconstruction on different trajectory bases as a function of the active basis size (i.e. the cardinality of $\mathcal{S}$). The sequence length used in this experiment is 150. We first applied an ℓ_2 objective for the DCT basis, and the number of harmonics dictating the basis size (denoted as DCT Basis+L_2). Then, for DCT, the learned sparse and convolutional sparse coding bases, an ℓ_1 LASSO objective was employed. The cardinality of $\mathcal{S}$ (i.e. the basis size) was controlled by adjusting the ℓ_1 penalty. $DCT + L_1$ is representative of the method proposed in [8]. It has been demonstrated that sparse coding applied to localized areas of natural signals (e.g natural images) generate Gabor "like" bases [18]. For completeness we have also included reconstruction results using a Gabor wavelet basis.

We also compared the results trajectory reconstruction bases with prior knowledge of the oracle set $\mathcal{S}$. The oracle set was obtained by finding the optimal set $\mathcal{S}$ for each test trajectory for each set of bases. We do exhaustively search on the whole trajectory bases to find the optimal set $\mathcal{S}$ as the oracle set to reconstruct each test trajectory. All reconstructions in Figure 2 that are appended by the notation "*" are estimated using the four different bases set and the canonical ℓ_2 in Equation 9 given a fixed oracle set $\mathcal{S}$.

Results demonstrate that the sparse learned bases, in particular the convolutional sparse coded basis could encode trajectories far more compactly than the canonical DCT basis or Gabor wavelet basis.

5.2.2 Reconstruction Comparison

In this section we look at the problem of reconstructing 3D trajectories from 2D projections given a known camera matrix. In Figure 3 we present the normalized RMS error of 3D motion reconstruction against circular camera speed for the canine sequences. We generated a synthetic orthographic camera with the y-axis pointed to the center of the object with the angle of rotation speed being varied (which effects reconstructability). Reconstruction results were obtained using the DCT, Gabor wavelets, sparse coded and convolutional sparse coded bases.

We compared results trajectory reconstruction bases with and without prior knowledge of the oracle set $\mathcal{S}$. The oracle set was obtained by finding the optimal set $\mathcal{S}$ for each test trajectory for each basis (as previously discussed in Section 5.2.1 and Figure 2). All reconstructions in Figure 3 that are appended by the notation "*" are estimated using the canonical ℓ_2 objective in Equation 9 with a fixed known oracle set $\mathcal{S}$.

These results are compared to reconstruction results with no prior information of the oracle set $\mathcal{S}$. Instead, we use an ℓ_1 style LASSO objective across all four bases to simultaneously estimate the set $\tilde{\mathcal{S}}$ and the coefficient vector $\tilde{\boldsymbol{\beta}}$. As a point of reference we include the performance for the DCT basis with a fixed set $\mathcal{S}$ that has been tuned to work well across all trajectories. The results shows that for the fast moving cameras, all methods obtain reasonable reconstruction performance. However, as the camera slows and reconstructability becomes poor there is a marked difference between strategies. As proposed earlier, having knowledge of the oracle set $\mathcal{S}$ gives superior performance compared to the ℓ_1 LASSO objective across all camera speeds.

Impressively, the sparse coding bases outperform the canonical DCT bases with and without prior knowledge of the oracle set. As expected the convolutional sparse coding basis gives superior performance to all other bases with and without knowledge of the oracle set. As expected sparse coding and convolutional sparse coding bases outperformed the DCT and Gabor wavelet bases. It is interesting to note that the Gabor wavelet basis outperformed the DCT basis across all camera speeds.

5.2.3 Reconstructability

Figure 4 presents the condition of the matrix of $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ as a function of camera speed across the DCT, sparse and convolutional sparse bases. The oracle set $\mathcal{S}$ for each trajectory and basis was used. As a point of reference we included the condition of a DCT basis for a fixed set $\mathcal{S}$ that had been tuned to work well across all trajectories.

This result gives some valuable insights into why the convolutional sparse coding basis performs so well across a wide variety of camera speeds. The condition of the matrix $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ remains much lower than the other bases

[5] We define RMS Err $= \sqrt{\frac{1}{PF}\sum_{t=1}^{F}\sum_{p=1}^{P}||\hat{\mathbf{x}}_{t,p} - \mathbf{x}_{t,p}||_2^2/||\mathbf{x}_{t,p}||_2^2}$, where $\mathbf{x}_{t,p}$ and $\mathbf{x}_{t,p}$ are the ground truth and estimated 3D points at frame t and point p. We normalize by the energy in the ground-truth signal to ensure a fair comparison across motion sequences.

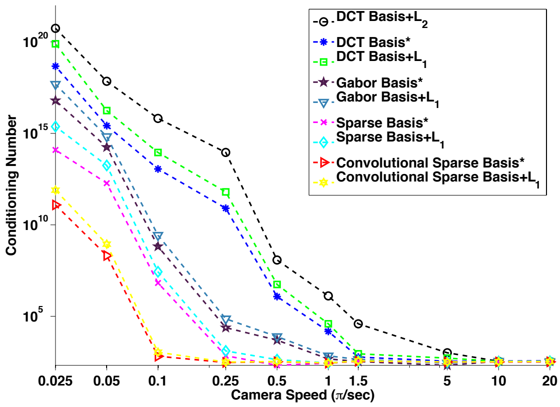


Fig. 4. The condition of matrix $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ versus different camera angle speeds (**Better viewed in color**).

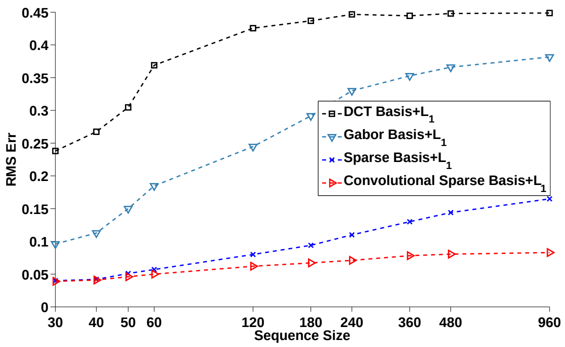


Fig. 5. Reconstruction (normalized RMS error) versus different length trajectories by sparse coding basis and convolutional sparse coding basis at camera speed 0.25 π/sec. The size of convolutional filter is 30 frames (**Better viewed in color**).

(until around 0.25π/sec). This result correlates strongly with the reconstruction results seen in Figure 3.

It is interesting to note that the condition of the matrix $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ is strongly correlated with the condition of the matrix $\boldsymbol{A}_\mathcal{S}^T \boldsymbol{A}_\mathcal{S}$ (i.e. if $\mathbf{Q}_\perp^T \mathbf{M}_\mathcal{S} \mathbf{Q}_\perp$ is singular it implies $\boldsymbol{A}_\mathcal{S}^T \boldsymbol{A}_\mathcal{S}$ is null). Further, having $\boldsymbol{A}_\mathcal{S}^T \boldsymbol{A}_\mathcal{S}$ well conditioned is a necessary (but not sufficient) condition for the RIP to hold. That is, if all possible $2F \times 2K$ submatrices of $\boldsymbol{A}$ are well conditioned then $\boldsymbol{A}_{\tilde{\mathcal{S}}}$ must be well conditioned since $|\tilde{\mathcal{S}}| = K$.

5.2.4 *Sequence Length*

A drawback to using a sparse coding basis in trajectory NRSfM reconstruction is the need to learn a different basis for each possible sequence length. Putting aside the issue of storing all these different bases, Figure 5 demonstrates for a camera speed of 0.25π/sec the reconstruction performance of sparse versus convolutional sparse coding bases as a function of sequence length. In these experiments we are not using the oracle sets, instead an ℓ_1 LASSO strategy is employed to estimate the set $\tilde{\mathcal{S}}$ and the trajectory coefficients $\tilde{\boldsymbol{\beta}}$. Interestingly, when the sequence length is small both bases obtain similar performance. However, as the sequence length gets longer the convolutional sparse coding basis exhibits much better generalization performance.

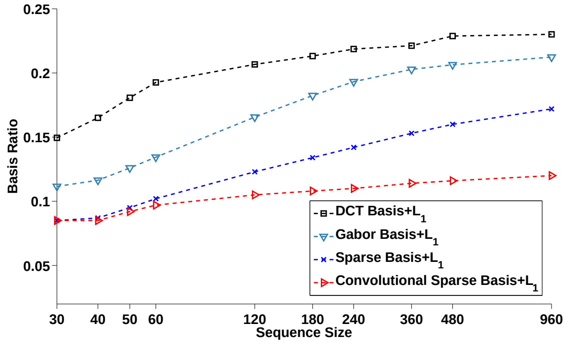


Fig. 6. Basis ratio (i.e. $K/3F$) versus sequence size (F) for the sparse coding and convolutional sparse coding bases. The size of convolutional filters is 30 frames (**Better viewed in color**).

An insight into why this superior generalization performance is obtained can be seen in Figure 6. The $y-$ axis plots the basis ratio $K/3F$ versus camera speed. We define the basis ratio as the cardinality K of the estimated set $\tilde{\mathcal{S}}$ versus the length $3F$ of the the trajectory. One can see for convolutional sparse coding this ratio remains relatively flat, however, for sparse coding it becomes higher as a function of the length of the sequence.

5.2.5 *Noise and Missing Data Test*

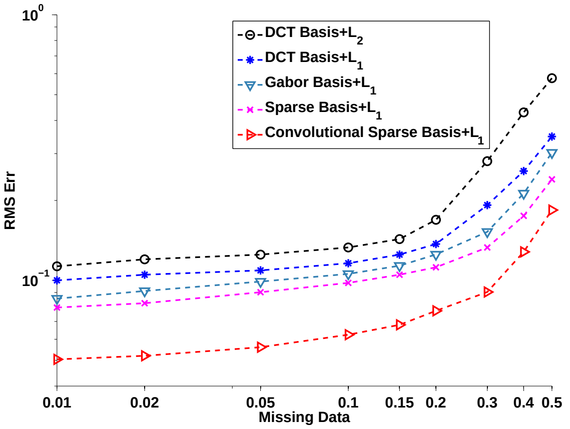


Fig. 7. The reconstruction (normalized RMS error) versus the ratio of missing data (**Better viewed in color**).

It is naive to think that the 2D projection vector $\mathbf{u} = \mathbf{Qx}$ of the 3D trajectory $\mathbf{x}$ is unaffected by noise and missing data. Our method's tolerance to missing data was evaluated by masking some of the 2D projections in a synthetic experiment. To ensure a realistic situation, occlusions were generated in blocks of adjacent frames, rather than uniformly distributed throughout the sequence. To synthesize the blocks of missing data, we use fixed size windows to select the part of continuous missing data. In our experiments we set 10 as our window size. The ratio of missing

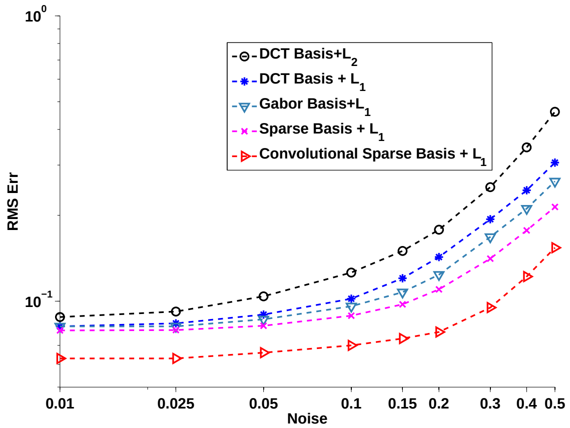


Fig. 8. Reconstruction (normalized RMS error) versus the noise magnitude (**Better viewed in color**).

data is controlled by the number of blocks.[6] The results are presented in Figure 7 for a camera speed at 20π/sec (high reconstructability). Performance of the convolutional sparse coded basis degrades gracefully with a larger fraction of occlusions. Low reconstruction error is obtained even with almost 30% missing data.

Tolerance to additive noise was also explored. We synthetically added Gaussian noise data to the 2D projection measurements at a camera speed of 20π/sec. Results are presented in Figure 8. Even though all methods degrade as a function of noise magnitude, our proposed convolutional sparse coding strategy preserves its margin of superior performance across all noise levels.

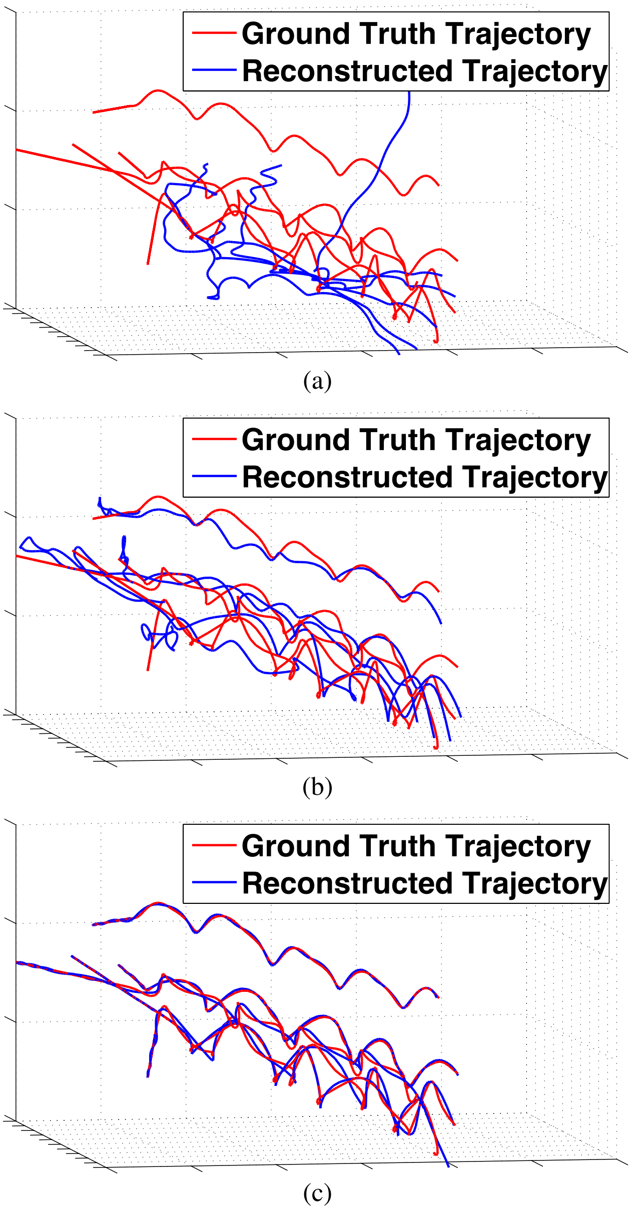
5.2.6 Visualization of the Reconstructed Trajectories

In Figure 9 we show a visualization of several trajectories taken from a running canine estimated from 2D point projections at a challenging camera speed of 0.25π/sec. In (a) we depict the reconstruction using the DCT basis with ℓ_1 penalty, (b) the sparse coding basis with ℓ_1 penalty, and (c) the convolutional sparse coding basis with ℓ_1 penalty. One can see our proposed method in (c) gives a superior reconstruction in comparison to (a) and (b).

Figure 10 shows the visualization of the reconstructed structure with the ground truth structure at randomly selected frames for the: (a) DCT basis with ℓ_1 penalty, (b) sparse coding basis with ℓ_1 penalty, and (c) convolutional sparse coding basis with ℓ_1 penalty. As expected our proposed convolutional sparse coding basis outperforms all the other strategies.

5.3 Real World Data Experiment

To fully evaluate our approach we employed a human motion dataset of "real-world" camera motion. The



(a)

(b)

(c)

Fig. 9. The reconstructed trajectories versus the ground truth trajectories at camera speed 0.25π/sec for the: (a) DCT, (b) sparse coding, and (c) convolutional sparse coding bases using the ℓ_1 (Equation 24) objective. As expected the convolutional sparse coding basis obtained superior performance (**Better viewed in color**).

Utrecht Multi-Person Motion (UMPM) benchmark [25] contains 2D video with 37 tracked points covering an articulated body. The dataset also has an accompanying 3D ground-truth stemming from motion sensors. The UMPM dataset is markedly different to CMU Motion Capture in that contains a different number of points, people, actions and interactions. All sequences from the UMPM benchmark dataset are between 50 seconds to 60 seconds (100 fps, 5000 frames to 6000 frames)and those sequences contain complicated and substantially nonrigid motion such as multiple human actions, human-object interactions and human-human interactions.

[6] The missing data is synthesized as $\hat{\mathbf{u}} = \mathbf{G}\mathbf{u} = \mathbf{G}\mathbf{Q}\mathbf{x}$, where $\mathbf{G} \in \mathbb{R}^{2F\times 2F}$ is a diagonal matrix for selecting points missed in frames. If the 2D point is missing at frame f, the element $\mathbf{G}(f, f)$ which refers to x-axis coordinate is set to 0 and $\mathbf{G}(2f, 2f)$ which refers to the y-axis coordinate is set to 0, otherwise, $\mathbf{G}(f, f) = 1$ and $\mathbf{G}(2f, 2f) = 1$.

For our reconstructions we used only the 2D point tracks within video portion of UMPM. We used 12 sequences (“p1_chair_2”, “p1_table_2”, “p1_grab_2”, “p2_orthosyn_1”, “p2_staticsyn_1”, “p2_free_2”, “p2_orthosyn_12”, “p4_ball_11”, “p4_free_11”, “p4_table_11”, “p4_table_12”, “p4_staticsyn_13”) in UMPM Benchmark dataset. All selected testing sequences are resampled from original 100 fps to 30 fps to ensure consistency between training and testing dataset. Those sequences contains missing data due to occlusions occurred at different body parts and frames. The ratio of missing data among all point trajectories of the testing sequences is between 2.5% to 10%. We estimated the camera motion by applying the Tomasi & Kanade factorization to the rigid torso as in [23].

One can see the normalized RMS reconstruction error using the DCT, sparse coding, and convolutional sparse coding bases with the ℓ_1 penalty for several real world sequences from the UMPM dataset in Figure 11. The relative motion between camera and people are slow and smooth for all these sequences. Our learned convolutional sparse coding basis achieved superior performance compared with the DCT and sparse coding bases.

5.3.1 *Visualization of Real-World Sequences*

Figure 12 shows a visualization of the reconstructed structure in comparison with ground truth structure at randomly selected frames for the: (a) DCT, (b) sparse coding, and (c) convolutional sparse coding bases. All methods used the ℓ_1 penalty for the reconstruction. Frames were taken from the “p1_table_2” sequence in the UMPM dataset [25]. As expected our proposed convolutional sparse coding method outperforms all the other strategies.

6 Discussion and Conclusions

Canonical ℓ_2 bounds on trajectory basis NRSfM reconstructability tell us that if a camera matrix composed with the trajectory basis is not well conditioned then it is impossible to obtain an exact reconstruction. In this paper, we investigate two insights associated with this bound. First, the requirement that one needs to know the optimal subset $\mathcal{S}$ of the trajectory basis a priori to be assured of an exact reconstruction. Second, that a more compact set $\mathcal{S}$ will allow for better reconstructions under a broader set of camera motions.

Inspired by recent work in sparse signal reconstruction we ask the question: under what conditions can one obtain an optimal trajectory reconstruction when one does not know the optimal set. We characterize theoretically that if the camera matrix composed with the trajectory basis satisfies the RIP (for a specified cardinality of the set $\mathcal{S}$) then the exact trajectory can be reconstructed without prior knowledge of the actual set.

We propose an ℓ_1 strategy for trajectory reconstruction that can obtain in polynomial time: (i) the subset $\mathcal{S}$ of the trajectory basis that is active based solely on the 2D projection, and (ii) the non-zero trajectory basis coefficients for 3D reconstruction. Our proposed ℓ_1 strategy can, under certain conditions, ascertain through mutual coherence (and no knowledge of the optimal set $\mathcal{S}$ or its cardinality) whether this is an exact reconstruction.

Finally, we explore the use of over-complete bases in trajectory NRSfM reconstruction. A central motivation here is by removing the shackles of orthonormality (e.g. DCT basis) from the problem we can explore bases that allow for more compact set $\mathcal{S}$. We demonstrate impressive reconstruction results using a learned convolutional sparse coding basis compared with DCT basis. An advantage of this learning filters, rather than basis, is their generalizability to any length signal (which is traditionally a drawback for learned bases).

References

[1] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *NIPS*, 2008. 1, 2, 3, 6

[2] 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. 2

[3] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000. 1, 2

[4] H. Bristow and S. Lucey. Fast convolutional sparse coding. In *CVPR*, 2013. 6

[5] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles:exact signal reconstruction from highly incomplete frequency information. *Communications on Pure and Applied Mathematics*, 59(8):1207–1233, 2006. 5

[6] E. Candes and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 2008. 4, 5

[7] E. Candies and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011. 2, 4, 5

[8] M. Chen, G. AlRegib, and B. Juang. Trajectory triangulation: 3d motion reconstruction with ℓ_1 optimization. In *ICASSP*, pages 4020–4023, 2011. 3, 7

[9] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. In *CVPR*, 2012. 6

[10] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 2

[11] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of 818 sparse overcomplete representations in the presence of noise. In *IEEE Transactions on Information Theory*, volume 52, pages 6–18, 2006. 5

[12] P. Gotardo and A. Martínez. Kernel non-rigid structure from motion. In *ICCV*, 2011. 3

[13] P. Gotardo and A. Martínez. Non-rigid structure from motion with complementary rank-3 spaces. In *CVPR*, 2011. 3

[14] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *NIPS*, pages 801–808, 2006. 5

[15] H. Park and Y. Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *ICCV*, 2011. 1

[16] H. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3D reconstruction of a moving point from a series of 2D projections. In *ECCV*, 2010. 1, 3

[17] V. Rabaud and S. Belongie. Linear embeddings in non-rigid structure from motion. In *CVPR*, 2009. 2

[18] E. Simoncelli and B. Olshausen. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 2001. 7

[19] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9:137–154, 1992. 2

[20] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3D shape from 2D motion. In *NIPS*, 2005. 2

[21] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure from motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):878–892, 2008. 2

[22] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *CVPR*, 2001. 2

[23] J. Valmadre and S. Lucey. Deterministic 3d human pose estimation using rigid structure. In *ECCV*, 2010. 10

[24] J. Valmadre and S. Lucey. General trajetory prior for nonrigid reconstruction. In *CVPR*, 2012. 1, 3, 4

[25] N. Van der Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp. UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *ICCV Workshops*, pages 1264–1269, 2011. 6, 9, 10, 12, 13

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), 2009. 4

[27] J. Xiao, J. Chai, and T. Kanade. A closed form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67:233–246, 2006. 2

[28] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus. Deconvolutional networks. In *CVPR*, 2010. 5

[29] Y. Zhu and S. Lucey. 3D motion reconstruction for real-world camera motion. In *CVPR*, 2011. 3

Yingying Zhu is a PhD student in University of Queensland and Commonwealth Science and Industrial Research Organization (CSIRO), Australia. She is working in the CSIRO ICT centre Computer Vision (CI2CV) lab. She is conducting research in the area of computer vision and machine learning, especially on Nonrigid Structure from Motion. She is a student memeber of IEEE.

Simon Lucey received the PhD degree from the Queensland University of Technology, Brisbane, Australia, in 2003. He is a science leader and a senior research scientist in the Commonwealth Science and Industrial Research Organization (CSIRO) and a current Futures Fellow Award recipient from the Australian Research Council. He holds adjunct professorial positions at the University of Queensland and Queensland University of Technology. His research interests include computer vision and machine learning and their application to human behavior. He is a member of the IEEE.

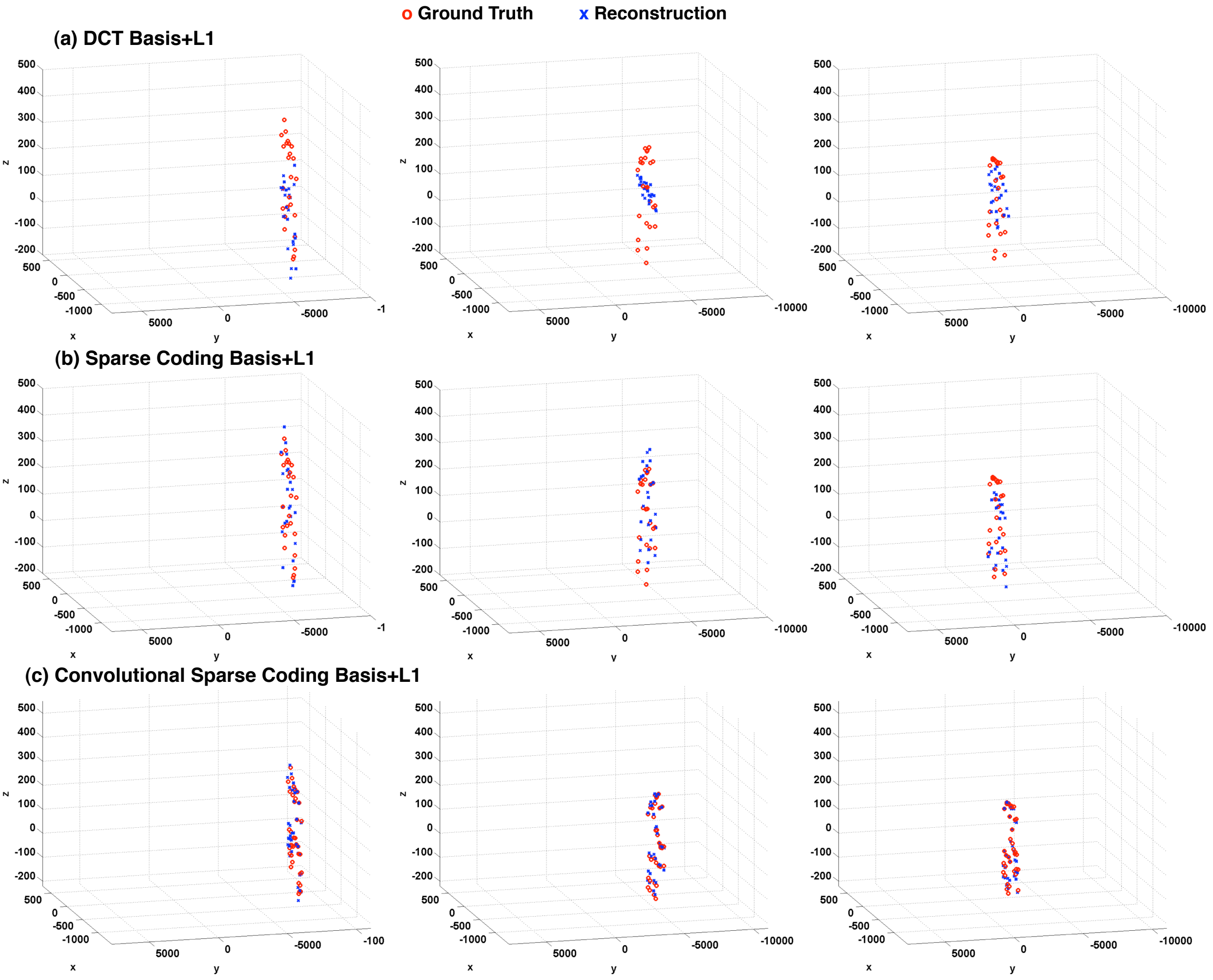


Fig. 10. The visualization of the reconstructed structure at camera speed 0.25π/sec. Visualizations of (a) the DCT, (b) sparse coding, and (c) convolutional sparse coding basis all employing the ℓ_1 objective (Equation 24) (**Better viewed in color**).

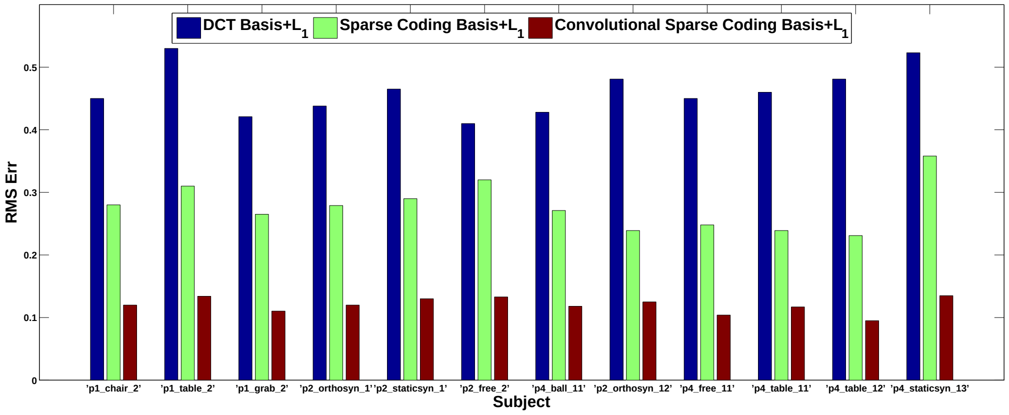
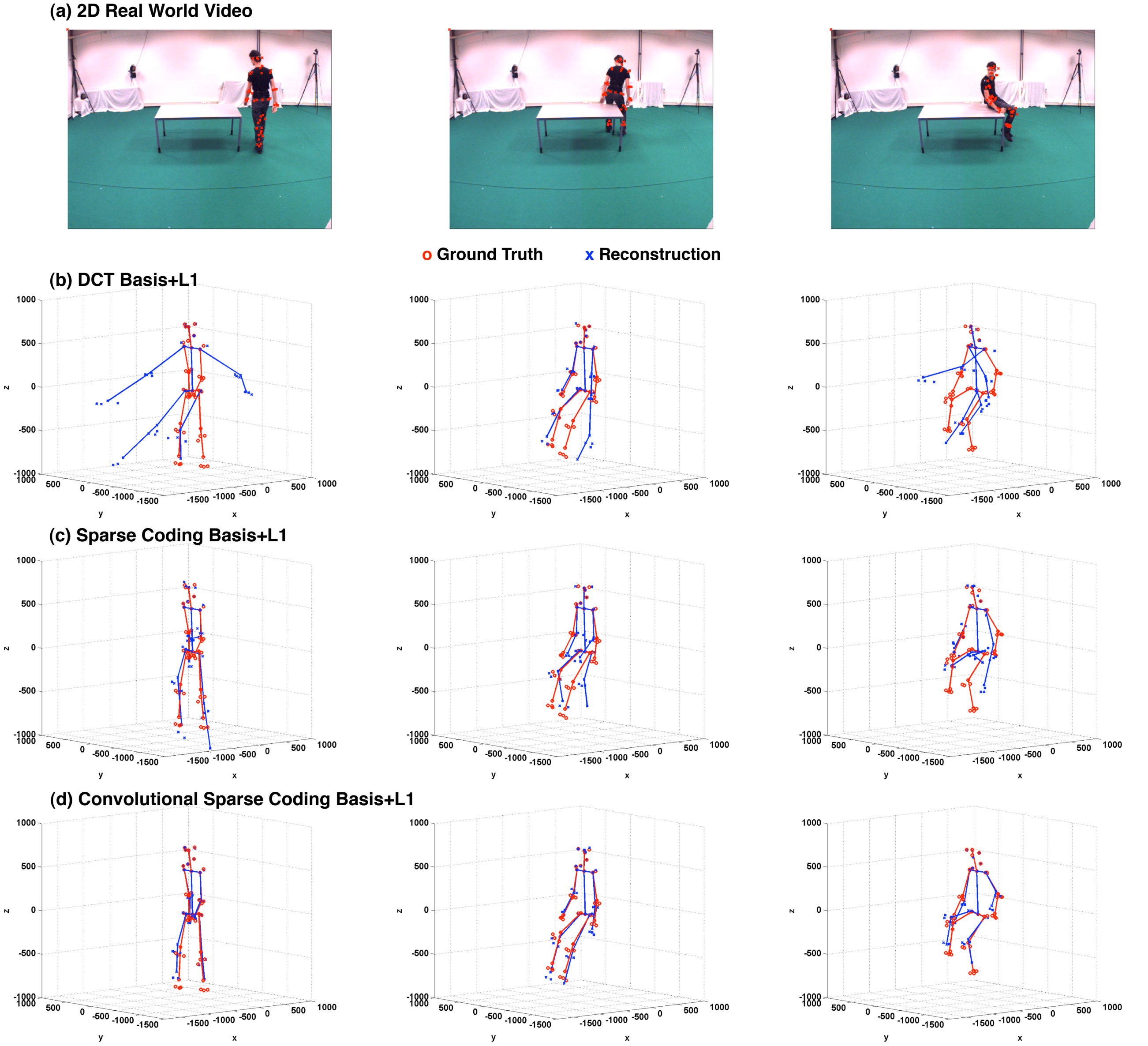


Fig. 11. Reconstruction error of DCT, sparse coding and convolutional bases with ℓ_1 penalty on selected sequences from the UMPM dataset [25] (**Better viewed in color**).



(c)

Fig. 12. Visualization of the reconstructed structure on a real world sequence ("p1_table_2") from the UMPM dataset [25]. Reconstructions of the (a) DCT basis, (b) sparse coding, and (c) convolutional sparse coding bases. The ℓ_1 objective (Equation 24) was used for all reconstructions (**Better viewed in color**).