Kronecker-Markov Prior for Dynamic 3D Reconstruction

Tomas Simon, Jack Valmadre, Iain Matthews, Yaser Sheikh

Abstract—Recovering dynamic 3D structures from 2D image observations is highly under-constrained because of projection and missing data, motivating the use of strong priors to constrain shape deformation. In this paper, we empirically show that the spatiotemporal covariance of natural deformations is dominated by a Kronecker pattern. We demonstrate that this pattern arises as the limit of a spatiotemporal autoregressive process, and derive a Kronecker Markov Random Field as a prior distribution over dynamic structures. This distribution unifies shape and trajectory models of prior art and has the individual models as its marginals. The key assumption of the Kronecker MRF is that the spatiotemporal covariance is separable into the product of a temporal and a shape covariance, and can therefore be modeled using the matrix normal distribution. Analysis on motion capture data validates that this distribution is an accurate approximation with significantly fewer free parameters. Using the trace-norm, we present a convex method to estimate missing data from a single sequence when the marginal shape distribution is unknown. The Kronecker-Markov distribution, fit to a single sequence, outperforms state-of-the-art methods at inferring missing 3D data, and additionally provides covariance estimates of the uncertainty.

Index Terms—Matrix normal distribution, Kronecker, trace-norm, spatiotemporal, missing data, generalized trace-norm.

1 Introduction

Dynamic 3D reconstruction is the problem of recovering the time-varying 3D configuration of points from incomplete observations. The theoretical and practical challenges in this problem center on the issue of missing data. In theory, dynamic 3D reconstruction is often an ill-posed problem because of *projection loss* due to the imaging of 3D information to 2D. In practice, a number of additional sources of missing data arise. First, occlusions, self-occlusions, and imaging artifacts (such as motion blur) can cause *detection loss* where points of interest are simply not detected in particular frames. Second, if points are not re-associated to their earlier detection, the system may break one trajectory into two separate trajectories, causing *correspondence loss*. While missing data issues are present in static 3D reconstruction, they are of greater significance in dynamic 3D reconstruction, as the observation system has only one opportunity to directly measure information about the structure at a particular time instant. Thus, the question at the core of dynamic 3D reconstruction is what internal model a system should refer to when there is insufficient information.

Ideally, a good model should capture all available correlations in the data—spatial, temporal, and spatiotemporal—as these correlations allow us to reason about the information that is missing. Because dynamic structure is high dimensional (e.g., 100 points over 120 frames is 36,000 degrees of freedom), the number of possible correlations is very large (i.e., $\sim$648 million parameters), and learning these correlations therefore requires a large quantity of samples, where each sample is a full spatiotemporal sequence. For most applications, such large numbers of sequences are not accessible. In this paper, we present a probabilistic model of 3D data that captures most salient correlations and can still be estimated from a few or even one sequence.

The correlations present in spatiotemporal sequences are primarily a result of separable correlations across time and correlations across structure or shape [1], [2]. Our model represents these correlations as a Matrix Normal Distribution (MND) over dynamic structure, which translate into a Kronecker pattern in the spatiotemporal covariance matrix. We show that this pattern is observed empirically for human motion sequences, and demonstrate that this pattern arises as the limit of a spatiotemporal random process under two simple assumptions. This limit explains why DCT-based bilinear basis models [2] capture a large percentage of the covariance of natural motions and provides guarantees of optimality of an analytical trajectory basis under certain conditions. However, an analytical expression for shape covariance is generally not available.

Instead, we place a prior over the shape covariance and derive a convex maximum *a posteriori* (MAP) solution to the dynamic 3D reconstruction problem in terms of the trace-norm. The model presented here applies to many dynamic 3D reconstruction problems, including nonrigid structure from motion, stereo, and multi-view trajectory reconstruction.

Contributions. (1) We are the first to identify the Kronecker pattern in time-varying 3D point cloud covariance matrices, and present a generative, probabilistic model of time-varying 3D points clouds based on the MND that explains this pattern. (2) We demonstrate that this model unifies a number of shape and trajectory models, both probabilistic and algebraic, used in prior art. (3) We establish a connection between MND and the trace-norm that leads to a convex MAP objective for reconstruction in the presence of missing data and show how the objective can be optimized using the Alternating Direction Method of Multipliers (ADMM). Empirically, our model outperforms previous approaches in handling missing data.

- *T. Simon, I. Matthews, and Y. Sheikh are with Carnegie Mellon University. Contact: http://cs.cmu.edu/$\sim$tsimon*
- *J. Valmadre is with the University of Oxford.*

This research was supported in part using an ONR grant N00014-15-1-2358.

2 Prior Art

The literature on reconstructing dynamic 3D structure is large and we focus our review on methods that directly deal with issues of information loss, either in the monocular or multi-camera case. There are largely two approaches: physically-based approaches, where ill-posed systems are conditioned according to a physically-grounded model, and statistically-based methods, where expected statistical properties of the data are used to regularize the ill-posed system without explicitly appealing to any physical grounding.

The earliest physically-based representation, in this context, was by Terzopoulos et al. [3]; subsequent work [4] presented a physically-based approach using nonlinear filtering over a superquadratic representation. Concurrently, Pentland and Horowitz [5] presented an approach where a finite element model described deformations in terms of a small number of free vibration modes, equivalent to a Kalman filter accounting for dynamics. Taylor et al. [6] revisited the idea of using rigidity but at a local scale using a minimal configuration orthographic reconstruction. Salzmann and Urtasun [7] described a number of physically-based constraints on trajectories of points that could be applied via convex priors. Investigation into statistically-based methods began with Tomasi and Kanade's rank 3 theorem [8], which established that image measurements of a rigidly rotating 3D object lay in a three dimensional subspace. The associated factorization algorithm was extended by Bregler et al. for nonrigid objects [9], positing that a shape space spanned the set of possible shapes. Unlike the rigid case, where the bilinear form could be solved using singular value decomposition (SVD), this formulation had a trilinear form. Bregler et al. proposed a nested SVD routine, which proved to be sensitive to initialization and missing data. A series of subsequent papers investigated various constraints to better constrain the solution or relax the optimization (a sample of major work includes [10], [11], [12], [13]). Recently, Dai et al. [14] presented a method that uses a trace-norm minimization to enforce a low rank shape space, and Garg et al. [15] showed that the method can be applied to recover dense, non-rigid structure. Lee et al. [16] expanded on the shape distribution model by explicitly including procrustes alignment as part of a probabilistic parameterization they call the Procrustean Normal Distribution (PND), later extended to PND mixtures by Cho et al. [17].

In conjunction, trajectory space representations were proposed by Sidenbladh et al. [18], which they referred to as *eigenmotions*. Akhter et al. [19] noted that, in trajectory space, a predefined basis could be used, which reduced the trilinear form to a bilinear form and allowed the use of SVD once again to recover the nonrigid structure. Unfortunately, the solution was shown to be sensitive to missing data and cases where the camera motion is smooth [20]. Park et al. [20] used static background structure to estimate camera motion, reducing the optimization into a linear system, and were able to handle missing data. Valmadre and Lucey [21] presented various priors on trajectories in terms of 3D point differentials, showing better noise performance than using a truncated basis.

A number of approaches have combined spatial and temporal constraints [1], [4], [5], [22], [23]. Torresani et al. [23] presented a probabilistic representation, using probabilistic PCA within a linear dynamical system, and, similarly, Lee et al. [24] combined the PND shape distribution with a temporal Markov process. The shape basis and trajectory basis approaches were combined within a single estimation procedure by Gotardo and Martinez [1], and later developed as a bilinear basis by Akhter et al. [2].

The model presented here is a probabilistic formulation of spatiotemporal bilinear basis models and was first published as [25], which we have updated here to include new theoretical insights and experiments. In contrast to prior work, our model describes an explicit parametric distribution over spatiotemporal data that can be estimated from a single sequence. This allows us to define a spatiotemporal covariance matrix relating any point in time to any other point in time, including covariance matrices for missing data. As summarized in Table 1 (Sect. 3.4), we take a step towards reconciling a number of recent statistically-based linear representations in nonrigid structure from motion [1], [9], [14], [19], [21], [22], [23], [26].

3 Modeling Time-Varying 3D Structures

The time-varying structure of a configuration of P 3D points across F frames can be represented by a matrix $\mathbf{X} \in \mathbb{R}^{F \times 3P}$. The row t corresponds to the 3D shape in frame t, and is formed by the horizontal concatenation of points $X_p^t \in \mathbb{R}^{1\times 3}$, denoting the p-th 3D point. We will denote by $\mathbf{x} = \mathrm{vec}(\mathbf{X})$ the column-major vectorization of the matrix $\mathbf{X}$, and we will interchangeably use lowercase bold letters to denote the vectorized matrices.

3.1 Observation Model

In practice, due to missing data and camera projection, only a reduced set of measurements of $\mathbf{X}$ are observed. We model observations linearly as

$$\mathbf{y} = \mathbf{O}\,\mathrm{vec}(\mathbf{X}) + \epsilon, \tag{1}$$

where $\mathbf{y}$ is a vector of observations of size n_{obs} (the number of observations), $\mathbf{O} \in \mathbb{R}^{n_{\text{obs}} \times 3FP}$ is the observation matrix, and ϵ is noise sampled from a normal distribution. In the simplest case of fully observed data, $\mathbf{O}$ is an identity matrix of size $3FP \times 3FP$. For entries x, y, or z that are missing, we would remove the corresponding rows of the identity matrix, yielding a matrix $\mathbf{O}_{\text{miss}}$ containing a subset of the rows.

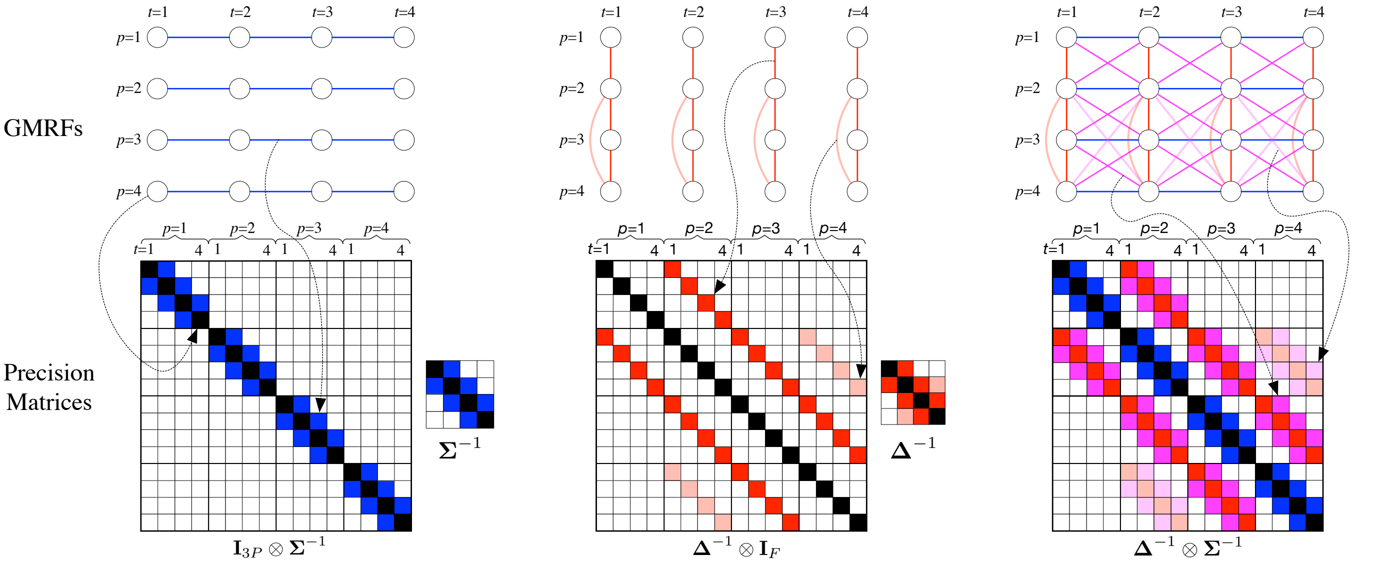
The action of camera projection can also be modeled by $\mathbf{O}$. For ease of notation, let us briefly consider the row-major vectorization $\mathrm{vec}(\mathbf{X}^T)$. For this arrangement, the effect of orthographic projection from a single camera can be expressed as a matrix $\mathbf{O}_{\text{ortho}}$ such that each of the P points in frame f is transformed by the first two rows of a rotation matrix, $\mathbf{R}_f \in \mathbb{R}^{2\times 3}$,

$$\mathbf{y} = \begin{pmatrix} \mathbf{R}_1 \otimes \mathbf{I}_P & & \\ & \ddots & \\ & & \mathbf{R}_F \otimes \mathbf{I}_P \end{pmatrix} \mathrm{vec}(\mathbf{X}^T) + \epsilon, \tag{2}$$

The case of a single camera observing the scene with unknown rotations $\mathbf{R}_f$ is the problem of non-rigid structure from motion (NRSfM). For multiview reconstruction, several $\mathbf{O}_{\text{ortho}}$ matrices can be stacked, one for each camera observing the scene. If some of the projected points are missing, we can concatenate the effect of the matrices: $\mathbf{O} = \mathbf{O}_{\text{miss}}\mathbf{O}_{\text{ortho}}$. In this paper, we assume that the observation matrix $\mathbf{O}$ is known (e.g., via rigid SfM [27] or inertial measurements); simultaneous recovery of the camera matrices (as in NRSfM) is not the focus of this paper. However, we will relax this constraint in Section 5.2 and jointly optimize the rotation matrices.

Our objective is to estimate the most likely spatiotemporal structure $\hat{\mathbf{X}}$ given the observations $\mathbf{y}$. Note that $n_{\text{obs}} \ll 3FP$, and the problem

$$\min_{\mathbf{X}}.\ \sigma^{-2}\|\mathbf{y} - \mathbf{O}\,\mathrm{vec}(\mathbf{X})\|_2^2 \tag{3}$$


GMRFs
Precision Matrices
t=1
t=2
t=3
t=4
p=1
p=2
p=3
p=4
$\mathbf{\Sigma}^{-1}$
$\mathbf{\Delta}^{-1}$
$\mathbf{I}_{3P} \otimes \mathbf{\Sigma}^{-1}$
$\mathbf{\Delta}^{-1} \otimes \mathbf{I}_F$
$\mathbf{\Delta}^{-1} \otimes \mathbf{\Sigma}^{-1}$

(a) AR-1 Trajectory Model (b) Point Distribution Shape Model (c) Kronecker Shape-Trajectory Model

Fig. 1. We define the Kronecker Shape-Trajectory GMRF as resulting from the Kronecker product of independent shape and trajectory precision matrices, describing shape-only and trajectory-only GMRF models of deformation respectively.

is therefore severely under constrained. We take a Bayesian approach to the estimation problem,

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmax}}\, p(\mathbf{X}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X})p(\mathbf{X}), \tag{4}$$

where $p(\mathbf{y}|\mathbf{X}) = \mathcal{N}\left(\mathbf{O}\,\mathrm{vec}(\mathbf{X}), \sigma^2\mathbf{I}\right)$ from Eq. (1).

Inference from Eq. (4) requires designing a prior for the dynamic 3D structure, $p(\mathbf{X})$, that models the data well while remaining amenable to global optimization. To allow for arbitrary objects and shapes, we decompose the time-varying structure $\mathbf{X} \in \mathbb{R}^{F\times 3P}$ as the sum of a mean component $\mathbf{M}$ (modeling the object's translation and mean shape) and a zero-mean matrix with any remaining deformations, which we call the non-rigid component $\mathbf{Z}$:

$$\mathbf{X} = \mathbf{M} + \mathbf{Z}. \tag{5}$$

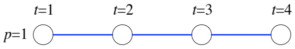
This decomposition allows us to set a prior over non-rigid deformations while leaving the mean shape unconstrained.

3.2 Shape, Trajectory, and Shape-Trajectory Priors

In the following, we will show how several widely-used temporal and spatial non-rigid priors can be unified into a single spatiotemporal prior over the non-rigid structure $\mathbf{Z}$. Similarly to $\mathbf{X}$, the arrangement of $\mathbf{Z}$ is such that each row is the non-rigid component of 3D shape at a particular time instant, $\mathbf{z}^t \in \mathbb{R}^{3P}$, (the t^{th} row arranged as a column vector), and each column $\mathbf{z}_p \in \mathbb{R}^F$ is the non-rigid time-trajectory of the x, y, or z coordinate of a particular point. Previous literature has considered shape priors (e.g., Torresani et al. [23]), temporal (or trajectory) priors (e.g., Ahkter et al. [19]), and spatiotemporal basis models (e.g., Gotardo and Martinez [1]). These approaches differ primarily in their independence assumptions; for example, shape basis methods assume that individual frames (i.e., rows of $\mathbf{Z}$) are conditionally independent, whereas temporal priors assume independence between point trajectories (columns of $\mathbf{Z}$). To expose these differences, we characterize each model in terms of the Markov Random Field (MRF) [28] structure that is implied by the assumptions made.

3.2.1 *Trajectory MRF*

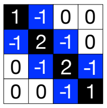
Points move smoothly over time as a direct consequence of objects having mass and following physical laws of motion. This intuition has been variously modeled as an autoregressive (AR) process (e.g., [23], [24]), temporal smoothing or filtering (e.g., [21], [29], [30]), physically-based energy minimization (e.g., [7]) or smooth basis approximations (e.g., [19], [20]). All of these examples can be expressed as an MRF where nearby frames are linked; for example, as depicted in the inset for 4 frames. Here, each node represents a location variable (a coordinate x, y, or z) at a particular point in time. This MRF forms an auto-regressive order 1 ($AR(1)$) model where the conditional dependence between adjacent frames is Gaussian, i.e., $Z_p^t = \phi Z_p^{t-1} + \epsilon$, where ϕ represents the partial correlation between frames and ϵ is Gaussian with variance σ^2, representing deviations from the model. This Gaussian MRF (GMRF) [28] modeling a single trajectory is completely defined by its precision matrix, $\mathbf{\Sigma}^{-1} \in \mathbb{R}^{F\times F}$.



Recall that the precision matrix is the inverse of the covariance matrix, and an entry (i, j) indicates the partial correlation between variables i and j. Consequently, the precision matrix is zero everywhere except between points that are connected in the graphical model (see Rue and Held [28]). If we now consider a set of P *independent* points, we can similarly express the model as a GMRF where only variables that are temporal neighbors are linked. This is illustrated in Fig. 1(a) for 4 point coordinates over 4 time instants. Each node represents an entry t, p in the spatiotemporal matrix $\mathbf{Z}$ and the lack of connections between different points indicates that this is a purely temporal prior. The corresponding GMRF describes a joint distribution over $\mathbf{Z}$ that is completely specified by the $3PF \times 3PF$ covariance matrix $\mathbf{\Phi}$ (or, equivalently, the precision matrix $\mathbf{\Phi}^{-1}$). Because we assumed spatial independence between points, the log-likelihood for a set of points is the sum of the P likelihoods, and so, the matrix is block diagonal where each block is the precision matrix of a single point. This can be expressed algebraically as $\mathbf{\Phi}^{-1} = \mathbf{I}_{3P} \otimes \mathbf{\Sigma}^{-1}$

where $\mathbf{I}_{3P}$ is the $3P \times 3P$ identity and $\otimes$ denotes the Kronecker product, and the choice of trajectory precision matrix $\mathbf{\Sigma}^{-1}$ gives rise to different trajectory priors.

For a single point with the 4-link chain above, the matrix $\mathbf{\Sigma}^{-1}$ is depicted in the inset, where white entries correspond to 0 and are independent when conditioned on the remaining variables. Ahmed, Natarajan, and Rao [31] proved that, for an $AR(1)$ process where ϕ tends to 1, the blue entries tend to become proportional to -1 as the black entries become proportional to 2. This corresponds to a DCT-2 matrix [31] (a tri-diagonal matrix with 2 on the diagonal and -1 on the off-diagonals, except the first and last entries). Therefore, for temporal processes where $\phi \approx 1$, the optimal trajectory basis in the L2 sense is given by the eigenvectors of this matrix, which correspond to cosines of different frequencies. This explains the effectiveness of trajectory-basis methods that use a truncated DCT basis (e.g., [1], [19], [20]).



1	-1	0	0
-1	2	-1	0
0	-1	2	-1
0	0	-1	1

More recently, Valmadre et al. [21] showed improved performance over a truncated DCT basis when using a filtering-based temporal smoothness prior. Interestingly, the model of Valmadre et al. is equivalent to minimizing the negative log-likelihood of the GMRF model given above, i.e.,

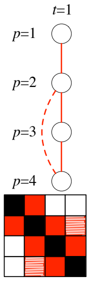
$$-\log(p(\mathbf{z}_p)) \propto \mathbf{z}_p^T \mathbf{\Sigma}^{-1} \mathbf{z}_p,$$

in the limit of $\phi \to 1$. In this case, the precision matrix corresponds to $\mathbf{\Sigma}^{-1} = \mathbf{D}^T\mathbf{D}$, with $\mathbf{D}$ the forward differences matrix[1]. The second-order differences model of [21] (an $AR(2)$ model) would in turn correspond to a GMRF with links to the adjacent frame and to the frame adjacent to that, (i.e., $\mathbf{\Sigma}^{-1} = \mathbf{D}^T\mathbf{D}\mathbf{D}^T\mathbf{D}$).

3.2.2 Shape MRF

However, the assumption that points move independently is intuitively wrong: nearby points on a surface will tend to move similarly, and their positions will therefore be correlated. A popular approach to capture this intuition is to model the covariance between shape coordinates as a *Point Distribution Model*, or PDM. Cootes et al. [32] described the standard procedure of combining generalized Procrustes analysis with PCA to model a distribution over shapes, but there exist many variants of essentially the same technique: shape basis models (e.g., [9], [23]), morphable models (e.g., [33]), low-rank priors (e.g., [14]), the Procrustean Normal Distribution (PND) [34], and modal analysis (e.g., [35]).

Assuming aligned shapes, the PDM prior models the shape distribution as a Gaussian and can therefore be expressed as a GMRF as well. As an illustrative example in the inset, we show links between adjacent points and between points 2 and 4, with the pattern of non-zero entries in the shape precision matrix as shown right.



In this matrix, an entry i, j indicates the partial correlation between point coordinates i and j, and white entries represent zeroes. In general, the connectivity between points in this MRF can be full, for example, when the shape precision matrix is computed via PCA (or PPCA) on Procrustes aligned shapes. In this case, the matrix is the inverse of the sample covariance matrix with $\mathbf{\Delta}^{-1} = \mathbf{U}^T\mathbf{S}^{-1}\mathbf{U}$, where $\mathbf{U}$ is the set of eigenvectors or deformation modes, and $\mathbf{S}$ is the diagonal matrix of the corresponding eigenvalues. For a single frame, the negative log-likelihood of the PDM prior at time t is proportional to $\mathbf{z}^{tT}\mathbf{\Delta}^{-1}\mathbf{z}^t$.

For a set of F *independent* frames, the connectivity is given by the MRF in Fig. 1(b) and the $3PF \times 3PF$ GMRF precision matrix is $\mathbf{\Phi}^{-1} = \mathbf{\Delta}^{-1} \otimes \mathbf{I}_F$. This matrix is visualized in Fig. 1(b). Note the symmetry with respect to the temporal model's precision matrix—indeed, this structure is also block-diagonal, but, in this case, for a *row-major* arrangement $\text{vec}(\mathbf{Z}^T)$, and would have matrices $\mathbf{\Delta}^{-1}$ along the diagonal.

3.2.3 Kronecker-Markov MRF

In the following, we show that by relaxing the independence assumptions of the above models, the natural combination of these priors results in an approximately *Kronecker* spatiotemporal covariance matrix. This result follows from simultaneously applying the shape and temporal priors:

1) At any time instant t, the configuration of points follows a zero-mean PDM:

$$\mathbf{z}^t = \mathbf{B}\mathbf{c}^t \quad \text{with} \quad \mathbf{c}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{6}$$

where $\mathbf{z}^t$ is the t^{th} row of matrix $\mathbf{Z}$ arranged as a column vector, $\mathbf{\Delta} = \mathbf{B}\mathbf{B}^T \in \mathbb{R}^{D\times D}$ is the shape covariance and $\mathbf{c}_t$ a vector of coefficients.

2) The dynamic evolution of the system follows a vector-valued $AR(1)$ process:

$$\mathbf{z}^t = \boldsymbol{\phi}\mathbf{z}^{t-1} + \mathbf{v}^t \quad \text{with} \quad \mathbf{v}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{7}$$

where $\boldsymbol{\phi} \in \mathbb{R}^{D\times D}$ describes the temporal dynamics, and $\mathbf{K} \in \mathbb{R}^{D\times D}$ describes i.i.d. Gaussian deviations from the model. Later, we will demonstrate that the matrix $\mathbf{K}$ is a scaled version of the spatial covariance matrix of the corresponding PDM.

From Eq. (7), using the Markov property and the chain rule we can write the joint distribution over a set of frames $[1, \ldots, F]$ as,

$$\begin{aligned} p(\mathbf{Z}) = p(\cdots, \mathbf{z}^{t-1}, \mathbf{z}^t, \mathbf{z}^{t+1}, \mathbf{z}^{t+2}, \cdots) = \\ \cdots p(\mathbf{z}^{t+1}|\mathbf{z}^t)p(\mathbf{z}^t|\mathbf{z}^{t-1})p(\mathbf{z}^{t-1}|\mathbf{z}^{t-2})\cdots, \end{aligned}$$

where each conditional distribution is independently Gaussian and is given by

$$p(\mathbf{z}^t|\mathbf{z}^{t-1}) = \frac{1}{C}\exp\left(-\frac{1}{2}(\mathbf{z}^t - \boldsymbol{\phi}\mathbf{z}^{t-1})^T\mathbf{K}^{-1}(\mathbf{z}^t - \boldsymbol{\phi}\mathbf{z}^{t-1})\right).$$

Taking the negative log-likelihood of the joint model, the general form is

$$\begin{aligned} -\log(p(\mathbf{Z})) = \cdots &+ (\mathbf{z}^{t+1} - \boldsymbol{\phi}\mathbf{z}^t)^T\mathbf{K}^{-1}(\mathbf{z}_{t+1} - \boldsymbol{\phi}\mathbf{z}^t) \\ &+ (\mathbf{z}^t - \boldsymbol{\phi}\mathbf{z}^{t-1})^T\mathbf{K}^{-1}(\mathbf{z}^t - \boldsymbol{\phi}\mathbf{z}^{t-1}) \\ &+ \cdots - \log(C), \end{aligned}$$

for some normalizing constant C. Letting $\mathbf{J} = (\mathbf{K}^{-1} + \boldsymbol{\phi}^T\mathbf{K}^{-1}\boldsymbol{\phi})$, and $\mathbf{H} = -\boldsymbol{\phi}^T\mathbf{K}^{-1}$, the negative log-likelihood of the set of frames can be written as a block tri-diagonal quadratic form,

$$\begin{pmatrix} \vdots \\ \mathbf{z}^{t+1} \\ \mathbf{z}^t \\ \mathbf{z}^{t-1} \\ \vdots \end{pmatrix}^T \begin{pmatrix} \ddots & & & & \\ \mathbf{H}^T & \mathbf{J} & \mathbf{H} & & \\ & \mathbf{H}^T & \mathbf{J} & \mathbf{H} & \\ & & \mathbf{H}^T & \mathbf{J} & \mathbf{H} \\ & & & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ \mathbf{z}^{t+1} \\ \mathbf{z}^t \\ \mathbf{z}^{t-1} \\ \vdots \end{pmatrix}. \tag{8}$$

[1] A matrix with 1 on the diagonal (except the last entry) and -1 on the upper diagonal. This matrix is rank deficient and the precision matrix and PDF do not (strictly) exist. However, note that this defines a proper distribution over a subspace that excludes the null-space of the matrix (which corresponds to a constant shape, explicitly excluded in Eq. (5)). Keeping this in mind, the inverses may be replaced with pseudo-inverses without affecting the results.

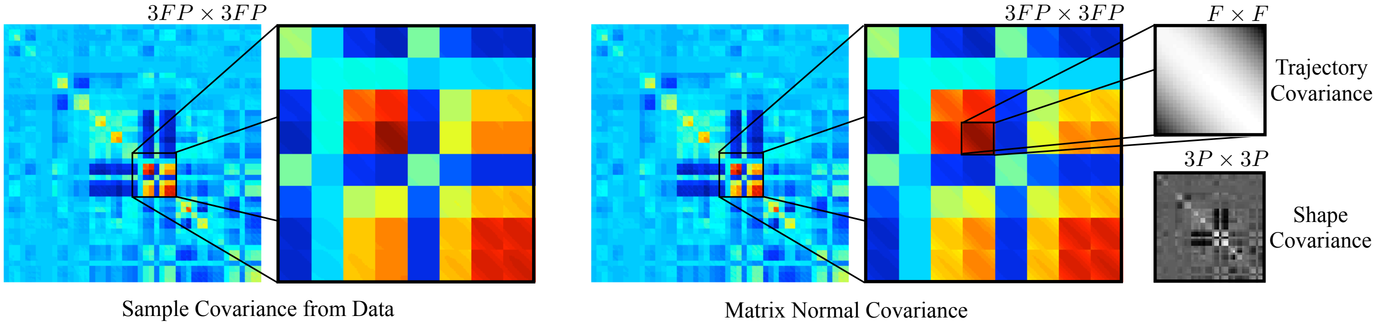


Fig. 2. Human spatiotemporal point cloud data exhibits a Kronecker structured covariance matrix, allowing us to model the distribution over sequences as matrix normal. (Left) The spatiotemporal covariance computed from 5402 vectorized sequences shows a distinct block structure, highlighted in the inset. (Right) The corresponding covariance of the matrix normal model, where the full $(3FP) \times (3FP)$ matrix is separable into two smaller covariance matrices, the $F \times F$ trajectory (row) and $3P \times 3P$ shape (column) covariances respectively. Here, $F = 30$ frames and $P = 16$ points.

When the transition matrix tends to $\boldsymbol{\phi} \to \mathbf{I}$, then $\mathbf{J} \approx 2\mathbf{K}^{-1}$, and $\mathbf{H} \approx -\mathbf{K}^{-1}$, and, in the limit, we can re-write the negative log-likelihood as $\mathrm{vec}(\mathbf{Z}^T)^T \Psi^{-1} \mathrm{vec}(\mathbf{Z}^T)$ with

$$\Psi^{-1} \approx \alpha \left[\underbrace{\begin{pmatrix} \star & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \vdots \\ \vdots & & & & \\ 0 & 0 & 0 & -1 & \star \end{pmatrix}}_{\boldsymbol{\Sigma}^{-1}} \otimes \mathbf{K}^{-1} \right], \qquad (9)$$

where the entries marked $\star$ depend on the boundary conditions chosen for the first and last time instants (see [36]), and α is a normalizing constant. This is a generalization of the result of Ahmed, Natarajan, and Rao (see Sect. 3.2.1), but for multivariate $AR(1)$ processes: the optimal (in the L2 sense) spatiotemporal basis for a vector-valued $AR(1)$ process with constant noise covariance will be given by the eigenvectors of Eq. (9) when $\boldsymbol{\phi} \to \mathbf{I}$.

In general, we will say that a vector-valued $AR(1)$ process $\mathbf{z}^t$ with $\boldsymbol{\phi} \approx \mathbf{I}$ has a spatiotemporal covariance Ψ that is approximately Kronecker-Markov, with covariance $\Psi \approx \boldsymbol{\Sigma} \otimes \mathbf{K}$, for some matrix $\mathbf{K}$, or equivalently $\boldsymbol{\Phi} \approx \mathbf{K} \otimes \boldsymbol{\Sigma}$ for the corresponding column-major arrangement.

The relationship between $\mathbf{K}$ and the PDM shape covariance $\boldsymbol{\Delta}$ can be derived by taking the marginal probability $p(\mathbf{z}^t)$, i.e., the probability of observing a particular shape at time t after marginalizing out all other time instants. In this case, we can see that the marginal shape distribution is

$$\mathbf{z}^t \sim \mathcal{N}(\mathbf{0}, \Sigma_{t,t}\mathbf{K}). \qquad (10)$$

From the assumption that individual frames follow a PDM distribution (Eq. (6)), we can conclude that $\mathbf{K} = \Sigma_{t,t}^{-1}\boldsymbol{\Delta}$, i.e., $\mathbf{K}$ is equal to the shape covariance up to a constant scale factor[2]. Note, however, that $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta}$ are not uniquely identifiable since $\boldsymbol{\Sigma} \otimes \boldsymbol{\Delta} = \frac{1}{\alpha}\boldsymbol{\Sigma} \otimes \alpha\boldsymbol{\Delta}$ for any non-zero scalar; we will therefore assume that we can find a scale factor such that $\Sigma_{t,t} = 1$.

We define the Kronecker-Markov prior as a GMRF with a precision matrix $\boldsymbol{\Phi}^{-1} = \boldsymbol{\Delta}^{-1} \otimes \boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}^{-1}$ is the DCT-2 matrix. Its connectivity diagram is shown in Fig. 1 (c), where we link a point p_1 at time instant t_1 and point p_2 at time instant t_2 iff there exists a link between p_1 and p_2 in the shape MRF, *and* a link between t_1 and t_2 in the temporal MRF. This can be generalized to higher order Markov chain models by allowing arbitrary temporal precision matrices $\boldsymbol{\Sigma}^{-1}$.

Fig. 2 illustrates the intuition for choosing this prior for dynamic 3D structures: the spatiotemporal covariance matrix of natural human motions is dominated by a Kronecker product block pattern (the specifics of this experiment are described in Sect. 6.1). This is a significant finding for the purposes of estimation because it allows us to parameterize the spatiotemporal covariance of a dynamic 3D structure with far fewer free variables than are needed for a general, unstructured covariance matrix. The number of covariance parameters in the Kronecker covariance is approximately $\frac{(3P)^2}{2} + \frac{(F)^2}{2}$, versus $\frac{(3FP)^2}{2}$ for a full covariance matrix. Even for small values of F=30 frames and P=31 points, this results in ~5000 variables for the Kronecker versus ~3.9 million for a full spatiotemporal covariance.

3.3 Matrix Normal Distributions for 3D Point Clouds

The Kronecker GMRF described above corresponds to a Matrix Normal Distribution (MND) [37]. We use this relationship to specify the prior over dynamic 3D structure, $p(\mathbf{X})$, in terms of distributions over $\mathbf{Z}$ (the non-rigid component) and $\mathbf{M}$ (the rigid, mean component), with $\mathbf{X} = \mathbf{M} + \mathbf{Z}$.

3.3.1 Modeling the Non-rigid Component $\mathbf{Z}$

From Eq. (9), the dynamic 3D structure will have an approximately Kronecker structured covariance matrix, and

$$\mathrm{vec}(\mathbf{Z}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta} \otimes \boldsymbol{\Sigma}). \qquad (11)$$

Equivalently, this model corresponds to a matrix normal distribution [38], [37] over the non-rigid component $\mathbf{Z}$, which we can write as

$$\mathbf{Z} \sim \mathcal{MN}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}), \qquad (12)$$

where $\mathcal{MN}$ denotes an MND with row covariance $\boldsymbol{\Delta}$=$\mathbf{B}\mathbf{B}^T$ (describing shape correlations) and column covariance $\boldsymbol{\Sigma}$=$\boldsymbol{\Theta}\boldsymbol{\Theta}^T$

[2] In practice, the first and last time instants can be scaled differently depending on the boundary conditions, see Eq. (9).

(describing trajectory correlations). This formulation exposes the relationship to bilinear spatiotemporal basis models [2], with

$$\mathbf{Z} = \mathbf{\Theta} \mathbf{C} \mathbf{B}^T, \qquad (13)$$

where $\mathbf{C} \in \mathbb{R}^{F \times 3P}$ is a matrix of mixing coefficients, $\mathbf{B} \in \mathbb{R}^{3P \times 3P}$ is a complete shape basis and $\mathbf{\Theta} \in \mathbb{R}^{F \times F}$ a complete trajectory basis such that $\mathbf{B} = \tilde{\mathbf{B}} \mathbf{W}_b$ and $\mathbf{\Theta} = \tilde{\mathbf{\Theta}} \mathbf{W}_t$ where $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Theta}}$ are orthonormal and $\mathbf{W}_b, \mathbf{W}_t$ are diagonal weighting matrices. When the distribution over coefficients $\mathbf{C}$ is $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then the distribution over $\mathbf{Z}$ is matrix normal as in Eq. (12).

This probabilistic formulation subsumes the bilinear basis models of [1] and [2], where truncated versions of the orthonormal basis matrices $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Theta}}$ are used. The limit of Eq. (9) explains the effectiveness of using a truncated DCT trajectory basis in a bilinear formulation. In fact, as $\phi \to 1$ (e.g., if we sample at increasing rates) the optimal L2 spatiotemporal basis vectors will be the eigenvectors of $\mathbf{\Delta} \otimes \mathbf{\Sigma}$, which are[3] $\{\tilde{\mathbf{B}}_p \otimes \tilde{\mathbf{\Theta}}_t\}$ for $p \in \{1, \ldots, 3P\}$ and $t \in \{1, \ldots, F\}$. This set corresponds directly to a bilinear basis model [2], with $\mathbf{\Theta}$ the DCT basis.

3.3.2 Modeling the Mean Component $\mathbf{M}$

In addition to the non-rigid component $\mathbf{Z}$ described above, we model the rigid shape of the object and its translational motion as a mean component $\mathbf{M}$. This component is

$\mathbf{M} = \mathbf{1}_F \mathbf{m}_{\text{shape}} + \mathbf{M}_{\text{trans}} \mathbf{P}_{\text{trans}}$,

where the zero-centered mean 3D shape is $\mathbf{m}_{\text{shape}} \in \mathbb{R}^{1 \times 3P}$, and the mean 3D trajectory is $\mathbf{M}_{\text{trans}} \in \mathbb{R}^{F \times 3}$ (containing the per-frame translation of the object), where $\mathbf{P}_{\text{trans}} = \text{blkdiag}(\mathbf{1}_P^T; \mathbf{1}_P^T; \mathbf{1}_P^T) \in \mathbb{R}^{3 \times 3P}$, with $\mathbf{1}_P$ denoting a column vector of ones of size P, and blkdiag produces a block diagonal matrix.

We do not have a preferred shape of objects, and so we do not set a prior over the mean shape $\mathbf{m}_{\text{shape}}$. However, the translational motion $\mathbf{M}_{\text{trans}}$ is necessarily smooth, and we will therefore favor smooth trajectories of the object using the trajectory prior[4]:

$$\mathbf{M}_{\text{trans}} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_3), \qquad (14)$$

where the row covariance $\mathbf{I}_3$ reflects that there are no a priori correlations between the x, y, and z components of motion.

3.4 Relationship to Previous Work

The model over dynamic 3D structure described in this paper can be related to shape, trajectory, and shape-trajectory representations used in prior art [1], [9], [18], [19], [21], [23], [32], [40] (see Table 1).

In the following, consider the MND prior over point cloud data $\mathbf{X} \sim \mathcal{MN}(\mathbf{M}, \mathbf{\Delta}, \mathbf{\Sigma})$ with known distribution parameters $\mathbf{M}$, $\mathbf{\Delta}$, and $\mathbf{\Sigma}$.

Trajectory Methods. The MND describes a joint shape-trajectory distribution, but the marginal distribution it induces for a particular trajectory $\mathbf{x}_p$ (a column p of $\mathbf{X}$) independent of all other points corresponds to a basis representation over trajectories, as described by Sidenbladh et al. [18]. The marginal distribution is $\mathbf{x}_p \sim \mathcal{N}(\mathbf{M}_p, \Delta_{p,p} \mathbf{\Sigma})$, where $\mathbf{\Sigma}$=$\mathbf{\Theta}\mathbf{\Theta}^T$ is the trajectory covariance matrix, and $\Delta_{p,p}$ loosely corresponds to the mass of each point. This expression is equivalent to the *filtering* solution proposed by Valmadre and Lucey [21], who observe that a combination of first and second-order differences fit natural motions well. See also [7] for a physically-based formulation of the same model.

TABLE 1
Comparison of linear methods for structure reconstruction. The symbols are explained in Section 3.2.

	Truncation	**Probabilistic**	**Low Rank**
Shape	Bregler et al. [9] $\mathbf{Z} = \mathbf{\Omega} \tilde{\mathbf{B}}^T$	Torresani et al. [23] $\mathbf{Z} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}, \mathbf{B}\mathbf{B}^T)$	Dai et al. [14] $\|\mathbf{Z}\|_*$
Trajectory	Akhter et al. [19] $\mathbf{Z} = \tilde{\mathbf{\Theta}} \mathbf{A}$	Valmadre et al. [21] $\mathbf{Z} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Theta}\mathbf{\Theta}^T, \mathbf{I})$	
Shape-Trajectory	Gotardo and Martinez [1] $\mathbf{Z} = \tilde{\mathbf{\Theta}} \mathbf{C} \tilde{\mathbf{B}}^T$	(This Paper) $\mathbf{Z} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Theta}\mathbf{\Theta}^T, \mathbf{B}\mathbf{B}^T)$ or $\|\mathbf{\Theta}^+ \mathbf{X} \mathbf{P}_\perp\|_*$	

Shape Methods. The marginal distribution of a particular shape $\mathbf{x}^t$ (a row t of $\mathbf{X}$ arranged as a column) independent of all other time instants corresponds exactly to shape-only distributions used in prior art, such as the Point Distribution Model (PDM) of Cootes et al. [32], and the shape basis model of Torresani et al. [23]. It follows from the matrix normal model that $\mathbf{x}^t \sim \mathcal{N}(\mathbf{M}^t, \Sigma_{t,t} \mathbf{\Delta})$,, where $\mathbf{\Sigma}_{t,t}$ is the entry (t,t) in $\mathbf{\Sigma}$ and $\mathbf{\Delta}$=$\mathbf{B}\mathbf{B}^T$ is the shape covariance matrix. An equivalent shape covariance is estimated using PCA by Cootes et al., where $\mathbf{B}$ is a shape basis [9], [19], [23], [40].

Similarly, the PND model of Lee et al. [16] is related in the same way save for two distinctions: firstly, the shape covariance in the PND model is restricted to exclude the subspace of small-angle rotations, scaling, and translation of the mean shape (i.e., adding the constraints that $\mathbf{P}_N^T \mathbf{\Delta} = \mathbf{0}$, where $\mathbf{P}_N$ is that subspace, $||\mathbf{m}_{\text{shape}}|| = 1$ and $\mathbf{m}_{\text{shape}} \mathbf{1} = 0$, see [16]); secondly, the shape covariance is rotated into the coordinate system of every frame (i.e., the PND models rotated shapes, whereas we model aligned shapes). While the Procrustean constraints can be incorporated into the MND model, this would prohibit the convex solution presented in Sect. 4.2. Similarly, modeling rotated shapes would make the $\mathbf{H}$ and $\mathbf{J}$ matrices in Eq. 8 become time-dependent, and we would loose the compact Kronecker expression of the covariance.

Spatiotemporal Methods. The model we present is a probabilistic formulation of the shape-trajectory basis models described in [1], [2]. These models describe spatiotemporal sequences as a linear combination of the outer product of a reduced set of trajectory basis vectors and a set of shape basis vectors. They rely on truncation of the basis to achieve compaction, while the probabilistic MND model describes the relative variance of each spatiotemporal mode with the weighting matrices $\mathbf{W}_t$ and $\mathbf{W}_b$. Additionally, the MND allows us to compute a confidence bound on the imputed missing data. We visualize this distribution in Fig. 10 on a facial motion capture sequence from [2].

As with the shape-only model, the PND Markov process (PMP) [24] is very closely related. With the same distinctions about the rotated coordinate space for the shape covariance discussed above, the Markov PND process is essentially the same as the Kronecker-Markov process in Sect. 3.2.3 but with a stationarity constraint $\Phi = \alpha \mathbf{I}$ rather than $\alpha \to 1$. The PMP model is therefore more general, but the trade-off is a non-convex optimization that requires careful initialization and explicitly solving for the parameter α.

[3] See [39] for properties of the Kronecker product.

[4] We use the same trajectory covariance $\mathbf{\Sigma}$ as for the non-rigid component, but, more generally, a different covariance matrix could be used.

4 Convex MAP Reconstruction for the Kronecker-Markov Prior

In this section, we derive convex estimation procedures to recover the most likely dynamic structure $\mathbf{X}$ given the measurements $\mathbf{y}$ using the Kronecker-Markov shape-trajectory prior, $p(\mathbf{y}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{y}|\mathbf{Z},\mathbf{M})p(\mathbf{Z})p(\mathbf{M})$, where the non-rigid and mean components are distributed according to Eqs. (12) and (14).

4.1 Known Distribution Parameters

With known covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta}$, the negative log-likelihood of the MND is quadratic, and inference under an MND prior is straightforward and can be posed as a least-squares problem:

$$\begin{aligned}
&\underset{\mathbf{M},\mathbf{Z}}{\operatorname{argmax}}\; p(\mathbf{y}|\mathbf{Z},\mathbf{M})p(\mathbf{Z})p(\mathbf{M}) = \\
&\quad\underset{\mathbf{M},\mathbf{Z}}{\operatorname{argmin}}\, \sigma^{-2}||\mathbf{y} - \mathbf{O}\,\mathrm{vec}(\mathbf{M}+\mathbf{Z})||_F^2 \\
&\qquad\qquad + \underbrace{\mathrm{tr}\left[\boldsymbol{\Delta}^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}\right]}_{-\log(p(\mathbf{Z}))+c_1} \\
&\qquad\qquad + \lambda\,\underbrace{\mathrm{tr}\left[\mathbf{M}_{\text{trans}}{}^T\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\text{trans}}\right]}_{-\log(p(\mathbf{M}))+c_2}, \qquad (15)
\end{aligned}$$

where λ is a scaling factor related to the mass of the object and the variance of its translational motion.

4.2 Unknown Distribution Parameters

We can approximate the trajectory distribution using a DCT−2 matrix $\boldsymbol{\Sigma}^{-1}$ (Sect. 3.2.3). However, the PDM shape distribution $\boldsymbol{\Delta}$ covariance depends on the object and is typically unknown a priori, and therefore needs to be estimated as well. The least-squares problem of Eq. (15) now becomes bilinear in $\mathbf{Z}$ and $\boldsymbol{\Delta}$, and seemingly non convex.

In the following, we show that there exists a convex solution when we set a hierarchical Wishart covariance [41] prior over $\boldsymbol{\Delta}$. Using the bilinear parameterization (Eq. (13)), $\mathbf{Z} = \boldsymbol{\Theta}\mathbf{C}\mathbf{B}^T$,

$$\begin{aligned}
p(\mathbf{X}|y) &= p(\mathbf{C},\mathbf{B},\mathbf{M}|\mathbf{y}) \propto \\
&\qquad p(\mathbf{y}|\mathbf{C},\mathbf{B},\mathbf{M})p(\mathbf{B}|\mathbf{C})p(\mathbf{C})p(\mathbf{M}). \qquad (16)
\end{aligned}$$

In this parameterization, $p(\mathbf{B}|\mathbf{C})$ is the only prior that remains to be specified.

To obtain a convex solution, we assume that $p(\mathbf{B}|\mathbf{C}) = p(\mathbf{B})$, i.e., the distribution over shape covariance is independent of the particular shape configurations observed. We use a normal prior over the entries of $\mathbf{B} \sim \mathcal{MN}(\mathbf{0},\mathbf{I}_{3P},\mathbf{I}_{3P})$, equivalent to a Wishart prior over the covariance $\boldsymbol{\Delta}$. Intuitively, this captures the low-rank characteristic of shape covariance matrices: that the singular values of the covariance matrix should decrease rapidly (see Sect. 6.1 for an illustration of this prior).

Combining these priors and writing this optimization in terms of the component negative log-likelihoods,

$$\begin{aligned}
&\underset{\mathbf{X}}{\operatorname{argmax}}\; p(\mathbf{X}|\mathbf{y}) = \\
&\quad\underset{\mathbf{M},\mathbf{C},\mathbf{B}}{\operatorname{argmin}}\; \sigma^{-2}||\mathbf{y}-\mathbf{O}\,\mathrm{vec}(\mathbf{M}+\boldsymbol{\Theta}\mathbf{C}\mathbf{B}^T)||_F^2 \\
&\qquad\qquad +||\mathbf{C}||_F^2 + ||\mathbf{B}||_F^2 + \lambda||\boldsymbol{\Theta}^+\mathbf{M}_{\text{trans}}||_F^2 \\
&\quad\text{s.t.}\quad \mathbf{X} = \mathbf{M}+\boldsymbol{\Theta}\mathbf{C}\mathbf{B}^T. \qquad (17)
\end{aligned}$$

This expression is bilinear in $\mathbf{C}$ and $\mathbf{B}$. However, we can transform this bilinear equation into a convex problem using the matrix trace-norm $\|\cdot\|_*$, where $\|\mathbf{R}\|_* = \min_{\mathbf{U},\mathbf{V}}\{\frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{1}{2}\|\mathbf{V}\|_F^2\}$ subject to $\mathbf{R} = \mathbf{U}\mathbf{V}^T$ with $\mathbf{R} \in \mathbb{R}^{m\times n}$, $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$. Mazumder et al. [42] show the equivalence of these two formulations when $r \geq \mathrm{rank}(\mathbf{R})$, and we can transform Eq. (17) by writing[5] $\mathbf{A} = \mathbf{C}\mathbf{B}^T$ and so

$$\begin{aligned}
&\underset{\mathbf{M},\mathbf{A}}{\operatorname{argmin}}\; \sigma^{-2}||\mathbf{y} - \mathbf{O}\,\mathrm{vec}(\mathbf{M}+\boldsymbol{\Theta}\mathbf{A})||_F^2 + \|\mathbf{A}\|_* \\
&\qquad\qquad +\lambda\|\boldsymbol{\Theta}^+\mathbf{M}_{\text{trans}}\|_F^2 \qquad (18)
\end{aligned}$$

By definition, $\mathbf{X}$ is parameterized into a mean shape component, $\mathbf{m}_{\text{shape}} = \frac{1}{F}\mathbf{1}_F^T\mathbf{X}$, a translational component, $\mathbf{M}_{\text{trans}} = \frac{1}{P}\mathbf{X}\mathbf{P}_{\text{trans}}^T$, and the remaining non-rigid component $\mathbf{Z} = \boldsymbol{\Theta}\mathbf{A}$ where

$$\mathbf{Z} = \mathbf{X} - \mathbf{M} = (\mathbf{I}_F - \frac{1}{F}\mathbf{1}_F\mathbf{1}_F^T)\mathbf{X}\underbrace{(\mathbf{I}_{3P} - \frac{1}{P}\mathbf{P}_{\text{trans}}^T\mathbf{P}_{\text{trans}})}_{\mathbf{P}_\perp}. \qquad (19)$$

Finally, note that $\mathbf{1}_F$ is in the left null-space of $\mathbf{Z}$, and right null-space of $\boldsymbol{\Theta}^+$, and so $\mathbf{Z} = \boldsymbol{\Theta}\boldsymbol{\Theta}^+\mathbf{Z}$. We can therefore write the change of variables $\boldsymbol{\Theta}^+\mathbf{X}\mathbf{P}_\perp = \mathbf{A}$ resulting in

$$\begin{aligned}
&\underset{\mathbf{X}}{\operatorname{argmin}}\quad \frac{1}{2\sigma^2}\underbrace{||\mathbf{O}\,\mathrm{vec}(\mathbf{X}) - \mathbf{y}||_2^2}_{\text{observations}} + \underbrace{||\boldsymbol{\Theta}^+\mathbf{X}\mathbf{P}_\perp||_*}_{\text{shape-trajectory prior}} \\
&\qquad\qquad + \underbrace{\lambda\frac{1}{2}||\boldsymbol{\Theta}^+\mathbf{X}\mathbf{P}_{\text{trans}}^T||_2^2}_{\text{translational regularizer}}. \qquad (20)
\end{aligned}$$

Relationship to Trace-norm Methods. The convex MAP minimization of Eq. (20). when using a normal prior over $\mathbf{B}$ can be related to the use of the trace-norm in rigid and non-rigid structure from motion [14], [43]. Note that the shape-trajectory prior component of this objective function is similar to the trace-norm energy term of Dai et al. [14], if we set $\boldsymbol{\Theta}^+$ to the identity. This amounts to assuming that every frame is independent and there exist no temporal correlations. The trace-norm method of Dai et al. can then be interpreted as a prior of non-rigid shape that corresponds to $\mathbf{C}\mathbf{B}^T$ with normal priors over coefficients $\mathbf{C}$ and shape basis $\mathbf{B}$. The effect of this is most easily understood for the case of interpolation: frames (rows) for which all points are missing will be set to zero by the $\|\mathbf{X}\|_*$ penalizer. This effect can result in abrupt changes in the reconstruction, and can be seen in the spiked blue curves in Fig. 7 (right). Making a similar observation, Angst et al. [43] proposed the "generalized trace-norm" for rigid SfM to incorporate temporal smoothness constraints in trace-norm approaches to SfM, resulting in a similar prior term. Compared to the rigid model of Angst et al., our work draws an explicit connection between the row and column spaces of an MND distribution of dynamic 3D structure.

5 Optimization via ADMMs

The objective of Eq. (20) lends itself to optimization by the Alternating Direction Method of Multipliers (ADMM) [44]. We discuss two cases: the convex solution, where the observation matrix $\mathbf{O}$ is fixed (and consequently, the camera rotation matrices are fixed), and a non-convex procedure that additionally optimizes the rotation matrices.

[5] In this case, $m = F$, $n = 3P$, and $r = 3P \geq \mathrm{rank}(\mathbf{A}) = \min(F, 3P)$.

5.1 Fixed Camera Matrices

Let $\mathbf{F} = \mathbf{P}_{\text{trans}} \otimes \mathbf{\Theta}^{+}$ and $\mathbf{G} = \mathbf{P}_{\perp}^{T} \otimes \mathbf{\Theta}^{+}$. Re-writing Eq. (20) in ADMM form (see Boyd et al. [44]),

$$\begin{aligned} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{subject to} \quad & \mathbf{G}\mathbf{x} - \mathbf{z} = \mathbf{0} \\ & f(\mathbf{x}) = \frac{1}{2\sigma^2}||\mathbf{O}\mathbf{x} - \mathbf{y}||_2^2 + \frac{\lambda}{2}||\mathbf{F}\mathbf{x}||_2^2 \\ & g(\mathbf{z}) = ||\operatorname{unvec}(\mathbf{z})||_* \end{aligned} \tag{21}$$

where $\mathbf{x} = \operatorname{vec}(\mathbf{X})$, $f(\mathbf{x})$ and $g(\mathbf{z})$ are convex, $\operatorname{unvec}(\cdot)$ reshapes the argument into the desired matrix[6] of size $F \times 3P$, and σ is the observation noise variance. Written in this more general form, we identify this as a *trace-norm regularized least squares* problem. The ADMM method iteratively updates the variables in two steps, according to the following subproblems:

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\text{argmin}} \left(f(\mathbf{x}) + \frac{\rho}{2}||\mathbf{G}\mathbf{x} - \mathbf{z}^k + \mathbf{u}^k||_2^2 \right) \tag{22}$$

$$\mathbf{z}^{k+1} = \underset{\mathbf{z}}{\text{argmin}} \left(g(\mathbf{z}) + \frac{\rho}{2}||\mathbf{G}\mathbf{x}^{k+1} - \mathbf{z} + \mathbf{u}^k||_2^2 \right) \tag{23}$$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \left(\mathbf{G}\mathbf{x}^{k+1} - \mathbf{z}^{k+1} \right) \tag{24}$$

with $\mathbf{u}$ the scaled dual variables of the augmented Lagrangian.

The $\mathbf{x}^{k+1}$ update Eq. (22) is a least-squares problem and is readily solvable. The $\mathbf{z}^{k+1}$ update Eq. (23) involves the nuclear norm and is more difficult to solve, but there exists a closed form solution [44] for problems of the form

$$\operatorname{prox}_{\lambda}(\mathbf{W}) = \underset{\mathbf{Z} \in \mathbb{R}^{m \times n}}{\text{argmin}} \frac{1}{2}||\mathbf{Z} - \mathbf{W}||_2^2 + \lambda||\mathbf{Z}||_*. \tag{25}$$

Define the *shrinkage* or *soft-thresholding* operator,

$$s_{\lambda}(x) = max(x - \lambda, 0) - max(0, -x - \lambda), \tag{26}$$

which we will apply entry-wise to vectors. The solution to this type of problems is then $\operatorname{prox}_{\lambda}(\mathbf{W}) = \mathbf{S}_{\lambda}(\mathbf{W})$, where the *matrix soft-thresholding operator* $\mathbf{S}_{\lambda}(\mathbf{W})$ will be

$$\mathbf{S}_{\lambda}(\mathbf{W}) = \mathbf{U}\mathbf{\Sigma}_{\lambda}\mathbf{V}^T, \tag{27}$$

where $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, and $\mathbf{\Sigma}_{\lambda}$ is diagonal with $(\mathbf{\Sigma}_{\lambda})_{ii} = s_{\lambda}(\mathbf{\Sigma}_{ii})$, the soft-thresholded singular values of $\mathbf{W}$ [44]. The solution to Eq. (23) is then

$$\begin{aligned} \mathbf{Z}^{k+1} &\leftarrow \operatorname{prox}_{\frac{1}{\rho}} \left(\operatorname{unvec}(\mathbf{G}\mathbf{x}^{k+1} + \mathbf{u}^k) \right) \\ \mathbf{z}^{k+1} &\leftarrow \operatorname{vec}\left(\mathbf{Z}^{k+1} \right), \end{aligned}$$

5.2 Optimizing the Camera Matrices

The ADMM procedure described in the preceding section suggests a simple way to incorporate the estimation of camera (or object) rotation into the optimization, at the cost of making the problem non-convex. For fixed camera matrices, we wrote the observation cost in matrix form as $\frac{1}{2\sigma^2}||\mathbf{O}\mathbf{x} - \mathbf{y}||$, where the matrix $\mathbf{O}$ was assumed to be constant. This expression only appears in the $\mathbf{x}^{k+1}$ update equation, which we can now rewrite more generally as

$$\{\mathbf{x}^{k+1}, \mathbf{p}^{k+1}\} = \underset{\mathbf{x},\mathbf{p}}{\text{argmin}} \left(f(\mathbf{x}, \mathbf{p}) + \frac{\rho}{2}||\mathbf{G}\mathbf{x} - \mathbf{z}^k + \mathbf{u}^k||_2^2 \right) \tag{28}$$

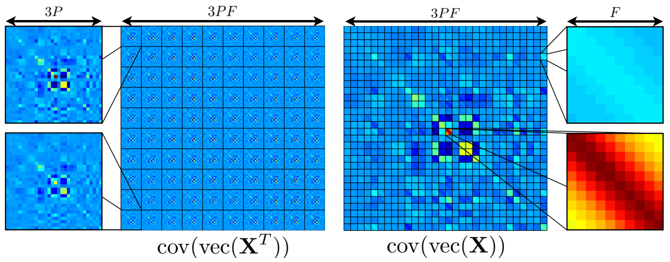


Fig. 3. Empirical spatiotemporal covariance matrix for a subset of face motion capture data (P=10, F=10), shown for two possible vectorizations of the matrix $\mathbf{X}$. (Left) The row-major arrangement shows blocks that are approximately scaled versions of the spatial or row covariance. (Right) The column-major arrangement shows more clearly the trajectory or column covariance.

where $\mathbf{p}^k$ is the current estimate of the camera parameters, and we redefine f as

$$f(\mathbf{x}, \mathbf{p}) = \mathcal{P}(\mathbf{x}, \mathbf{p}, \mathbf{y}) + \frac{\lambda}{2}||\mathbf{F}\mathbf{x}||_2^2 \tag{29}$$

where the function $\mathcal{P}$ measures the total observation cost and can be any smooth differentiable function. Without going into the verbose particulars of how to index each observation and its corresponding camera parameters, we parameterize each camera at each time instant as an axis-angle rotation vector and a translation vector (when camera and object motion are not ambiguous). The function $\mathcal{P}$ computes the reprojection error residuals for each of the observed points. We solve the ADMM $\mathbf{x}^{k+1}$ update Eq. (28) for both $\{\mathbf{x}, \mathbf{p}\}$ using Levenberg-Marquadt and the ceres-solver [45] within the ADMM framework. As in [44], we use the previous $\mathbf{x}^k$ and $\mathbf{p}^k$ to warm-start the optimization, and only run a few iterations (5, in our experiments) for each $\mathbf{x}^{k+1}$ update.

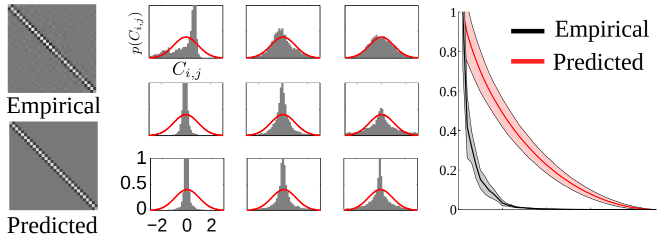
5.3 Implementation details

The choice of the augmented Lagrangian parameter ρ greatly affects the convergence. We follow the heuristic described in Boyd et al. [44], halving or doubling ρ when the r-norm and s-norm ratios are greater than 2, and we use a maximum of 500 iterations with the stopping criteria described in Section 3.3.1 of [44].

The algorithm relies on two crucial operations: solving a large linear system of equations, and computing an SVD. For typical matrix completion problems, it is usually assumed that computing the SVD is the most time consuming operation. For our problem, it is typically the case that solving the system of equations is more difficult: our problem size is one with $3FP$ unknowns, and the matrices $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{O}$ are not necessarily sparse. The matrices $\mathbf{G}$ and $\mathbf{F}$ do have a limited number of non-zero entries: $\mathbf{\Theta}^{+}$ is the forward differences matrix (each residual involves at most two time instants) and $\mathbf{P}_{\text{trans}}$ has bandwidth $3P$ (each row involves summing over P points at a single time instant).

To solve each linear problem, we therefore either pre-compute the Cholesky-factors for quicker per-iteration solves, or, if the factors are too large to build, we use an iterative solver relying on matrix-vector products. In particular, we use Matlab's `lsqr`: for a problem of size F=2000 frames and P=49 points, there are 294000 unknowns, and each linear solve dominates the ADMM iteration time and takes 4 to 8 seconds on an Intel Core i7 at 2.7GHz, for a total runtime of 38 min. or 1.1 s per frame.

[6] The development is valid even if $\mathbf{Z}$ is not the same size as $\mathbf{X}$. In particular, the two arrangements described by Dai et al. [14], $3F \times P$ and $F \times 3P$, are options to consider.



(a) Trajectory (b) Coefficients (c) Shape

Fig. 4. (Left) Empirical and predicted model parameter distributions. (a) Top, the empirical trajectory precision matrix. Below, the DCT−2 matrix from Sect. 3.2.3. (b) Each plot corresponds to a coefficient $C_{i,j}$ in the matrix $\mathbf{C}$. The red curve shows the predicted standard normal pdf, the histogram shows the empirical distribution. (c) Distribution of singular values for empirical shape covariances (black), compared to the predicted fall-off induced by $p(\mathbf{B})$ (red).

6 Results

6.1 Validation on Natural Motions

We validate the proposed distribution and the four components of our model by computing statistics on a large set of natural motions. We use the CMU Motion Capture database, where we subsample the data to retain point tracks for 15 joint locations on the body, yielding $N = 5402$ 30-frame sub-sequences $\mathbf{X}_n$ which we also align using Procrustes analysis and center around the mean.

I. Kronecker-Markov Covariance Structure. (Sect. 3.2)
Fig. 2(left) shows the empirical sample covariance matrix $\frac{1}{N}\sum_n \operatorname{vec}(\mathbf{X}_n)\operatorname{vec}(\mathbf{X}_n)^T$ computed on the full set of sequences. On the right, we show the covariance associated with the matrix normal distribution, i.e., $\mathbf{\Delta} \otimes \mathbf{\Sigma}$, where $\mathbf{\Delta}$ is computed[7] as the covariance of the rows $\mathbf{\Delta} = \frac{1}{NF}\sum_n \mathbf{X}_n^T\mathbf{X}_n$, and $\mathbf{\Sigma} = \frac{1}{vN3P}\sum_n \mathbf{X}_n\mathbf{X}_n^T$, with $v = \frac{1}{3P}\operatorname{tr}(\mathbf{\Delta})$. Note that this separable approximation captures most of the structure and energy in the covariance using far fewer parameter than a full covariance matrix. Fig. 3 shows the empirical covariance matrix for a dataset of face motion capture data (an example frame can be seen in Fig. 10), for a subset of $P = 10$ points and $F = 10$ frames at 30Hz sampled from 158s of data.

II. Analytical Trajectory Distribution. (Sect. 3.2.1, 3.2.3)
Fig. 4(a) shows that the empirical precision matrix computed over trajectories (the inverse of the sample covariance, $\mathbf{\Sigma}^{-1}$) closely resembles the regularizer predicted by the DCT−2 matrix. Most correlations in the data are captured by the analytical model.

III. Distribution of Coefficients. (Sect. 3.3)
The matrix normal model assumes a standard normal distribution over the latent coefficients, i.e., $C_{i,j}{\sim}\mathcal{N}(0,1)$. Given a large set of natural motion sequences, we can verify the accuracy of this assumption by fitting the model coefficients $\mathbf{C}_n \in \mathbb{R}^{F\times 3P}$ to each sequence $\mathbf{X}_n$, and plotting the resulting histogram of coefficient values. Fig. 4(b) shows that the empirical distribution can be more spiked, closer to Laplacian or Cauchy in practice.

IV. Hierarchical Prior on Shape Covariance. (Sect. 4.2)
We sample shape covariance matrices from the prior $\mathbf{B}{\sim}\mathcal{MN}(\mathbf{0}, \mathbf{I}_{3P}, \mathbf{I}_{3P})$ and compute their singular values (SVs). Fig. 4(c) compares the energy fall-off in SVs from sampled matrices to that of empirically computed covariance matrices. The plot shows the mean SVs and ±3 standard deviations. The fall-off in energy of the singular values by the prior on $\mathbf{B}$ is not as quick as observed in data, but the choice allows convex optimization.

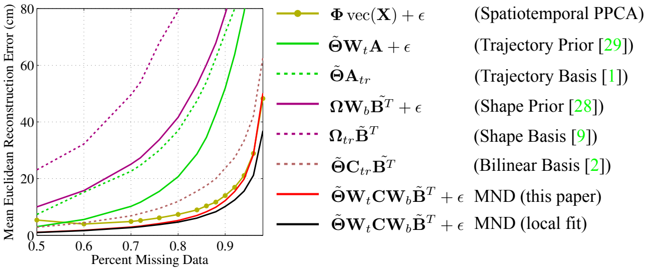


Fig. 5. Inference of missing data with learned distribution parameters. Subscript tr indicates an orthonormal truncation method.

6.2 Missing Data in Motion Capture

To characterize the models' resilience to the patterns of missing data encountered in dynamic reconstruction, we simulate different patterns of occlusion, and we decouple the problem from that of projection loss and reconstructibility [21] by studying inference in 3D. The task is to infer or complete a dynamic 3D point cloud from a reduced set of 3D observations—a practical application would be filling in missing markers in motion capture data. We use the observation model $\mathbf{O}_{\text{miss}}$ as per Sect. 3.1.

I. Known Distribution Parameters. (Sect. 4.1)
When 3D training data is available, we can learn the parameters for MND distribution and perform inference with Eq. (15). We compare with the models corresponding to *probabilistic* and *truncated* versions of shape, trajectory, and shape-trajectory distributions (summarized in Table 1). Additionally, we evaluate against a probabilistic PCA model trained on the vectorized spatiotemporal sequences, i.e., $\mathbf{y} = \Phi\operatorname{vec}(\mathbf{X}) + \epsilon$. We report mean 3D error in Fig. 5. As a reference, the error incurred when using the mean shape at every frame as an estimation is ∼175 cm.

For this experiment, we use data from the CMU Motion Capture database. We take 50 random sequences of 20 s in duration, sample them at 30 Hz and Procrustes align and mean center them. There are 31 markers on the body, and we subdivide each sequence into $1s$ chunks resulting in $F{=}30$ and $P{=}31$. We train all models on 49 of the sequences, and test on a random $1s$ segment of the left out sequence. We simulate random occlusion on a percentage of the points and report the average over 50 trials. For the probabilistic models, we set the noise variance to 0. For models relying on truncation of the basis, we sweep over all possible levels of truncation and pick the best number *a posteriori*. Note that the MND model with factored covariance performs equally well or better than PCA on the vectorized sequences, while requiring less training data (50 times less in this experiment). This allows us to train a *local* model only on subsequences neighboring the test data; the model is more specific and results in lower error.

II. Unknown Distribution Parameters. (Sect. 4.2)
When no training data is available, we perform inference with Eq. (20). In Figure 6, we compare our approach with three different priors: (1) a trajectory-only prior, (2) a trace-norm prior, and (3) an additive combination of the trace-norm and trajectory priors. We assume Gaussian observation noise with standard deviation

[7] ML estimates of the parameters for noiseless data can be obtained using a "flip-flop" algorithm [38], but in practice we obtained better results with the described procedure.

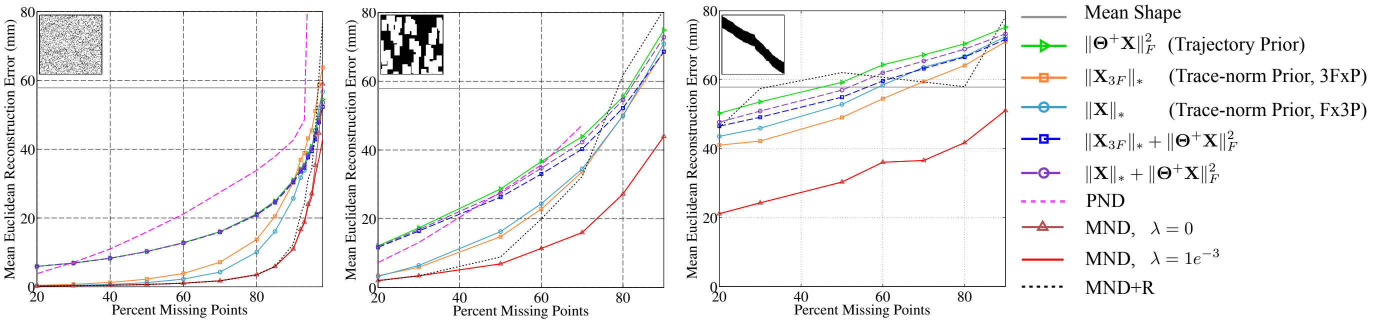


Fig. 6. Inferring missing data under three different occlusion patterns when the shape distribution is unknown. The graphs show mean Euclidean error in the reconstruction under the occlusion models discussed in Section 6.2. The bottom two results correspond to the method of Sect. 4.2. We investigate two different arrangements for the data matrix, $3F \times P$ and $F \times 3P$, which capture different correlations of the data. For this experiment, $3F \times P$ usually offered better performance, which we report on our method. The data is from dense human motion capture originally intended to measure non-rigid skin deformation while running in place.

σ=1 mm for all methods. We use dense motion capture data from Park and Hodgins [46]. The sequences are captured at 120 Hz with a dense spatial sampling across the body. We downsample by four spatially and temporally, yielding a point cloud of 118 points at 30 Hz across 162 frames. We measure reconstruction error as mean Euclidean distance over all points, under three different patterns of missing data: **(a) Random:** We occlude points (x,y,z) at random until we achieve a percentage of missing data. This pattern of occlusion is not common in practical situations, but is of interest theoretically: theorems about performance of the trace norm as an approximation to the rank are based on this pattern. **(b) Detection loss:** We model detection loss by occluding spatially proximal points during 1 second durations (30 frames), simulating an occlusion. We superimpose these simulated occlusions to increase the amount of missing data. **(c) Correspondence loss:** We duplicate every point trajectory. Each of the resulting trajectories is observable during a non-overlapping duration, resulting in a pattern similar to that observed when tracking from visual features.

The resulting occlusion patterns are shown as insets in Fig. 6. We note that *correspondence loss* results in a much harder problem. Independently of the occlusion pattern, the proposed approach shows improved results. The performance drop when additionally optimizing rotations (Sect. 5.2) is explained by the nature of the data, which contains almost no rotation and very little translation. We expect the reduced performance of PND[8] in this experiment is for the same reason. Because there is no temporal smoothness constraint on the rotations, for high percentages of missing data the rotation estimation overfits the observed points.

6.3 Non-rigid Structure from Motion

We compare the performance of our time-varying point cloud reconstruction method using Eq. (20) on a standard set of structure from motion sequences, where the only data loss is from projection. In Table 2, we report normalized mean 3D error as computed in [1] for four methods, (1) KSTA [1], a non-linear kernelized shape-trajectory method, (2) Dai et al. [14], (3) a trajectory-only prior, (4) PND [16], and (5) our approach. For our methods (MND and MND+R), we compute the camera matrices as in Dai et al. [14] [9], and set σ=1 and λ=0 (the sequences are translationally mean-centered). For Dai et al. and KSTA, the optimal parameter k was chosen for each test.

TABLE 2
Comparison on zero-noise standard NRSfM sequences using normalized mean 3D error [1], [14].

Dataset	KSTA	Dai	Traj.	PND	MND	MND+R
Drink	0.0156	0.0266	0.0102	0.0868	**0.0099**	0.0898
Pick-up	0.2322	0.1731	0.1707	0.1188	0.1707	**0.0935**
Yoga	0.1476	0.1150	0.1125	**0.1040**	0.1114	0.1084
Stretch	**0.0674**	0.1034	0.0972	0.0908	0.0940	0.1213
Dance	0.2504	0.1842	0.1385	0.6394	**0.1347**	0.1598
Face2	0.0339	0.0303	0.0408	0.0306	**0.0299**	0.0333
Walking2	**0.1029**	0.1298	0.3111	0.2948	0.1615	0.705
Shark2	**0.0160**	0.2358	0.1380	0.6166	0.1297	0.0684

We also evaluated the robustness with respect to missing data compared to the Procrustean Normal Distribution (PND) of Lee et al. [16]. We report these results in Table 3 using the metric used in [16]. Note that this metric is different from that used in [14], computing normalized mean 3D error on the mean-centered trajectories including camera motion (see [16]). When using the camera matrix optimization procedure of Sect. 5.2 (MND+R), we see similar or a slight improvement in performance for most sequences. However, the improvements using MND+R are much smaller than we expected; and the solutions have larger error variance. We attribute this to two factors: (1) the estimated rotations are completely unconstrained and may not be smooth, and (2) the optimization procedure of Sect. 5.2 is no longer being convex and the solution can stagnate at a poor local minimum.

6.4 Multiview Dynamic Reconstruction

We perform a qualitative evaluation of the method of Sect. 4.2 on a dynamic reconstruction sequence from Park et al. [20]. This sequence is observed very sparsely by multiple cameras taking snapshots of the scene at a rate of around 1 per second. We aim to reconstruct the original motion at 30 Hz. Because the observations

[8] The original code was modified to compensate for translations of the object. Large amounts of missing data also proved problematic, resulting in numerical issues for some data points in the graph.

[9] For KSTA [1], the camera matrices are computed as per Akhter et al. [19]. Our method shows improved performance on 5 of 8 sequences, while the non-linear KSTA method can achieve better performance on some sequences. The implementation of Dai et al. and KSTA was provided by the respective authors.

TABLE 3
Comparison with the PND method of Lee et al. [16] for 0%, 30%, and 60% missing data. We show results using fixed cameras (MND), and optimized using the algorithm of Section 5.2 (MND+R).

0% missing data					30% missing data				60% missing data			
Name	PND	PMP	MND	MND+R	Name	PND	MND	MND+R	Name	PND	MND	MND+R
yoga	0.0140	**0.0128**	0.0137	0.0145	yoga	0.0324	0.0430	**0.0179**	yoga	0.0277	0.0519	**0.0246**
pickup	0.0372	**0.0127**	0.0154	0.0142	pickup	0.0366	**0.0141**	0.0145	pickup	0.0267	0.0675	**0.0161**
stretch	0.0156	0.0124	**0.0116**	0.0170	stretch	0.0151	**0.0138**	0.0173	stretch	0.0308	0.0447	**0.0236**
drink	0.0037	**0.0018**	0.0021	0.0022	drink	0.0055	0.0027	**0.0024**	drink	0.0169	0.0519	**0.0051**
dance	0.1834	0.1278	**0.1035**	0.1205	dance	0.1768	**0.1020**	0.1205	dance	0.1512	**0.1072**	0.1212
face	**0.0165**	0.0166	0.0177	0.0195	face	**0.0177**	0.0200	0.0251	face	**0.0208**	0.0279	0.0487
walking	0.0465	**0.0424**	0.1360	0.3756	walking	**0.0459**	0.1256	0.3567	walking	**0.0608**	0.1293	0.3564
jaws	0.0134	**0.0099**	0.0882	0.0687	jaws	**0.0154**	0.0825	0.0696	jaws	**0.0139**	0.0813	0.0713

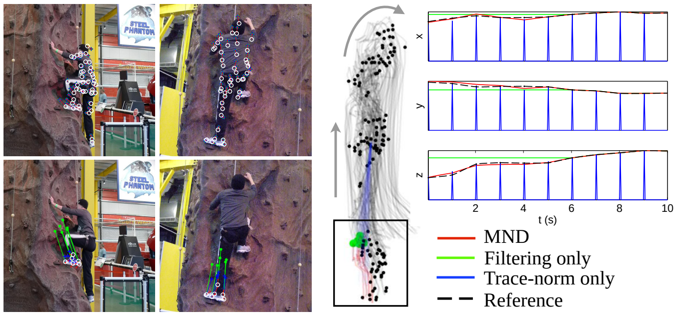


Fig. 7. Multiview reconstruction on the "Rock Climbing" sequence from [20]. Annotated labels are shown in white. (Left) Qualitative comparison. The top row shows a result on the full data (104 camera snapshots of 45 points). All methods perform similarly for fully observed frames. The bottom row shows a result on a simulated occlusion (see text). (Center) Reconstructed 3D trajectories of the points, side view of the climbing wall. The arrows denote the direction of motion of the climber. (Right) x,y,z-plot of the mean trajectories of the imputed points.

are now 2D image measurements under 3D-to-2D perspective projection, we use an observation model $\mathbf{O}_{\text{proj}}$ corresponding to a matrix re-arrangement of the observation model described in [20].

Fig. 7 shows reconstructions on a climbing sequence, where we have simulated occlusion of the left foot. Because ground truth is not available, to obtain a reference reconstruction we first run all methods on the full data and average the resulting structure.

This result is shown in black. Fig. 7(left) shows a simulated occlusion of the points on the left foot during the first 6 seconds of the sequence. The trajectory-only prior $\|\boldsymbol{\Theta}^{+}\mathbf{X}\|_F^2$ gives a smooth solution, but the foot is not at a coherent location with respect to the body. Conversely, all trace-norm based methods are able to infer the position of the left foot (bottom row of images) fairly plausibly in the shape domain. However, when we look at the temporal domain Fig. 7(right), we observe that the trace-norm penalization $\|\mathbf{X}\|_*$ results in temporal artifacts—rows in the matrix with no observations are set to zero. This model is not adequate for data interpolation: as observed in the matrix completion literature, the non-uniformity of the missing entries (as happens when interpolating a sparsely observed signal at 30 Hz) negatively affects the performance of trace-norm methods. Our method is able to combine both properties and achieve a smoother interpolation while maintaining a low-rank structure.

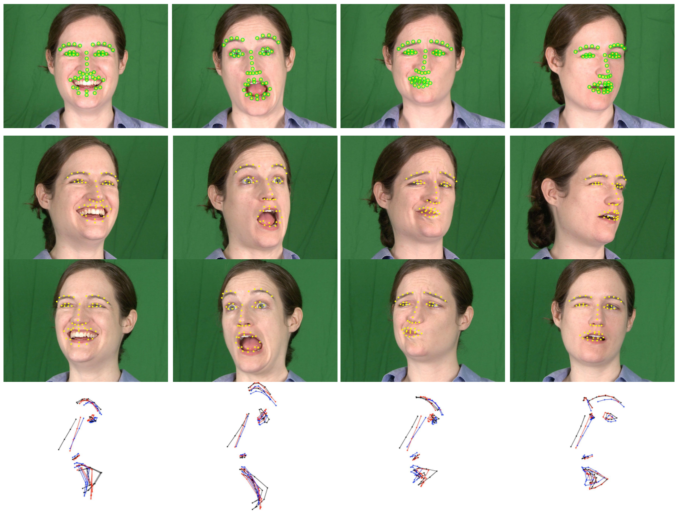


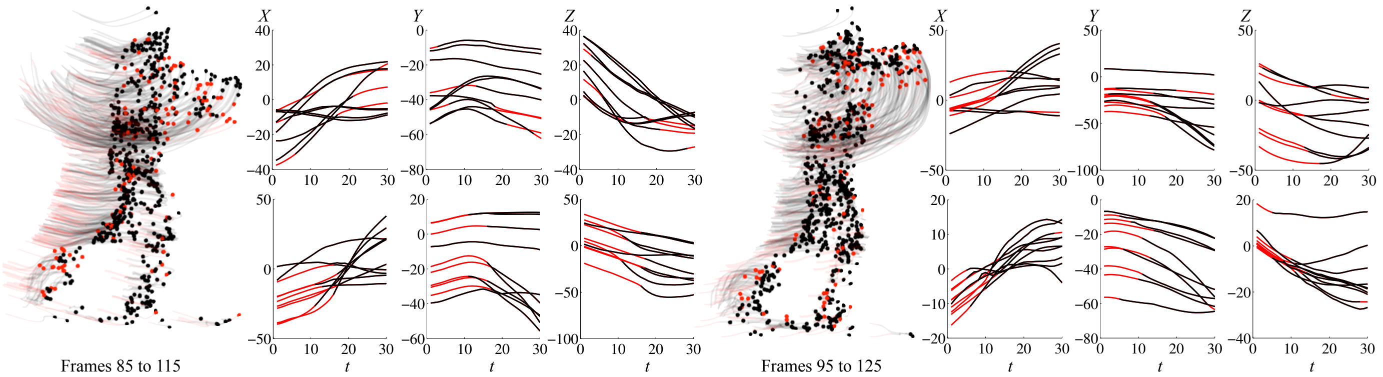
Fig. 8. Reconstructing a dynamic face from a frontal view. The top row shows frames from a video with superimposed detected 2D landmarks (green circles). We reconstruct the face in full 3D using Eq. (20) and show the reprojection onto three other (held out) views for comparison (yellow). Bottom: ground truth (black), MND (red), trace-norm (blue).

6.5 Monocular reconstruction

In Fig. 8 we show a 3D point cloud reconstruction example from a frontal view of a face using 2D landmark detections provided by IntraFace [47]. The original video is around ∼1500 frames long, which we reconstruct simultaneously. Only a subset of frames is shown here. We directly use the model of Sect. 4.2 and build an observation matrix $\mathbf{O}_{\text{ortho}}$ using the head pose estimation matrices provided by IntraFace. Our method recovers a time-varying 3D point cloud of the face, which we can project onto three other views (not used during reconstruction) to evaluate the accuracy. As a quantitative comparison, the ground truth was computed by running the face detector on all views and triangulating the position of each point. The mean 3D error after Procrustes alignment to the ground truth shape was 3.3 mm for MND ($\lambda{=}1e^{-3}$), compared to 3.8 mm for the trace-norm prior (choosing the best weight $\lambda{=}0.025$), and 5.4 mm for MND+R.

6.6 3D Time-varying Point Cloud Reconstruction

In Fig. 9 we show a reconstruction of the baseball sequence acquired by Joo et al. [48]. The input is a set of 3D point trajectories obtained from a multi-camera system. Each trajectory is only



Frames 85 to 115

Frames 95 to 125

Fig. 9. Reconstructing a baseball motion sequence. Black lines indicate observed points, red lines are inferred trajectories. Two motion trail diagrams of 30-frame overlapping parts of a baseball swing are shown. The graphs show a close up reconstruction for different subsets of the points.

partially observed (i.e., once a point cannot be tracked forwards or backwards, its coordinates in subsequent frames are missing). These sequences are 30-frames in duration and have around ∼800 points, which where occluded on average ∼15% of the time. The goal is to obtain complete trajectories for the entire duration of the video. Here, we show two qualitative reconstructions for two overlapping 30-frame subsets of these sequences. The graphs show the trajectories for subsets of points. Note how the recovered trajectories are smooth, and motion occurs in groups because of the low-rank effect of the shape prior.

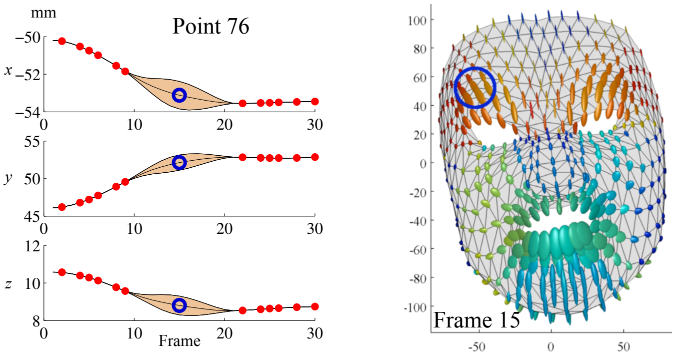


Fig. 10. The matrix normal model allows us to compute the expected value and spatiotemporal covariance of missing data. For this 30 frame sequence, points have been removed completely from frames 10–20. Observed points are marked by red dots. We infer missing values and visualize the mean and 95% confidence bound.

7 Conclusions

We have identified the Kronecker-Markov structure of the covariance of time-varying 3D point cloud data and presented a generative, probabilistic model based on the MND that explains this pattern. The model unifies a number of shape and trajectory models, both probabilistic and algebraic, used in prior art. When training data is available, the prior is easy to use in a least-squares framework and greatly outperforms using either shape or trajectory models independently.

When no training data is available, we show how a connection between the MND and the trace norm leads to a convex MAP objective for missing data reconstruction. The advantage of our convex method is that finding a good solution to the shape factorization problem is guaranteed—however, this comes at the expense of employing a prior over shape covariance that is not as concentrated as observed in practice (see Fig. 4(c)), and not being able to optimize rotations within the same convex framework. Determining under precisely which conditions a generalized trace norm regularization implies a Kronecker-Markov covariance structure is a possible direction of future work. Conversely, the PND-based optimization procedures, particularly the closely related Markov PND [24] prior, show very good results in practice, despite requiring a non-convex optimization and being sensitive to initialization. Ideally, a combination of the properties of both models would result in a stable, highly accurate prior for spatiotemporal data that can be estimated reliably. Along these lines Cabral et al. [49] have shown that under certain conditions a bilinear factorization approach using a non-convex procedure can still converge to the global optimum, and that this optimization can also be simpler and faster. This seems to suggest that, at least for 3D dynamic reconstruction problems, procedures for finding good solutions with guarantees that are laxer than convexity might be found.

References

[1] P. Gotardo and A. Martinez, “Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion,” *PAMI*, 2011.

[2] I. Akhter, T. Simon, I. Matthews, S. Khan, and Y. Sheikh, “Bilinear spatiotemporal basis models,” *ACM Transactions on Graphics*, 2012.

[3] D. Terzopoulos, A. Witkin, and M. Kass, “Constraints on deformable models: Recovering 3d shape and nonrigid motion,” *Artificial Intelligence*, 1988.

[4] D. Metaxas and D. Terzopoulos, “Shape and nonrigid motion estimation through physics-based synthesis,” *PAMI*, 1993.

[5] A. Pentland and B. Horowitz, “Recovery of nonrigid motion and structure,” *PAMI*, 1993.

[6] J. Taylor, A. Jepson, and K. Kutulakos, “Non-rigid structure from locally-rigid motion,” *CVPR*, 2010.

[7] M. Salzmann and R. Urtasun, “Physically-based motion models for 3d tracking: A convex. formulation,” *ICCV*, 2011.

[8] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *IJCV*, 1992.

[9] C. Bregler, A. Hertzmann, and H. Biermann, “Recovering non-rigid 3d shape from image streams,” *CVPR*, 2000.

[10] M. Brand, “A direct method for 3d factorization of nonrigid motion observed in 2d,” *ICCV*, 2005.

[11] J. Yan and M. Pollefeys, “A factorization-based approach to articulated motion recovery,” *CVPR*, 2005.

[12] R. Vidal and D. Abretske, “Nonrigid shape and motion from multiple perspective views,” *ECCV*, 2006.

[13] J. Fayad, A. Del Bue, L. Agapito, and P. Aguiar, “Non-rigid structure from motion using quadratic deformation models,” *BMVC*, 2009.

[14] Y. Dai, H. Li, and M. He, “A simple prior-free method for non-rigid structure-from-motion factorization,” *CVPR*, 2002.

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

[16] M. Lee, J. Cho, C.-H. Choi, and S. Oh, “Procrustean normal distribution for non-rigid structure from motion,” in *CVPR*, 2013, pp. 1280–1287.

[17] J. Cho, M. Lee, and S. Oh, “Complex non-rigid 3d shape recovery using a procrustean normal distribution mixture model,” *IJCV*, 2016.

[18] H. Sidenbladh, M. Black, and D. Fleet, “Stochastic tracking of 3d human figures using 2d image motion,” *ECCV*, 2000.

[19] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, “Nonrigid structure from motion in trajectory space,” *NIPS*, 2008.

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

[21] J. Valmadre and S. Lucey, “A general trajectory prior for non-rigid reconstruction,” *CVPR*, 2012.

[22] S. Olsen and A. A. Bartoli, “Implicit non-rigid structure-from-motion with priors,” *Journal of Mathematical Imaging and Vision*, 2008.

[23] L. Torresani, A. Hertzmann, and C. Bregler, “Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors,” *PAMI*, 2008.

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

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

[26] R. Angst and M. Pollefeys, “A unified view on deformable shape factorizations,” *ECCV*, 2012.

[27] A. Bue, X. Llad, and L. Agapito, “Non-rigid face modelling using shape priors,” in *Analysis & Modelling of Faces & Gestures*, 2005.

[28] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, ser. Monographs on Statistics and Applied Probability. London: Chapman & Hall, 2005, vol. 104.

[29] S. Olsen and A. A. Bartoli, “Using priors for improving generalization in non-rigid structure-from-motion,” *BMVC*, 2007.

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

[31] N. Ahmed, T. Natarajan, and K. Rao, “Discrete cosine transform,” *IEEE Transactions on Computers*, 1974.

[32] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models, their training and application,” *CVIU*, 1995.

[33] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3d faces,” *SIGGRAPH ’99*, 1999.

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

[35] A. Agudo, L. Agapito, B. Calvo, and J. Montiel, “Good vibrations: A modal analysis approach for sequential non-rigid structure from motion,” in *CVPR*, 2014.

[36] G. Strang, “The discrete cosine transform,” *SIAM review*, 1999.

[37] G. Allen and R. Tibshirani, “Transposable regularized covariance models with an application to missing data imputation,” *Annals of Applied Statistics*, 2010.

[38] P. Dutilleul, “The mle algorithm for the matrix normal distribution,” *Statistical Computation and Sim.*, 1999.

[39] A. J. Laub, *Matrix Analysis For Scientists And Engineers*. Society for Industrial and Applied Mathematics, 2004.

[40] J. Xiao, J. Chai, and T. Kanade, “A closed-form solution to non-rigid shape and motion recovery,” *ECCV*, 2004.

[41] C. R. Rao, *Linear Statistical Inference and its Applications*. Wiley-Interscience, 1973.

[42] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization for learning large incomplete matrices,” *JMLR*, 2010.

[43] R. Angst, C. Zach, and M. Pollefeys, “The generalized trace-norm and its application to structure-from-motion problems,” *ICCV*, 2011.

[44] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, 2011.

[45] S. Agarwal, K. Mierle, and Others, “Ceres solver,” http://ceres-solver.org.

[46] S. I. Park and J. K. Hodgins, “Data-driven modeling of skin and muscle deformation,” *ACM Transactions on Graphics*, 2008.

[47] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, “Intraface,” in *FG*, 2015, pp. 1–8.

[48] H. Joo, H. Park, and Y. Sheikh, “Optimal visibility estimation for large-scale dynamic 3d reconstruction,” *CVPR*, 2014.

[49] R. Cabral, F. D. L. Torre, J. P. Costeira, and A. Bernardino, “Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition,” *ICCV*, 2013.

Tomas Simon Tomas Simon is a Ph.D. student in computer vision at Carnegie Mellon University, working with Yaser Sheikh and Iain Matthews. He holds a B.Sc. in Telecommunications from Universidad Politcnica de Valencia and an M.S. in Robotics from Carnegie Mellon University, which he obtained while working at the Human Sensing Lab under the supervision of Fernando De la Torre. His research interests lie mainly in computer vision and graphics, especially as applied to the modeling of face and body motion.

Jack Valmadre Jack Valmadre is currently a post-doc with Philip Torr at the University of Oxford. Within computer vision, his research interests include object tracking, correspondence, 3D non-rigid reconstruction, Fourier techniques and deep learning. Jack studied mechatronic engineering at the University of Queensland in Australia and obtained his PhD from Queensland University of Technology. His doctoral work was done at CSIRO under the supervision of Simon Lucey.

Iain Matthews Iain Matthews is a Research Scientist at Oculus Research working on social virtual reality. His research interests include computer vision and facial tracking, modeling, and animation. Iain received a BEng degree in electronic engineering and a PhD in computer vision from the University of East Anglia, UK. He then joined Carnegie Mellon University, first as a post-doctoral fellow then as faculty in the Robotics Institute. In 2006 he spent two years in New Zealand at visual effects company Weta Digital creating the facial motion capture system for the movies Avatar and Tintin. He joined the newly formed Disney Research Pittsburgh in 2008 to lead the computer vision group, and in 2013 he became the Associate Director of Disney Research Pittsburgh.

Yaser Sheikh Yaser Sheikh is an Associate Professor at the Robotics Institute at Carnegie Mellon University, with appointments in the Mechanical Engineering Department and the Quality of Life Technology Center. He also heads Oculus Research Pittsburgh, a Facebook Research lab focused on virtual reality research. He has served as a senior committee member at ICCP (2011), SIGGRAPH (2013, 2014), CVPR (2014, 2015), and ICRA (2014, 2015), and served as an Associate Editor of CVIU. He has won the Honda Initiation Award (2010), best paper awards at SAP (2012), WACV (2012), SCA (2010), and ICCV THEMIS (2009), and the Hillman Fellowship for Excellence in Computer Science Research (2004). Yaser Sheikh received his PhD in 2006 from the University of Central Florida, advised by Mubarak Shah and completed a postdoctoral fellowship in 2008 at Carnegie Mellon University under the mentorship of Takeo Kanade.