Separable Spatiotemporal Priors for Convex Reconstruction of Time-Varying 3D Point Clouds

Tomas Simon,[1] Jack Valmadre,[2,3] Iain Matthews,[4,1] and Yaser Sheikh[1]

[1]Carnegie Mellon University
[2]Queensland University of Technology, Australia
[3]Commonwealth Scientific and Industrial Research Organisation, Australia
[4]Disney Research Pittsburgh
{tsimon, yaser, iainm}@cs.cmu.edu, j.valmadre@qut.edu.au

Abstract. Reconstructing 3D motion data is highly under-constrained due to several common sources of data loss during measurement, such as projection, occlusion, or miscorrespondence. We present a statistical model of 3D motion data, based on the Kronecker structure of the spatiotemporal covariance of natural motion, as a prior on 3D motion. This prior is expressed as a matrix normal distribution, composed of separable and compact row and column covariances. We relate the marginals of the distribution to the shape, trajectory, and shape-trajectory models of prior art. When the marginal shape distribution is not available from training data, we show how placing a hierarchical prior over shapes results in a convex MAP solution in terms of the trace-norm. The matrix normal distribution, fit to a single sequence, outperforms state-of-the-art methods at reconstructing 3D motion data in the presence of significant data loss, while providing covariance estimates of the imputed points.

Keywords: Matrix normal, trace-norm, spatiotemporal, missing data

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., ∼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 [15, 1]. 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. Additionally, we show that analytical models of the trajectory covariance capture most of the covariance of natural motions. Because such an analytical distribution over shape or structure is generally not available, we instead place a prior over the shape covariance, and derive a convex MAP solution to this problem in terms of the trace-norm. The model presented here applies to any dynamic 3D reconstruction problem, including nonrigid structure from motion, stereo, and multi-view dynamic 3D reconstruction.

Summary. In Sect. 4, we identify the Kronecker pattern in time-varying 3D point cloud covariance matrices, and present a generative probabilistic model based on the MND that explains this pattern. In Sect. 5, we establish a connection between MND and the trace-norm that leads to a convex MAP objective for 3D reconstruction. In Sect. 7, we show how this model unifies a number of shape and trajectory models, both probabilistic and algebraic, used in prior art.

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. [31]; subsequent work [19] presented a physically-based approach using nonlinear filtering over a superquadratic representation. Concurrently, Pentland and Horowitz [23] 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. [30] revisited the idea of using rigidity but at a local scale using a minimal configuration orthographic reconstruction. Salzmann and Urtasun [26] 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 [32], 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 [8], 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 [7, 38, 35, 13, 25]). Recently, Dai et al. [10] presented a method that uses a trace-norm minimization to enforce a low rank shape space, and Garg et al. [14] showed that the method can be applied to recover dense, non-rigid structure. Lee et al. [17] formulated a normal distribution over shapes in a Procrustes aligned space.

In conjunction, trajectory space representations were proposed by Sidenbladh et al. [27], which they referred to as *eigenmotions*. Akhter et al. [2] 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 [21]. Park et al. [21] 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 [34] presented various priors on trajectories in terms of 3D point differentials, showing better noise performance.

A number of approaches have combined spatial and temporal constraints [23, 19, 20, 33, 15]. Torresani et al. [33] presented a probabilistic representation, using probabilistic PCA within a linear dynamical system. The shape space and trajectory space approach were combined by Gotardo and Martinez [15], and Lee et al. [18] embedded the Procrustean distribution within a Markov process.

In contrast to prior work, our model describes an explicit parametric distribution over spatiotemporal data that allows us to define a spatiotemporal covariance matrix relating any point in time to any other point in time. The distribution can be estimated from a single sequence and used to calculate covariance estimates for missing data. As summarized in Table 1, we take a step towards reconciling a number of recent statistically-based linear representations in nonrigid structure from motion [8, 33, 20, 2, 15, 34, 10, 4].

3 Observation Model for Time-Varying 3D Point Clouds

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 f corresponds to the

3D shape in frame f, and is formed by the horizontal concatenation of points $X_p^f \in \mathbb{R}^{1\times 3}$, denoting the p-th 3D point. We will denote by $\text{vec}(\mathbf{X})$ the column-major vectorization of the matrix $\mathbf{X}$, and we will interchangeably use lowercase bold letters to denote the vectorized matrices, i.e., $\mathbf{x} = \text{vec}(\mathbf{X})$.

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

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

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 $\text{vec}_r(\mathbf{X})$. For this arrangement, the effect of orthographic projection from a single camera can be expressed as a matrix $\mathbf{O}_{\text{ortho}}$ such that

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

i.e., each of the P points is transformed by a camera matrix for frame f, equal to $\mathbf{R}_f \in \mathbb{R}^{2\times 3}$. A rearrangement of this matrix can be used with the column-major vectorization $\text{vec}(\mathbf{X})$. The case of a single camera observing the scene with unknown rotations $\mathbf{R}_f$ is the problem of 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 [11] or IMUs); simultaneous recovery of the camera matrices (as in NRSfM) is not the focus of this paper.

Our objective is to estimate the most likely spatiotemporal structure $\hat{\mathbf{X}}$ given the observations $\mathbf{y}$. Note, however, that $n_{\text{obs}} \ll 3FP$, and the problem $\min_{\mathbf{X}} \sigma^{-2}\|\mathbf{y} - \mathbf{O}\,\text{vec}(\mathbf{X})\|_2^2$ is therefore severely under constrained. We therefore take a Bayesian approach to the estimation problem,

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\text{argmax}}\, p(\mathbf{X}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X})p(\mathbf{X}), \tag{3}$$

where $p(\mathbf{y}|\mathbf{X}) = \mathcal{N}\left(\mathbf{O}\,\text{vec}(\mathbf{X}), \sigma^2\mathbf{I}\right)$ from Eq. (1). The goal is then to design a prior $p(\mathbf{X})$ that models the data well while remaining amenable to global optimization.

4 Spatiotemporal Prior for Time-Varying 3D Point Clouds

We model the time-varying structure $\mathbf{X} \in \mathbb{R}^{F\times 3P}$ as the sum of a mean component $\mathbf{M}$ and a residual non-rigid component $\mathbf{Z}$,

$$\mathbf{X} = \mathbf{M} + \mathbf{Z}. \tag{4}$$

While this does not reduce the number of variables to estimate, this decomposition will allow us to set sensible priors over the individual components.

Mean Component M. The purpose of the mean component is to capture the rigid shape of the object and its translational motion. We model this component as $\mathbf{M} = \mathbf{1}_F\mathbf{m}_{\text{shape}} + \mathbf{M}_{\text{trans}}\mathbf{P}_{\text{trans}}$, where the 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[1] $\mathbf{P}_{\text{trans}} = \text{blkdiag}(\mathbf{1}_P^T; \mathbf{1}_P^T; \mathbf{1}_P^T) \in \mathbb{R}^{3\times 3P}$.

We set a uniform prior over the mean shape: a priori, we do not have a preferred shape for objects. Because translational motion of objects that have mass is necessarily smooth, we will choose a prior for the translational component that encourages smooth motion of the object. We specify this prior using a complete trajectory basis $\mathbf{\Theta} \in \mathbb{R}^{F\times F}$, where $\mathbf{\Theta} = \tilde{\mathbf{\Theta}}\mathbf{W}_t$ with $\tilde{\mathbf{\Theta}}$ an orthonormal basis and $\mathbf{W}_t$ a diagonal weighting matrix. The basis vectors and corresponding weights in $\mathbf{W}_t$ are chosen such that smooth trajectories are more likely, resulting in a covariance over trajectories $\mathbf{\Sigma} = \mathbf{\Theta}\mathbf{\Theta}^T$ that characterizes the prior distribution over trajectories:

$$\mathbf{M}_{\text{trans}} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_3), \tag{5}$$

where $\mathcal{MN}$ denotes the Matrix Normal Distribution (MND) [12], with mean $\mathbf{0}$, column covariance $\mathbf{\Sigma}$ (describing correlations across time), and row covariance $\mathbf{I}_3$ (describing that there are no a priori correlations between the x, y, and z components).

Residual Component Z. We model the residual non-rigid deformations of the object as

$$\mathbf{Z} = \mathbf{\Theta}\mathbf{C}\mathbf{B}^T, \tag{6}$$

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 such that $\mathbf{B} = \tilde{\mathbf{B}}\mathbf{W}_b$ where $\tilde{\mathbf{B}}$ is orthonormal and $\mathbf{W}_b$ a diagonal weighting matrix. Additionally, we model the distribution over coefficients $\mathbf{C}$ as $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This corresponds to a probabilistic formulation of the bilinear model of Akhter et al. [1], and results in a matrix normal distribution

$$\mathbf{Z} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Delta}), \tag{7}$$

where $\mathbf{\Delta} = \mathbf{B}\mathbf{B}^T$ is the row covariance (describing shape correlations) and $\mathbf{\Sigma} = \mathbf{\Theta}\mathbf{\Theta}^T$ is the column covariance (describing trajectory correlations). Equivalently, the distribution over dynamic 3D structure is multivariate normal with

[1] $\mathbf{1}_P$ denotes a column vector of ones of size P, and blkdiag produces a block diagonal matrix.

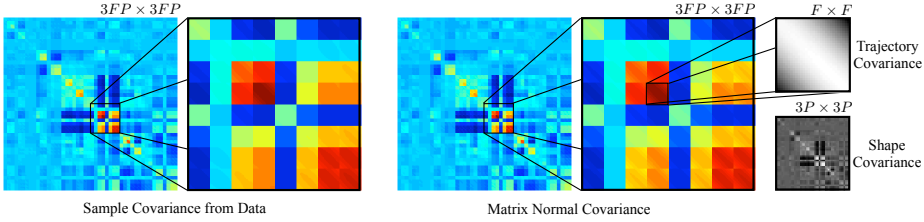


Fig. 1. Human spatiotemporal point cloud data exhibits a Kronecker structured covariance matrix, allowing us to model the distribution over sequences as matrix normal. (Left) The spatiotemporal covariance computed from 5402 vectorized sequences shows a distinct block structure, highlighted in the inset. (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.

a Kronecker structured covariance matrix [3], with $\mathbf{z} = (\mathbf{B} \otimes \mathbf{\Theta})\mathbf{c}$ and,

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Delta} \otimes \mathbf{\Sigma}). \tag{8}$$

Fig. 1 illustrates the intuition for choosing this prior over dynamic 3D structure: the spatiotemporal covariance matrix of natural motions is dominated by a Kronecker product block pattern. This Kronecker pattern of the covariance is precisely the one induced by the MND distribution.

This is a significant finding for the purposes of estimation because the MND model 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 an MND distribution is approximately $(3P)^2/2 + (F)^2/2$, versus $\sim(3FP)^2/2$ for a full covariance matrix. Even for small values of F=30 frames and P=31 points, this results in ∼5000 variables for the MND versus ∼3.9 million for a full spatiotemporal covariance.

5 Convex MAP Reconstruction

Reconstructing the 3D shape of the object can now be formulated as finding the most likely spatiotemporal configuration of points $\mathbf{X}$ given the image measurements $\mathbf{y}$ under our new probabilistic parameterization for dynamic structures,

$$p(\mathbf{y}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{y}|\mathbf{M}, \mathbf{Z})p(\mathbf{M}, \mathbf{Z}). \tag{9}$$

We assume independence between the mean and non-rigid components, $p(\mathbf{M}, \mathbf{Z}) = p(\mathbf{M})p(\mathbf{Z})$, with each of the priors described by an MND as defined above. The negative log-likelihood of the MND is quadratic, and inference under an MND

prior can be posed as a least-squares problem:

$$\underset{\mathbf{M},\mathbf{Z}}{\operatorname{argmax}}\ p(\mathbf{y}|\mathbf{M},\mathbf{Z})p(\mathbf{M})p(\mathbf{Z}) = \underset{\mathbf{M},\mathbf{Z}}{\operatorname{argmin}}\, \sigma^{-2}||\mathbf{y} - \mathbf{O}\,\mathrm{vec}(\mathbf{M}+\mathbf{Z})||_F^2 + \lambda \underbrace{\mathrm{tr}\left[{\mathbf{M}_{\mathrm{trans}}}^T \boldsymbol{\Sigma}^{-1}\mathbf{M}_{\mathrm{trans}}\right]}_{-\log(p(\mathbf{M}))+c_2} + \underbrace{\mathrm{tr}\left[\boldsymbol{\Delta}^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}\right]}_{-\log(p(\mathbf{Z}))+c_1}, \quad (10)$$

where λ is a scaling factor related that depends on the variance of the object's translational motion.

Recall that the distribution over non-rigid structures is defined by the generative model $\mathbf{Z} = \boldsymbol{\Theta}\mathbf{C}\mathbf{B}^T$, where $\boldsymbol{\Theta}$ and $\mathbf{B}$ parameterize the shape and trajectory covariances. These covariances may depend on the object and are unknown a priori, and therefore need to be estimated:

$$p(\boldsymbol{\Theta},\mathbf{C},\mathbf{B},\mathbf{M}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\Theta},\mathbf{C},\mathbf{B},\mathbf{M})p(\boldsymbol{\Theta}|\mathbf{C},\mathbf{B})p(\mathbf{B}|\mathbf{C})p(\mathbf{C})p(\mathbf{M}). \quad (11)$$

At this point, we have added optimization variables without reducing the number of unknowns. The benefit of this seemingly more complex parameterization is that we can set priors over the individual terms. The two priors that remain to be specified are:

(1) $p(\boldsymbol{\Theta}|\mathbf{C},\mathbf{B})$. Because the MND covariance is separable into shape and trajectory covariances, we can make use of a generic analytical model for the trajectory covariance $\boldsymbol{\Sigma} = \boldsymbol{\Theta}\boldsymbol{\Theta}^T$. Consider the trajectory $\mathbf{X}_p \in \mathbb{R}^{F\times 3}$ of a point p. Minimizing the kinetic energy is equivalent to minimizing $\mathrm{tr}\left[\mathbf{X}_p^T\mathbf{D}^T\mathbf{D}\mathbf{X}_p\right]$, where $\mathbf{D}$ is a first order difference matrix, which is proportional to the negative log-likelihood of a Gaussian distribution over trajectories. Define $\mathbf{G} = \mathbf{D}^T\mathbf{D}$, the second order difference matrix. It is known that $\mathbf{G} = \tilde{\boldsymbol{\Theta}}\boldsymbol{\Lambda}\tilde{\boldsymbol{\Theta}}^T$, where $\tilde{\boldsymbol{\Theta}}$ is the orthogonal DCT transform and $\boldsymbol{\Lambda}$ is a diagonal matrix (subject to boundary conditions), and therefore $\boldsymbol{\Theta} = \tilde{\boldsymbol{\Theta}}\boldsymbol{\Lambda}^{-1/2}$ [29]. The term $\boldsymbol{\Theta}$ is therefore known and drops from the expression

$$\underset{\mathbf{M},\mathbf{C},\mathbf{B}}{\operatorname{argmax}}\, p(\mathbf{y}|\boldsymbol{\Theta},\mathbf{C},\mathbf{B},\mathbf{M})p(\mathbf{B}|\mathbf{C})p(\mathbf{C})p(\mathbf{M}). \quad (12)$$

(2) $p(\mathbf{B}|\mathbf{C})$. 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 in a given sequence. We choose a normal prior over the entries of $\mathbf{B}$ (equivalently, a Wishart prior over $\boldsymbol{\Delta}$). This is computationally convenient, but more importantly, the effect is similar to the traditional low-rank shape assumption. Intuitively, the prior minimizes non-rigid deformations by encouraging that the singular values of the shape covariance matrix should decrease rapidly (see Sect. 6.1).

Using the specified priors and writing this optimization in terms of the component negative log-likelihoods,

$$\underset{\mathbf{M},\mathbf{C},\mathbf{B}}{\operatorname{argmin}}\, \sigma^2||\mathbf{y} - \mathbf{O}\,\mathrm{vec}(\mathbf{M}+\boldsymbol{\Theta}\mathbf{C}\mathbf{B}^T)||_F^2 + \|\mathbf{C}\|_F^2 + \|\mathbf{B}\|_F^2 + \lambda\|\boldsymbol{\Theta}^+\mathbf{M}_{\mathrm{trans}}\|_F^2. \quad (13)$$

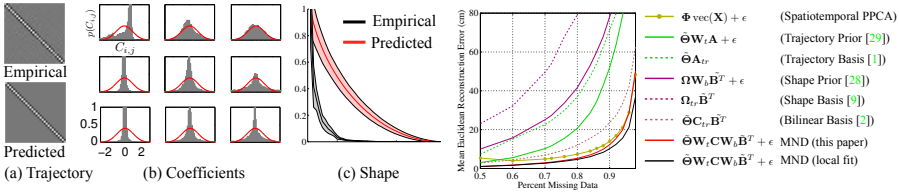


Fig. 2. (Left) Empirical and predicted parameter distributions. (a) Top: the empirical trajectory precision matrix. Below, the second order differences matrix predicted by energy minimization. (b) Each plot corresponds to a coefficient $C_{i,j}$ in the matrix $\mathbf{C}$. The red curve shows the predicted 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). (Right) Inference of missing data with known distribution parameters. Subscript tr indicates truncation.

This expression is bilinear in $\mathbf{C}$ and $\mathbf{B}$. A change of variables suffices to transform this bilinear equation into a convex problem using the matrix trace-norm $\|\cdot\|_*$. Using a result from [28], the trace-norm can also be written as $\|\mathbf{R}\|_* = \min_{\mathbf{U},\mathbf{V}}\{\frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{1}{2}\|\mathbf{V}\|_F^2\}$ subject to $\mathbf{R} = \mathbf{U}\mathbf{V}^T$. With a change of variables $\mathbf{\Theta}^+\mathbf{X}\mathbf{P}_\perp^T = \mathbf{C}\mathbf{B}^T$ we have

$$\underset{\mathbf{X}}{\operatorname{argmax}}\, p(\mathbf{X}|\mathbf{y}) = \underset{\mathbf{X}}{\operatorname{argmin}}\, \sigma^{-2}\|\mathbf{y}-\mathbf{O}\operatorname{vec}(\mathbf{X})\|_2^2 + \|\mathbf{\Theta}^+\mathbf{X}\mathbf{P}_\perp\|_* + \frac{\lambda}{\sqrt{P}}\|\mathbf{\Theta}^+\mathbf{X}\mathbf{P}_{\text{trans}}^T\|_F^2, \tag{14}$$

where $\mathbf{P}_\perp$ is a projection matrix that removes the per-frame translation component (i.e., $\mathbf{P}_\perp = \mathbf{I} - \mathbf{P}_{\text{trans}}^T(\mathbf{P}_{\text{trans}}\mathbf{P}_{\text{trans}}^T)^{-1}\mathbf{P}_{\text{trans}}$). Note that this is the inverse operation of $\mathbf{P}_{\text{trans}}^T$, which isolates the per-frame translation such that $\mathbf{X}\mathbf{P}_{\text{trans}}^T = P\mathbf{M}_{\text{trans}}$. This objective lends itself to optimization by the Alternating Direction Method of Multipliers (ADMM) [6], being decomposable into readily solvable sub-problems (see supplementary materials for details), or as a generalized trace-norm problem [5].

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 their mean shape.

I. Kronecker Covariance Structure. (Sect. 4) Fig. 1(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[2] 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 $v = \frac{1}{3P}\mathrm{tr}(\mathbf{\Delta})$. Note that this separable approximation captures most of the structure and energy in the covariance using far fewer parameter than a full covariance matrix.

II. Analytical Trajectory Distribution. (Sect. 5) Fig. 2(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 energy minimization. Most correlations in the data are captured by the analytical model.

III. Distribution of Coefficients. (Sect. 4) 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. 2(b) shows that the empirical distribution can be more spiked, closer to Laplacian or Cauchy.

IV. Hierarchical Prior on Shape Covariance. (Sect. 5) We sample shape covariance matrices from the prior $\mathbf{B} \sim \mathcal{MN}(0, \mathbf{I}_{3P}, \mathbf{I}_{3P})$ and compute their singular values (SVs). Fig. 2(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 the energy of the singular values by the induced prior on $\mathbf{B}$ is not as quick as that observed from data, but this particular choice allows for a convex optimization. Finding priors with faster fall-off but that still remain amenable to global minimization is an interesting direction for future research.

6.2 Missing Data in Motion Capture

The objective of these experiments is to characterize the resilience of the model to typical patterns of missing data encountered in dynamic reconstruction. We decouple the problem of missing data from projection loss and reconstructibility [34] by studying inference on 3D observations (e.g., the output from a motion capture system). The task is to infer the complete sequences from a reduced set of 3D observations. We use the observation model $\mathbf{O}_{\text{miss}}$ as per Sect. 3.

Known Distribution Parameters. When 3D training data is available, we can learn the parameters for MND distribution and perform inference with Eq. (10). 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 Principal Component Analysis (PCA) model trained on the vectorized spatiotemporal sequences, i.e., $\mathbf{y} = \Phi\,\mathrm{vec}(\mathbf{X}) + \epsilon$. We report mean 3D error in Figure 2. As a reference, the error incurred when using the mean shape at every frame as an estimation is $\sim$175cm.

For this experiment, we use data from the CMU Motion Capture database. We take 50 random sequences of $20s$ in duration, sample them at 30Hz and Procrustes align and mean center them. There are 31 markers on the body, and

[2] ML estimates of the parameters for noiseless data can be obtained using a "flip-flop" algorithm [12], but in practice we obtained better results with this procedure.

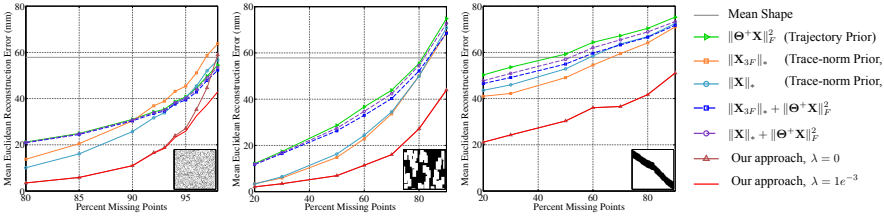


Fig. 3. Inferring missing data under three different occlusion patterns when the shape distribution is unknown. The graphs show mean Euclidean error in the reconstruction under the occlusion models discussed in Section 6.2. The bottom two results correspond to the method of Sect. 5. We investigate two 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.

we subdivide each sequence into $1s$ windows 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 the subsequences neighboring the test subsequence; the model is more specific and results in lower error.

Unknown Distribution Parameters. When no training data is available, we rely on the convex inference procedure described in Sect. 5. We compare our approach with three different priors: (1) a trajectory-only prior, (2) a trace-norm prior, and (3) a naive combination of the trace-norm and trajectory priors. We assume σ=1mm for all methods. We use dense motion capture data from Park and Hodgins [22]. The sequences are captured at 120Hz with a dense spatial sampling across the body. We downsample by four spatially and temporally, yielding a point cloud of 118 points at 30Hz 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. Nonetheless, it is of interest here because under this pattern, the trace-norm is known to provide minimum rank solutions with high probability [24]. **(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 track length is modified to achieve a particular level of missing data (with respect to the original sequence).

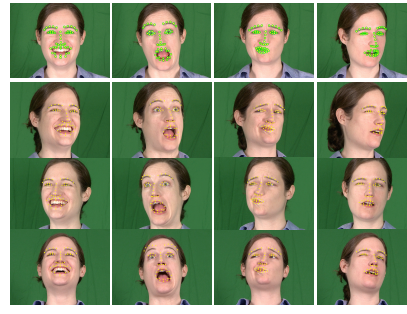
The resulting occlusion patterns are shown as insets on the graphs in Fig. 3, laid out as a matrix of frames by points, where black denotes observed entries. We note that *correspondence loss* results in a much harder problem. Independently of the occlusion pattern, the proposed approach improves results.

6.3 Non-rigid Structure from Motion

We compare the performance of our time-varying point cloud reconstruction method using Eq. (10) on a standard set of structure from motion sequences, where the only data loss is from projection. We report normalized mean 3D error as computed in [15] for four methods, (1) KSTA [15], a non-linear kernelized shape-trajectory method, (2) Dai et al. [10], (3) a trajectory-only prior, and (3) our approach. For our method, we compute the camera matrices as in Dai et al. [10] [3], and set σ=1 and λ=0. For Dai et al. and KSTA, the optimal parameter k was chosen for each test.

Dataset	KSTA	Dai	Traj.	Ours
Drink	0.0156	0.0266	0.0102	**0.0099**
Pick-up	0.2322	0.1731	**0.1707**	**0.1707**
Yoga	0.1476	0.1150	0.1125	**0.1114**
Stretch	**0.0674**	0.1034	0.0972	0.0940
Dance	0.2504	0.1842	0.1385	**0.1347**
Face2	0.0339	0.0303	0.0408	**0.0299**
Walking2	**0.1029**	0.1298	0.3111	0.1615
Shark2	**0.0160**	0.2358	0.1380	0.1297
Capoeira	**0.2376**	0.3931	0.4394	0.3786

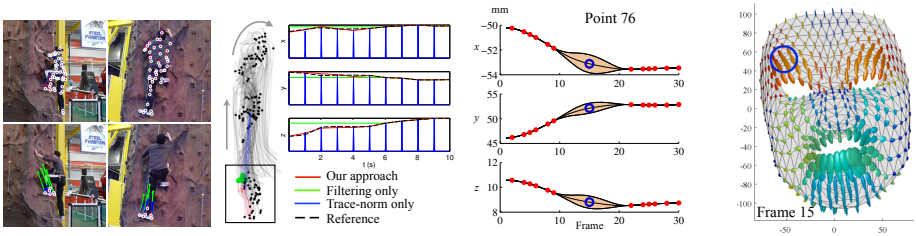
(a) Performance on NR-SfM data sets



(b) Frontal face 3D reconstruction

Fig. 4. (a) Comparison on standard NRSfM sequences using normalized mean 3D error as reported by [10] and [15]. For our method, we compute the camera matrices as Dai et al. Our method shows improved performance on 5 of 8 sequences, while the non-linear KSTA method can achieve better performance on some sequences. (b) Reconstructing a dynamic face from a frontal view. The top row shows frames from a video with superimposed detected 2D landmarks (green circles) provided by IntraFace [37]. We reconstruct the face in full 3D using Eq. (14) and show the reprojection onto three other (held out) views for comparison (yellow dots).

[3] For KSTA [15], the camera matrices are computed as per Akhter et al. [2]. The implementation of Dai et al. and KSTA was provided by the respective authors.



(a) Multiview Reconstruction (b) Confidence Bounds

Fig. 5. (a) Multiview reconstruction on the "Rock Climbing" sequence from [21]. 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. (b) The matrix normal prior 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.

6.4 Monocular reconstruction

In Fig. 4(b) we show a 3D point cloud reconstruction example from a frontal view of a face using 2D landmark detections provided by IntraFace [37]. The original video is around 1500 frames long and is reconstructed simultaneously. Only a subset of frames is shown here. We directly use the model of Eq. (14) 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.

6.5 Multiview Dynamic Reconstruction

We perform a qualitative evaluation of the method of Sect. 5 on a dynamic reconstruction sequence from Park et al. [21]. 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 30Hz. Because the observations 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 [21]. Fig. 5(a) shows reconstructions on two sequences, where we have simulated two types of occlusion. Because ground truth is not available, we first run all methods on the full data to obtain a reference reconstruction and we average the resulting structure. This result is shown in black. Fig. 5(a)(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. However, when we look at the temporal domain Fig. 5(a)(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 30Hz) negatively affect the performance of trace-norm based methods. Our method is able to achieve a smoother interpolation while maintaining a low-rank structure.

6.6 3D Time-varying Point Cloud Reconstruction

In Fig. 6 we show a reconstruction of the baseball sequence acquired by Joo et al. [16]. The sequence is given as a set of 3D point trajectories obtained from a multi-camera system. Each trajectory is only 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 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.

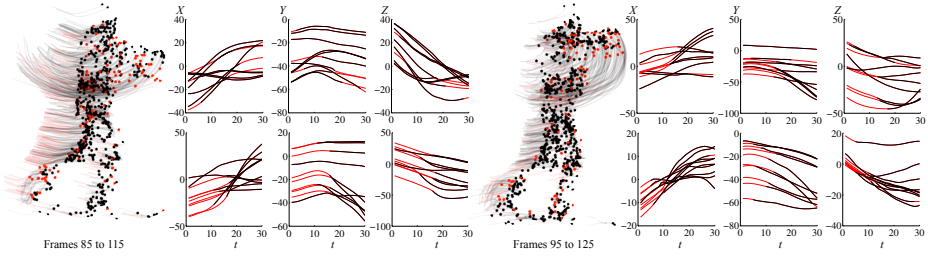


Fig. 6. Reconstructing a baseball motion sequence. Black lines indicate observed points, red lines are inferred trajectories. Two motion trail diagrams of 30-frame overlapping parts of a baseball swing are shown. The graphs show a close up reconstruction for different subsets of the points.

7 Discussion

The model over dynamic 3D structure we describe can be related to shape, trajectory, and shape-trajectory representations used in prior art [15, 8, 9, 27, 36,

	Truncation	Probabilistic	Convex Approx.
Shape	Bregler et al. [8] $\mathbf{X} = \mathbf{\Omega}\tilde{\mathbf{B}}^T$	Torresani et al. [33] $\mathbf{X} = \mathbf{\Omega}\mathbf{W}_b\tilde{\mathbf{B}}^T + \epsilon$	Dai et al.[10] $\|\mathbf{X}\|_*$
Trajectory	Akhter et al. [2] $\mathbf{X} = \tilde{\mathbf{\Theta}}\mathbf{A}$	Valmadre et al. [34] $\mathbf{X} = \tilde{\mathbf{\Theta}}\mathbf{W}_b\mathbf{A} + \epsilon$	
Shape-Trajectory	Gotardo and Martinez [15] $\mathbf{X} = \tilde{\mathbf{\Theta}}\mathbf{C}\tilde{\mathbf{B}}^T$	(This Paper) $\mathbf{X} = \tilde{\mathbf{\Theta}}\mathbf{W}_t\mathbf{C}\mathbf{W}_b\tilde{\mathbf{B}}^T + \epsilon$ $\|\mathbf{\Theta}^+\mathbf{X}\mathbf{P}_\perp\|_*$	

Table 1: Comparison of linear methods for structure reconstruction. See Sect. 4.

33, 2, 34] (see Table 1). The convex MAP minimization of Eq. (14). 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 [5, 10]. 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 it is illustrative to consider the marginal distribution it induces for a particular trajectory $\mathbf{x}^j$ (a column j of $\mathbf{X}$) independent of all other points. This corresponds to an equivalent basis representation over trajectories, as described by Sidenbladh et al. [27]. The marginal distribution is then $\mathbf{x}_j \sim \mathcal{N}(\mathbf{M}^j, \mathbf{\Delta}_{j,j}\mathbf{\Sigma})$, where $\mathbf{\Sigma}{=}\mathbf{\Theta}\mathbf{\Theta}^T$ is the trajectory covariance matrix, and $\mathbf{\Delta}_{j,j}$ loosely corresponds to the mass of each point. This expression is equivalent to the *filtering* solution proposed by Valmadre and Lucey [34], who observe that a combination of first and second-order differences fit natural motions well. See also [26].

Shape Methods. The marginal distribution of a particular shape $\mathbf{x}^i$ (a row i 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. [9], and the shape basis model of Torresani et al. [33]. It follows from the matrix normal model that $\mathbf{x}^i \sim \mathcal{N}(\mathbf{M}^i, \mathbf{\Sigma}_{i,i}\mathbf{\Delta})$,, where $\mathbf{\Sigma}_{i,i}$ is the entry (i, i) in $\mathbf{\Sigma}$ and $\mathbf{\Delta}{=}\mathbf{B}\mathbf{B}^T$ is the shape covariance matrix. An equivalent shape basis $\mathbf{B}$ is usually computed with PCA [8, 36, 2, 33].

Bilinear Spatiotemporal Methods. The model we present is a probabilistic formulation of the shape-trajectory basis models described in [15, 1]. 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. 5(b) on a facial motion capture sequence from [1].

Trace-norm Methods. The trace-norm term in Eq. (14) can be written in terms of the "generalized trace-norm" developed by Angst et al. for rigid SfM [5]. 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 a time-varying 3D structure. Compared to the trace-norm regularizer of Dai et al. [10],

we obtain an equivalent minimization if we assume that the temporal covariance is an identity matrix (and set λ=0). 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. 5(a) (right).

References

1. Akhter, I., Simon, T., Matthews, I., Khan, S., Sheikh, Y.: Bilinear spatiotemporal basis models. TOG (2012)
2. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid structure from motion in trajectory space. NIPS (2008)
3. Allen, G., Tibshirani, R.: Transposable regularized covariance models with an application to missing data imputation. Annals of Applied Statistics (2010)
4. Angst, R., Pollefeys, M.: A unified view on deformable shape factorizations. ECCV (2012)
5. Angst, R., Zach, C., Pollefeys, M.: The generalized trace-norm and its application to structure-from-motion problems. ICCV (2011)
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning (2011)
7. Brand, M.: A direct method for 3d factorization of nonrigid motion observed in 2d. ICCV (2005)
8. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. CVPR (2000)
9. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models, their training and application. CVIU (1995)
10. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. CVPR (2002)
11. Del Bue, A., Llad, X., Agapito, L.: Non-rigid face modelling using shape priors. In: Analysis & Modelling of Faces & Gestures (2005)
12. Dutilleul, P.: The mle algorithm for the matrix normal distribution. Statistical Computation and Simulation (1999)
13. Fayad, J., Del Bue, A., Agapito, L., Aguiar, P.: Non-rigid structure from motion using quadratic deformation models. BMVC (2009)
14. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. CVPR (2013)
15. Gotardo, P., Martinez, A.: Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. PAMI (2011)
16. Joo, H., Park, H., Sheikh, Y.: Optimal visibility estimation for large-scale dynamic 3d reconstruction. CVPR (2014)
17. Lee, M., Cho, J., Choi, C., Oh, S.: Procrustean normal distribution for non-rigid structure from motion. CVPR (2013)
18. Lee, M., Choi, C., Oh, S.: A procrustean markov process for non-rigid structure recovery. CVPR (2014)
19. Metaxas, D., Terzopoulos, D.: Shape and nonrigid motion estimation through physics-based synthesis. PAMI (1993)
20. Olsen, S., A. Bartoli, A.: Implicit non-rigid structure-from-motion with priors. Journal of Mathematical Imaging and Vision (2008)

21. Park, H.S., Shiratori, T., Matthews, I., Sheikh, Y.: 3D reconstruction of a moving point from a series of 2D projections. ECCV (2010)
22. Park, S.I., Hodgins, J.K.: Data-driven modeling of skin and muscle deformation. TOG (2008)
23. Pentland, A., Horowitz, B.: Recovery of nonrigid motion & structure. PAMI (1993)
24. Recht, B., Fazel, M., Parrillo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM (2010)
25. Russell, C., Fayad, J., Agapito, L.: Energy based multiple model fitting for non-rigid structure from motion. CVPR (2011)
26. Salzmann, M., Urtasun, R.: Physically-based motion models for 3d tracking: A convex. formulation. ICCV (2011)
27. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. ECCV (2000)
28. Srebro, N., Rennie, J., Jaakkola, T.: Maximum margin matrix factorizations. NIPS (2005)
29. Strang, G.: The discrete cosine transform. SIAM review (1999)
30. Taylor, J., Jepson, A., Kutulakos, K.: Non-rigid structure from locally-rigid motion. CVPR (2010)
31. Terzopoulos, D., Witkin, A., Kass, M.: Constraints on deformable models: Recovering 3d shape and nonrigid motion. Artificial Intelligence (1988)
32. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. IJCV (1992)
33. Torresani, L., Hertzmann, A., Bregler, C.: Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. PAMI (2008)
34. Valmadre, J., Lucey, S.: A general trajectory prior for non-rigid reconstruction. CVPR (2012)
35. Vidal, R., Abretske, D.: Nonrigid shape and motion from multiple perspective views. ECCV (2006)
36. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. ECCV (2004)
37. Xiong, X., De La Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR (2013)
38. Yan, J., Pollefeys, M.: A factorization-based approach to articulated motion recovery. CVPR (2005)