MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

A Direct Method for 3D Factorization of Nonrigid Motion Observed in 2D

Brand, M.

TR2005-052 December 2005

Abstract

The nonrigid structure-from-motion (NSFM) problem seeks to recover a sequence of 3D shapes, shape articulation parameters, and camera view matrices from 2D correspondence data. Factorization approaches relate the principal subspaces of the data matrix to the desired parameters through a linear corrective transform. Current methods for finding this transform are heuristic or depend on strong assumptions about the data. We show how to solve for this transform by directly minimizing deviation from the required orthogonal structure of the projection/articulation matrix. The solution is exact for noiseless data and an order of magnitude more accurate than state-of-the-art methods for noisy data.

IEEE International Conference on Computer Vision and Pattern Recognition

A direct method for 3D factorization of nonrigid motion observed in 2D

Matthew Brand
Mitsubishi Electric Research Labs,
Cambridge MA, USA

Abstract

The nonrigid structure-from-motion (NSFM) problem seeks to recover a sequence of 3D shapes, shape articulation parameters, and camera view matrices from 2D correspondence data. Factorization approaches relate the principal subspaces of the data matrix to the desired parameters through a linear corrective transform. Current methods for finding this transform are heuristic or depend on strong assumptions about the data. We show how to solve for this transform by directly minimizing deviation from the required orthogonal structure of the projection/articulation matrix. The solution is exact for noiseless data and an order of magnitude more accurate than state-of-the-art methods for noisy data.

1 Nonrigid structure from motion

Nonrigid SFM is the well-known problem of reconstructing 3D shape and deformations of a nonrigid surface from a set of correspondences across 2D views. Typically the solution is expressed as a linear basis for 3D shape, articulation weights for the observed shape in each frame, and projective parameters for each 2D camera. The basis, often called a morphable model, is particularly useful for the analysis and synthesis of organic shapes such as faces, internal organs, and patches of skin. Demand for a factorization solution is driven by rapid progress toward robust monocular 2D trackers. There is also applicability to the analysis of stereo views and 3D motion capture data.

Parke [6] and Terzopoulos [8] and Forchheimer [5] pioneered the use of morphable models graphics, vision, and video coding, respectively. After the landmark rigid-structure-from-motion factorization of Tomasi and Kanade [9], it was conjectured that morphable models could be automatically acquired by factorization. Bregler *et alia* [2] led community interest in the problem with a proposed nonrigid generalization of the Tomasi-Kanade method. Brand [1] presented a counter-example and pointed out that the problem of upgrading an algebraic factorization to geometric orthogonality was underconstrained. Papers by Brand [1], Torresani, Hertzmann, & Bregler [12, 11, 10], and Del Bue & Agapito [3] suggested various additional constraints and heuristics for regularizing the problem, and began to show substantial empirical results. A second landmark paper from Xiao, Chai, and Kanade [13] (hereafter XCK) analyzed the problem's main degeneracy and introduced a set of additional constraints that enables a closed-form solution that is exact for noiseless data.

This paper shows how the problem can be solved without any such additional constraints. Like the XCK method, our solution is exact for noiseless data. However, the error surface contemplated by all methods is quartic in the unknowns, and nested least-squares solutions such as XCK's essentially ignore some of the terms. With noisy data or clean data with a long-tailed singular value spectrum, these terms can make a substantial contribution to the error, leading to suboptimal factorizations. We recast nonrigid SFM as a constrained optimization problem and show how to efficiently and directly minimize the error, thereby obtaining substantially better factorizations of both synthetic and real-world data.

2 Formal problem statement

Let matrix $\mathbf{P} \in \mathbb{R}^{2F\times N}$ record the 2D image locations of N 3D points observed in F frames from a single deforming surface having K linear degrees of freedom. W.l.o.g., assume that $\mathbf{P}$ is row-centered, such that $\mathbf{P1} = \mathbf{0}$. Submatrix $\mathbf{P}_f \in \mathbb{R}^{2\times N}$ records the observed 2D locations of N features in a single frame. Under weak perspective projection, the data-generation model is $\mathbf{P}_f = \mathbf{R}_f \sum_{k=1}^{K} c_{kf}\mathbf{S}_k$, where truncated row-orthonormal rotation matrix $\mathbf{R}_f \in \mathbb{R}^{2\times 3}$ is applied to weighted sum of basis shapes, each shape $\mathbf{S}_k \in \mathbb{R}^{3\times N}$ scaled by a scalar weight c_{kf}. We seek to factor the observation data $\mathbf{P}$ into motion and shape parameters

$$\mathbf{P} = \mathbf{MS} = \begin{bmatrix} \mathbf{c}_1^T \otimes \mathbf{R}_1 \\ \vdots \\ \mathbf{c}_F^T \otimes \mathbf{R}_F \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_K \end{bmatrix} \tag{1}$$

with shape basis $\mathbf{S} \in \mathbb{R}^{3K\times N}$, and motion matrix $\mathbf{M} \in \mathbb{R}^{2F\times 3K}$ consisting of frame-specific weight vectors $\mathbf{c}_f^T = [c_{1f}, \cdots, c_{Kf}] \in \mathbb{R}^K$ and projective rotations $\mathbf{R}_f$ satisfying $\mathbf{R}_f\mathbf{R}_f^T = \mathbf{I} \in \mathbb{R}^2$. The Kronecker (outer) product $\otimes$ and related operators are reviewed in appendix B.

Bregler *et alia* [2] proposed to solve the problem by generalizing the Tomasi-Kanade rigid structure-from-motion factorization: Factor $\mathbf{P}$ via singular value decomposition

(SVD): $\mathbf{P} \rightarrow \tilde{\mathbf{M}}\tilde{\mathbf{S}}$, where proto-motion matrix $\tilde{\mathbf{M}}$ and proto-shape matrix $\tilde{\mathbf{S}}$ are related to $\mathbf{M},\mathbf{S}$ through an unknown *corrective transform* $\mathbf{G} \in \mathbb{R}^{3K\times 3K}$, with $\mathbf{M} = \tilde{\mathbf{M}}\mathbf{G}$ restoring the orthogonal structure of the rotations (such that $\mathbf{R}_f\mathbf{R}_f^T = \mathbf{I}_2$) and $\mathbf{S} = \mathbf{G}^{-1}\tilde{\mathbf{S}}$ restoring the shape basis. Solving for $\mathbf{G}$ when $K > 1$ has turned out to be a thorny problem.

Let us review how $\mathbf{G}$ is obtained in the rigid $K = 1$ case. Denote the pair of rows in $\mathbf{M}$ giving the projection for frame f by $\mathbf{x}_f^T$ and $\mathbf{y}_f^T$. These are rows of a scaled rotation matrix and thus must be orthogonal ($\mathbf{x}_f^T\mathbf{y}_f = 0$) and of equal norm ($\mathbf{x}_f^T\mathbf{x}_f = \mathbf{y}_f^T\mathbf{y}_f$; equivalently $(\mathbf{x}_f - \mathbf{y}_f)^T(\mathbf{x}_f + \mathbf{y}_f) = \mathbf{x}_f^T\mathbf{x}_f - \mathbf{y}_f^T\mathbf{y}_f = 0$). These equalities can be expressed as linear constraints on the elements of the gram matrix $\mathbf{G}\mathbf{G}^T$ using the following notation: Let $\text{vech}(\mathbf{X})$ to be a vector that serializes the elements on the lower triangle of symmetric matrix $\mathbf{X}$ and define $\text{vc}(\mathbf{x},\mathbf{y}) \doteq \text{vech}(\mathbf{x}\mathbf{y}^T + \mathbf{y}\mathbf{x}^T - \text{diag}(\mathbf{x}\circ\mathbf{y}))$, with $\circ$ denoting Hadamard (elementwise) product. Then we can write the constraints on $\mathbf{G}$ as

$$\forall_f \, [\text{vc}(\tilde{\mathbf{x}}_f,\tilde{\mathbf{y}}_f), \text{vc}(\tilde{\mathbf{x}}_f - \tilde{\mathbf{y}}_f, \tilde{\mathbf{x}}_f + \tilde{\mathbf{y}}_f)]^T \text{vech}(\mathbf{G}\mathbf{G}^T) = \mathbf{0} \quad (2)$$

with $\tilde{\mathbf{M}}_f = [\tilde{\mathbf{x}}_f, \tilde{\mathbf{y}}_f]^T$ a row-pair in $\tilde{\mathbf{M}}$.

In the case of rigid motion, the resulting homogeneous system of linear constraints suffices to fully determine the $g(K) \doteq \binom{3K}{2}$ unknowns of the symmetric gram matrix $\mathbf{G}\mathbf{G}^T$. The (possibly approximate) nullspace of this system is a single vector $\mathbf{n} \in \mathbb{R}^{g(1)}$ containing these $g(1) = 6$ values. One way to do this calculation is to collect all $[\cdot]$-forms from equation 2 as the columns of a matrix $\mathbf{L} \in \mathbb{R}^{g(K)\times 2F}$, solve for the minimizing eigenvector $\mathbf{n}$ of gram matrix $\mathbf{Q}_A \doteq \mathbf{L}\mathbf{L}^T$, form the symmetric matrix $\mathbf{Q}_B \doteq \text{vech}^{-1}\mathbf{n} = \mathbf{G}\mathbf{G}^T \in \mathbb{R}^{3\times 3}$, and factor $\mathbf{Q}_B$ via SVD to obtain an estimate of $\mathbf{G}$ (up to an arbitrary scaling and rotation of its row-space). Because $\mathbf{n}$ spans the (possibly approximate) nullspace of $\mathbf{L}$, this minimizes the sum-squared deviation from orthogonality in the final motion matrix,

$$\mathit{OrthError}_{\mathbf{Q}_A}(\mathbf{G}) \doteq \text{vech}(\mathbf{G}\mathbf{G}^T)^T\mathbf{Q}_A\text{vech}(\mathbf{G}\mathbf{G}^T), \quad (3)$$

and thereby restores the physical constraint that the 3D geometry of the scene is independent of the camera position. Minimizing this error is the *sine qua non* of a correct structure-from-motion algorithm. If it is not minimized, then some of the motion observed in 2D can be falsely attributed to camera motion rather to 3D shape (and vice versa), resulting in degraded estimates of both $\mathbf{M}$ and $\mathbf{S}$.

2.1 Underdetermined variables for $K > 1$

In the nonrigid case, motion matrix $\mathbf{M}$ grows to $3K$ columns, organized in K triads. Attempts to generalize the Tomasi-Kanade solution for $\mathbf{G}$ to nonrigidity typically focus on just one triad of columns in $\mathbf{M}$, which would be obtained from proto-motion matrix $\tilde{\mathbf{M}}$ by the corresponding column-triad in $\mathbf{G}$. We will write this as $\mathbf{M}_{1:3} = \tilde{\mathbf{M}}\mathbf{G}_{1:3} \in \mathbb{R}^{2F\times 3}$, with the subscripts denoting a range of columns. The orthogonality constraints on $\mathbf{M}_{1:3}$ are the same as in the rigid case, but are expressed here by writing equation 2 with the $\text{vc}(\cdot)$ operator applied to entire rows of $\tilde{\mathbf{M}}$. Then $\mathbf{G}$ is replaced with $\mathbf{G}_{1:3}$ in equations 2–3 to yield a system of linear constraints on the elements of gram matrix $\mathbf{Q}_B = \mathbf{G}_{1:3}\mathbf{G}_{1:3}^T \in \mathbb{R}^{3K\times 3K}$. Unfortunately, the linear system does not sufficiently determine $\mathbf{Q}_B$: In the case of nonrigid nonconstant motion ($K > 1$ and $\text{rank}(\mathbf{C}) > 1$), XCK showed that there exist many possible values of $\mathbf{Q}_B$ that satisfy the orthogonality constraints without admitting a real-valued factorization $\mathbf{Q}_B \rightarrow \mathbf{G}_{1:3}\mathbf{G}_{1:3}^T$.

This is partly because the system of linear constraints on the elements $\mathbf{Q}_B$ can be have a multidimensional nullspace, making $\mathbf{n} = \text{vech}\,\mathbf{Q}_B = \text{null}\,\mathbf{Q}_A$ underdetermined. In general, the solution matrices $\mathbf{M}$ and $\mathbf{S}$ (and therefore $\mathbf{G}$) are determinable up to the Kronecker product of an arbitrary $K \times K$ nonsingular transform $\mathbf{F} \in GL_K(\mathbb{R})$ of the collected weights matrix $\mathbf{C} \doteq [\mathbf{c}_1, \cdots, \mathbf{c}_F] \in \mathbb{R}^{K\times F}$ and an arbitrary 3D rotation $\mathbf{E} \in SO_3(\mathbb{R})$ of the basis shapes $\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_K$. This implies that there is an $(\mathbb{R}^K \backslash \mathbf{0}) \times \mathbb{S}^2$ manifold of equally correct values for $\mathbf{G}_{1:3}$ (or $\mathbb{S}^{K-1} \times \mathbb{S}^2$ if we ignore global scalings). As in the rigid case, it is not necessary to fix the rotational ($\mathbb{S}^2$) component of this invariance, but in the nonrigid case it is necessary to specify *which* of $\mathbf{M}$'s column-triads (or combinations thereof) we wish $\mathbf{G}_{1:3}$ to extract from $\tilde{\mathbf{M}}$.

One way to do this is by choosing a set of K frames and specifying *a priori* the deformation weights that $\mathbf{G}_{1:3}$ should assign to them. *I.e.*, we fix the first element of modified weight vector $\mathbf{c}_i'$ in each of K frames and thus determine the column of $\mathbf{F}$ that generates $\mathbf{G}_{1:3}$ (assuming $\mathbf{c}_i', \mathbf{c}_j', \cdots$ are linearly independent). XCK showed how to do this in the factorization setting by proposing that we assume that the matrix of all deformation weights $\mathbf{C}$ has been transformed such that K of its columns form the identity matrix $\mathbf{I}_K$. *I.e.*, $[\mathbf{c}_f', \mathbf{c}_g', \cdots] \doteq \mathbf{F}[\mathbf{c}_f, \mathbf{c}_g, \cdots] = \mathbf{I}_K$ for some subset of weight vectors $[\mathbf{c}_f, \mathbf{c}_g, \cdots] \subset \mathbf{C}$. This zeros out a large swath of $\mathbf{M}$ and adds the constraints

$$\forall_b \, \text{vc}(\tilde{\mathbf{m}}_a, \tilde{\mathbf{m}}_b)^T \text{vech}(\mathbf{G}_{1:3}\mathbf{G}_{1:3}^T) = 0 \quad (4)$$

for each row $\tilde{\mathbf{m}}_a^T$ in $\tilde{\mathbf{M}}$ corresponding to a zeroed basis weight $c_{ka} = 0$, and any other row $\tilde{\mathbf{m}}_b^T \subset \tilde{\mathbf{M}}$. One basis weight in each column of $\mathbf{C}$ is also set to $c_{kf} = 1$; yielding the unit constraint

$$\text{vc}(\tilde{\mathbf{x}}_f, \tilde{\mathbf{y}}_f)^T \text{vech}(\mathbf{G}_{1:3}\mathbf{G}_{1:3}^T) = 1 \quad (5)$$

for one frame. XCK showed that if the data has the expected algebraic structure, the collected set of constraints from equations (2,4,5) will have a unique solution $\mathbf{n} \in \mathbb{R}^{g(K)}$ that parameterizes a symmetric rank-3 matrix $\mathbf{Q}_B \doteq \text{vech}^{-1}\mathbf{n} = \mathbf{G}_{1:3}\mathbf{G}_{1:3}^T$. Then $\mathbf{G}_{1:3}$ can be factored from

$\mathbf{Q}_B$ via truncated SVD. Repeating this procedure for each row of $\mathbf{F}[\mathbf{c}_f, \mathbf{c}_g, \cdots] = \mathbf{I}_K$ yields K successive column-triads $\mathbf{G}_{4:6}, \mathbf{G}_{7:9}, \cdots$, each determined up to a 3D rotation.

The XCK solution, which we will call the **basis method** because it employs assumptions about the basis, is correct when (A) the data generated noiselessly via the forward model from a full-rank basis ($\text{rank}(\mathbf{S}^{(3)}) = K$), (B) the exact number of deformational modes K can be deduced from the rank of the data matrix ($\text{rank}(\mathbf{P}) = 3K$), and (C) the 3D shapes in the selected K basis frames are linearly independent (after rotational alignment). It begins to break down in other circumstances, particularly when the data is noisy or the wrong value of K is used. In these cases, the estimated $\mathbf{n}$ almost certainly does not parameterize a rank-3 positive semidefinite gram matrix $\mathbf{Q}_B$, and therefore some information about the orthogonality constraints will be lost when $\mathbf{G}_{1:3}$ is obtained from $\mathbf{Q}_B$ by truncated SVD. This is problematic because the resulting estimate of $\mathbf{G}_{1:3}\mathbf{G}_{1:3}^T$ is *not* the rank-3 approximation of $\mathbf{Q}_B$ that minimizes the orthogonality error $OrthError_{\mathbf{Q}_A}(\mathbf{G}_{1:3})$; it is just the minimum squared error rank-3 approximation of $\mathbf{Q}_B$.

The culprit is a change of error norm: $\mathbf{Q}_A$ specifies the correct problem-specific Mahalanobis (elliptic) error metric in $\mathbb{R}^{g(K)}$ but the rank-3 factorization (SVD) of $\mathbf{Q}_B$ optimizes a problem-independent spherical error measure in $\mathbb{R}^{3K}$, ignoring impact on the orthogonality error. In short, barring perfect data, the XCK estimate of $\mathbf{G}_{1:3}$ is a suboptimal approximation of an approximation, each made under different and inconsistent error norms. This problem is well known in eigenvalue methods for fitting higher-order algebraic surfaces to data [4]: Nested EVD's optimize algebraic error (here, distance from $\mathbf{G}_{1:3}\mathbf{G}_{1:3}^T$ to $\mathbf{Q}_B$) rather than the geometric objective objective (here, orthogonal structure of the scene).

3 Directly minimizing error

We now introduce a **direct method** that solves for the elements of $\mathbf{G}_{1:3}$ rather than for the elements of its gram matrix $\mathbf{Q}_B = \mathbf{G}_{1:3}\mathbf{G}_{1:3}^T$. In avoiding $\mathbf{Q}_B$, we sidestep all the pathological indeterminacies discussed above in section 2.1, and thus can rely purely on the original orthogonality constraints. (The nonpathological invariances are fixed by initial conditions.)

We will solve for a single column-triad $\mathbf{G}_{1:3} \subset \mathbf{G}$; section 4 below will show that this suffices to determine the entire solution. Thus we have only $9K$ unknowns for $\mathbf{G}_{1:3}$, as opposed to $\binom{3K}{2} = O(K^2)$ unknowns for indirect methods that first solve for $\mathbf{Q}_B$. Since the factorization is ultimately invariant to scalings of $\mathbf{G}_{1:3}$, we begin with a constrained optimization problem: Minimize $OrthError_{\mathbf{Q}_A}(\mathbf{G}_{1:3})$ subject to $\|\mathbf{G}_{1:3}\|_F = 1$. The optimization domain is therefore the surface of the unit sphere $\mathbb{S}^{9K-1}$ embedded in $\mathbb{R}^{9K}$.

To optimize, we will take maximal-gradient slices through this sphere and solve for the subspace optimum in closed form. Because the sphere is "laced" with an embedded submanifold $\mathbb{S}^{K-1} \otimes \mathbb{S}^2 \subset \mathbb{S}^{9K-1}$ of zero-error solutions (noiseless case), the error surface is akin to a balloon bulging through a net; one merely needs to descend to one of the "strings" where the error is globally (but not uniquely) minimized.

For the remainder of this section we will simplify the notation $OrthError_{\mathbf{Q}_A}(\mathbf{G}_{1:3})$ to $E(\mathbf{Z})$ with $\mathbf{Z} \doteq \mathbf{G}_{1:3}$. The partial gradient of $E(\mathbf{Z})$ w.r.t. any element of $\mathbf{Z}$ is

$$\partial_{Z_{ij}} E(\mathbf{Z}) = 2\text{vech}(\mathbf{Z}\mathbf{Z}^T)^T \mathbf{Q}_A \text{vech}(\mathbf{Z}\mathbf{J}_{ij}^T + \mathbf{J}_{ij}\mathbf{Z}^T) \qquad (6)$$

where $\mathbf{J}_{ij} \in \{0,1\}^{3K\times 3}$ is all zeros except for element $J_{ij} = 1$. An optimization on a sphere is easily converted to an unconstrained optimization problem by modifying the error to be invariant to the norm of the optimization variable. The norm-invariant error is $E'(\mathbf{Z}) \doteq E(\mathbf{Z}) \cdot \|\mathbf{Z}\|_F^{-4}$ with gradient

$$\partial_{Z_{ij}} E'(\mathbf{Z}) = ((\partial_{Z_{ij}} E(\mathbf{Z})) \cdot \|\mathbf{Z}\|_F^2 - 4Z_{ij}E(\mathbf{Z})) \cdot \|\mathbf{Z}\|_F^{-6}. \quad (7)$$

This allows us to construct a variable-metric quasi-Newton method akin to BFGS (see [7, section 10.7]), which is reputed to be remarkably effective for multidimensional low-degree polynomial problems such as ours. Like BFGS, we will perform line searches for minima along the gradient. However, our problem affords a much more efficient and accurate strategy: An explicit *non-local* solution for the line minimum

$$x^* = \min_x E'(\mathbf{Z} + x \cdot \mathbf{D}) \qquad (8)$$

where $\mathbf{D} \propto \nabla E'(\mathbf{Z})$ is a unit vector in the direction of the gradient. Appendix A gives the optimal x^* in closed form; because $E'(\cdot)$ is norm-invariant one can then project by scaling from the line minimum $\mathbf{Z} + x^* \cdot \mathbf{D}$ back onto the surface $\|\mathbf{Z}\|_F = 1$ without changing the error. Furthermore, because the error has 180° symmetry ($E(\mathbf{Z}) = E(-\mathbf{Z})$) and the line projects to a half great-circle, the minimum is global over the entire subspace, not just the immediate valley, and that gradient $\mathbf{D}$ will never need to be explored again. In short, in each step we compute in closed form the globally optimal hop in the direction of maximal error reduction.

Because the XCK algorithm gives a different result for each subset of K linearly independent frames, some better, some worse, we have reason to expect that the error surface has local optima in the noisy case. The empirical question of whether we find high-quality optima is answered in the experimental section. We are studying the question of whether there are local optima given noiseless "perfect" data. We suspect that the question is mooted in practice by the nonlocality of the line search. The XCK result implies that the error surface of perfect data is convex in the vicinity of a solution, and indeed, all 10^5 Monte Carlo trials with perfect data produced global optima.

4 The full correction matrix

In the noiseless case, only one estimate of $\mathbf{G}_{1:3}$ is needed to solve the problem, because one can obtain all rotation information from the first three columns of the motion matrix $\mathbf{M}_{1:3} = \tilde{\mathbf{M}}\mathbf{G}_{1:3}$, and from these rotations solve for the rest of $\mathbf{G}$, as follows: Let $\mathbf{z}_f^T$ be the "missing" third row of the truncated rotation matrix $\mathbf{R}_f \propto \tilde{\mathbf{M}}_f\mathbf{G}_{1:3} \in \mathbb{R}^{2\times 3}$. $\mathbf{z}_f^T$ gives the projection in z (depth), and is thus the cross-product of the two rows in $\tilde{\mathbf{M}}_f\mathbf{G}_{1:3}$. Because $\mathbf{R}_f\mathbf{z}_f = \mathbf{0}$. we have $c_{1f}\mathbf{R}_f\mathbf{z}_f = \tilde{\mathbf{M}}_f\mathbf{G}_{1:3}\mathbf{z}_f = \mathbf{0}$, $c_{2f}\mathbf{R}_f\mathbf{z}_f = \tilde{\mathbf{M}}_f\mathbf{G}_{4:6}\mathbf{z}_f = \mathbf{0}$, etc. These constraints can be collectively expressed as $\forall_f\,(\mathbf{z}_f^T \otimes \tilde{\mathbf{M}}_f)\mathbf{N} = \mathbf{0}$ with $\mathbf{N} \doteq [\text{vec}\,\mathbf{G}_{1:3}, \text{vec}\,\mathbf{G}_{4:6}, \cdots]$. Thus the (possibly approximate) row-nullspace

$$\mathbf{N} \doteq \text{null}\left(\begin{bmatrix} \mathbf{z}_1^T \otimes \tilde{\mathbf{M}}_1 \\ \vdots \\ \mathbf{z}_F^T \otimes \tilde{\mathbf{M}}_F \end{bmatrix}\right) \in \mathbb{R}^{9K\times K} \tag{9}$$

specifies the corrective transform, with $\mathbf{G} = \text{vec}_{3K}\mathbf{N}$ for any rotation and rescaling of the columns of $\mathbf{N}$. (Note that since $\mathbf{N}$ is semi-orthogonal, $\mathbf{G}$ has spectral radius $\rho(\mathbf{G}) \leq 1$.) The orthogonalized motion matrix $\tilde{\mathbf{M}}\mathbf{G}$ can then be factored into rotations and weights.

5 Combining multiple estimates

When the data is noisy, we can improve the final result by combining information from diverse optimal of the error surface. We make multiple estimates of column-triads of $\mathbf{G}$, rotate them to a common coordinate frame, and combine their corrections to get an improved estimate of the rotations by factoring the matrix $\tilde{\mathbf{M}}[\mathbf{G}_{1:3}, \mathbf{G}'_{1:3}, \mathbf{G}''_{1:3}, \cdots]$. To get diverse estimates of $\mathbf{G}_{1:3}$, one can either begin the optimization with different random initial conditions, or add the constraint that the deformation weights in K frames are orthogonal, *i.e.*, $[\mathbf{c}_f, \mathbf{c}_g, \cdots] \in \mathbb{R}^{K\times K}$ has orthogonal rows. (This is a weaker assumption that the XCK method, which sets $[\mathbf{c}_f, \mathbf{c}_g, \cdots] = \mathbf{I}_K$, but has the same effect of producing successive column-triads of $\mathbf{G}$.) This constraint can be expressed in terms of proto-motion matrix $\tilde{\mathbf{M}}$, an already estimated $\mathbf{G}_{1:3}$, and a new estimate $\mathbf{Z}$ of an alternate column-triad in $\mathbf{G}$, as follows: Let $\tilde{\mathbf{M}}'_x \subset \tilde{\mathbf{M}}$ and $\tilde{\mathbf{M}}'_y \subset \tilde{\mathbf{M}}$ be row-subsets of $\tilde{\mathbf{M}}$ corresponding to the x and y projections observed in K frames. Then from the orthogonality structure of $\mathbf{M}$, $\text{trace}((\tilde{\mathbf{M}}'_x\mathbf{G}_{1:3})^T(\tilde{\mathbf{M}}'_x\mathbf{Z})) = \text{trace}((\tilde{\mathbf{M}}'_y\mathbf{G}_{1:3})^T(\tilde{\mathbf{M}}'_y\mathbf{Z})) = 0$. Rearranging, we obtain the constraint $0 = \mathbf{H}^T\text{vec}(\mathbf{Z})$ with

$$\mathbf{H} \doteq [\text{vec}(\mathbf{G}_{1:3}^T\tilde{\mathbf{M}}_x'^T\tilde{\mathbf{M}}'_x), \text{vec}(\mathbf{G}_{1:3}^T\tilde{\mathbf{M}}_y'^T\tilde{\mathbf{M}}'_y)] \in \mathbb{R}^{9K\times 2}, \tag{10}$$

which is trivially incorporated into the optimization of section 3 by projecting $\mathbf{Z}$ and $\mathbf{D}$ into the nullspace of $\mathbf{H}$.

5.1 Rotating to a common frame

Each estimate of $\mathbf{G}_{1:3}$ is determined up to a global 3D rotation. We make them all rotationally consistent by enforcing orthogonality constraints: Let $\mathbf{G}_{1:3}$ and $\mathbf{G}'_{1:3}$ be two estimates of the first three columns of the correction matrix. Taking row-vector $\mathbf{x}_f \in \tilde{\mathbf{M}}\mathbf{G}_{1:3}$ and $\mathbf{y}'_f \in \tilde{\mathbf{M}}\mathbf{G}'_{1:3}$ the rotation $\mathbf{E} \in \mathbb{R}^{3\times 3}$ that aligns the two estimates should satisfy $0 = \mathbf{x}_f\mathbf{E}\mathbf{y}_f'^T = \mathbf{y}_f\mathbf{E}\mathbf{x}_f'^T = \mathbf{x}_f\mathbf{E}(\mathbf{x}'_f \times \mathbf{y}'_f)^T$ in a least-squares sense. Rewriting $\mathbf{x}_f\mathbf{E}\mathbf{y}_f'^T = (\mathbf{x}_f \otimes \mathbf{y}'_f)(\text{vec}\,\mathbf{E})$, we see that vec $\mathbf{E}$ is orthogonal to all the vectors formed by these Kronecker products. If the data were noiseless we could collect all such constraint vectors into a matrix $\mathbf{L} \in \mathbb{R}^{6F\times 9}$, compute its 1-dimensional row-nullspace, and reshape to a 3×3 matrix to obtain $\mathbf{E}$ (up to scale).

With noise, the matrix of accumulated orthogonality constraints $\mathbf{L}$ most likely will not have a nullspace; even if it does, it is not likely to form an orthonormal matrix. The gram matrix $\mathbf{Q}_C \doteq \mathbf{L}^T\mathbf{L}$ gives us the error metric $E_{\mathbf{Q}_C}(\mathbf{E}) = (\text{vec}\,\mathbf{E})^T\mathbf{Q}_C(\text{vec}\,\mathbf{E})$ in the space of orthonormal matrices. Projecting the approximate 1D nullspace (the minimizing eigenvector of $\mathbf{Q}_C$) to the nearest orthonormal matrix does not minimize the error, because the result has nonzero projection onto the remaining eigenvectors of $\mathbf{Q}_C$. In some cases it is not even a good approximation.

Again, this is a constrained optimization problem: $\min_{\mathbf{E}\in SO_n(\mathbb{R})} E_{\mathbf{Q}_C}(\mathbf{E})$. This turns out to be solvable in closed form for SO_2 rotations and the solution can be leveraged into a fast numerical method for SO_3 rotation matrices (where there probably is no closed form solution). We make an initial estimate of $\mathbf{E}$ by projecting the approximate nullspace of $\mathbf{Q}_C$ to the nearest orthonormal matrix. We then refine this estimate by solving for a series of plane rotations of $\mathbf{E}$ that each minimize the error. Let $\mathbf{E}' \doteq \mathbf{E}\begin{bmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{bmatrix}$ with circle constraint $c^2 + s^2 = 1$. The derivatives of $E_{\mathbf{Q}_C}(\mathbf{E}')$ with respect to c, s are linear; setting these to zero we obtain constraints of the form $\begin{bmatrix} a & b \\ d & e \end{bmatrix}\begin{bmatrix} c \\ s \end{bmatrix} = \begin{bmatrix} d \\ f \end{bmatrix}$. Solving this in a least-squares sense under the circle constraint yields a quartic polynomial in c. At the real roots of this polynomial the Laplacian $\bigtriangledown^2 E_{\mathbf{Q}_C}(\mathbf{E}')$ is zero; we select the root minimizing the error and apply the corresponding plane rotation to $\mathbf{E}$. This converges after just a few iterations through the three orthogonal planes of $\mathbf{E}$ (and could be applied to any plane in $\mathbf{E}$ by rotating the coordinate system).

6 Experiments

We implemented the XCK basis method and, with some trial and error, devised "cube-and-points" data that qualitatively reproduces the error curves reported in [13]. The direct

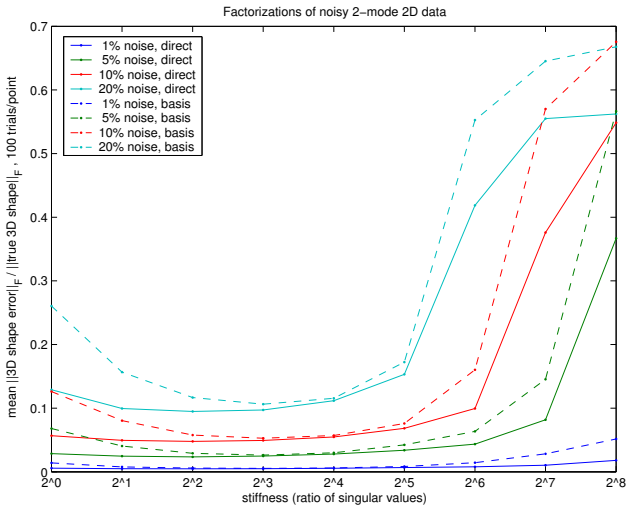

Figure 1: Reconstruction error of direct solution (solid lines) and basis solution (dashed lines) on $F = 32$ frames of $N = 40$ point synthetic $K = 2$ mode data, plotted as a function of surface stiffness (horizontal axis) and data noise (each curve). Each point averages 100 trials. Graph shows 3D shape reconstruction error as a fraction of the data norm.

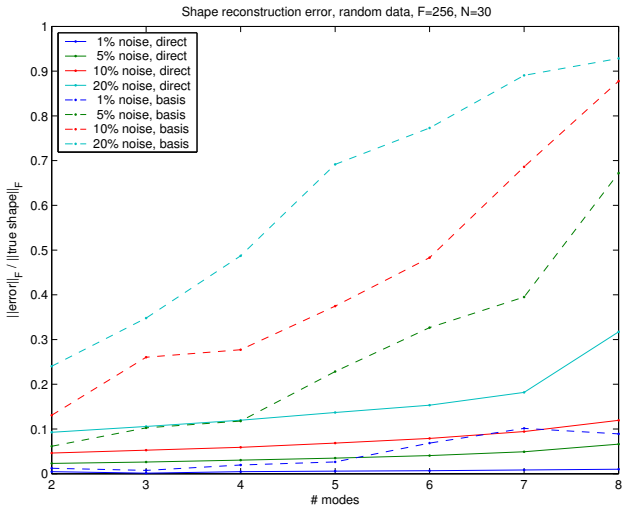

Figure 2: Reconstruction error of direct solution (solid lines) and basis solution (dashed lines) on $F = 256$ frames of synthetic K-mode data, plotted as a function of the number of modes (horizontal axis) and data noise (each curve). Graph shows 3D shape reconstruction error as a fraction of the data norm. Each point averages 100 trials.

method produces slightly lower error curves, however, we will use the space here to detail results using $\sim 10^4$ randomly generated shapes, in order to more broadly stress-test the algorithms. In these experiments the direct method combines estimates of $\mathbf{G}_{1:3}$, each obtained from random initial conditions. For fairness, we also took the best of K different outputs from the basis method. XCK do not say how their estimates are rotated to a common coordinate frame; we tried several methods and kept the one that gave best results.

As in [13], we generated synthetic 2-mode data that varied by stiffness and noise content (figure 1). Stiffness is measured as the ratio of the singular values of the shape variation matrix $\mathbf{S}^{(3)}\mathbf{C} \in \mathbb{R}^{3N\times F}$—in this experiment a rank $K = 2$ matrix containing full noiseless 3D information for N points in F frames. White gaussian noise $\mathbf{Y}$ added to the observations $\mathbf{P}$ is measured as $\|\mathbf{Y}\|_F/\|\mathbf{P}\|_F$. Reconstruction error is reported as $\|\hat{\mathbf{X}}-\mathbf{X}\|_F/\|\mathbf{X}\|_F$, where $\hat{\mathbf{X}},\mathbf{X}$ are the estimated and true 3D locations of all points in all frames. Both methods give exact solutions for noiseless data, modulo some floating-point error; the direct solution is more susceptible to floating-point error due to the evaluation of quartics but this can be remediated using extended-precision floats. The situation reverses even with a tiny amount of data noise: The direct method strongly outperforms the basis method for all noise settings, as shown in all figures. Both algorithms used double-precision floating point numbers. The average performance gaps are statistically significant at $p < 10^{-3}$ levels.

The algorithms share a failure mode, shown at the right side of the graphs, where error trends up sharply: When the data contains a very minor deformation mode that is polluted by noise, both algorithms can produce garbage. Because the noisy mode has such small magnitude, it can have smallest residual w.r.t. *any* combination of orthogonality and basis constraints, thereby polluting *all* estimated column-triads of $\mathbf{G}$ and defeating either factorization. At present the only remedy is to further truncate the initial data SVD in hopes of dropping the noisy mode.

With $K = 2$ modes the direct solution can be seen to be slightly more robust to noise; this gap widens dramatically in figure 2, where the direct solution proves to be substantially more robust to noisy high-DOF problems. Increasing the number of points and frames will lower the error of both methods; with sufficient frames we can coax correct solutions from the basis method.

One thing that distinguishes real-world problems from the experiments above is that real faces have many, many degrees of freedom, and the singular value spectrum of facial tracking data does not often suggest a clear cutoff value for the number of deformation modes K. Arguably, this case was not explored in the XCK experiments because their tracking was done with an active appearance model—a subspace-constraint tracker that guarantees a clear and early cutoff in the singular value spectrum. To compare factorization methods on more realistic data, we tested them on unconstrained visual tracking data and motion capture data.

Figure 3 depicts a model obtained from 1000 frames of

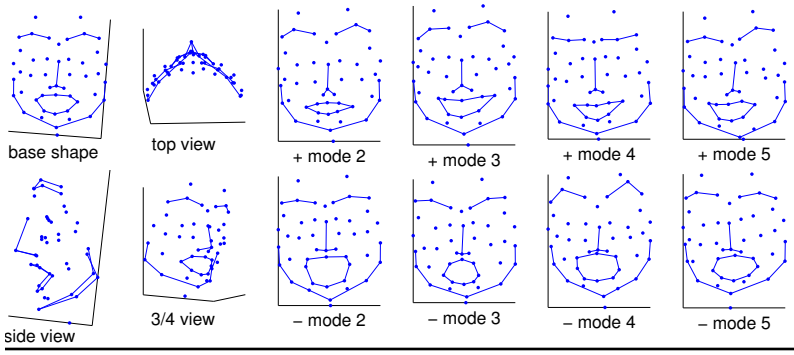

Figure 3: $K = 5$ mode model factored from $F = 1000$ frames of $N = 46$ points 2D tracking data. Eyebrows, jawline, and mouth are outlined. Lines also connect the bridge, nose tip, and two points directly above the upper lip. At the left are four views of the first mode, the average 3D shape. At right the four remaining modes are applied at (considerably exaggerated) ± 3 standard deviation magnitudes.

2D point data by the direct method. In order to have ground truth, we took 1000 frames of noisy 3D motion-capture data and made random 2D projections with up to 22.5° horizontal rotation and 11.25° vertical rotation—small enough to guarantee no occlusions. The resulting tracking data matrix masses 99% of its variance in the first 13 singular values, suggesting a $K \geq 5$ mode basis in which two deformations might be rank-deficient (planar). Consequently, we were unable to coax a nondegenerate result from the basis method, which depends on all deformations having full rank (being 3D). In contrast, the direct method succeeded in finding estimates of $\mathbf{G}_{1:3}$ based on the nondegenerate modes, and produced a solution that reconstructed the ground-truth 3D data with less than 0.2% relative error.

Figure 4 compares the base shapes factored from $F = 790$ frames of visually tracked features. To obtain this data, a student's face was marked so that optical flow succeeded without introducing any constraints that might artificially reduce the rank or otherwise alter the singular value spectrum of the tracking data. Although imperfect, most tracks were quite good. An SVD suggested $K \geq 21$ modes of deformation (at 99% variance); we got best results from both methods factoring at $K = 6$ to retain 93% of the data variance. The base shapes were obtained via principal components analysis of the output shape bases, weighted by the norms of their associated deformation coefficient vectors. As figure 4 shows, the direct method produced a reasonable profile, while the basis method output has several depth-inverted facial features.

7 Comments

We have presented a solution to the nonrigid structure-from-motion problem. By addressing the orthogonal structure of the problem in a constrained optimization setting, we can solve directly for the key variable of interest, sidestepping the degeneracies that arise in attempts to generalize the orthogonalizing step of the Tomasi-Kanade rigid SFM factorization. By replacing the "search" component of this optimization with a closed-form solution, we obtain a fast and efficient algorithm. Numerical experiments confirm that it produces exact solutions from perfect (*i.e.*, algebraically conforming) data and superior solutions from real and/or noisy data.

This approach has two limitations: Computing quartic roots reduces numerical precision, and attempting to model deformations whose magnitudes are near noise levels can contaminate all estimates. We are investigating the possibility that the latter problem can be fixed by modifying the constrained optimization to avoid solutions near the nullspace of the proto-motion matrix $\tilde{\mathbf{M}}$.

References

[1] M. Brand. Morphable 3D models from video. In *Proc. CVPR'01*, 2001.

[2] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *Proc. CVPR'00*, 2000.

[3] A. D. Bue and L. Agapito. Non-rigid 3d shape recovery using stereo factorization. In *Proc. ACCV'04*, pages 25–30, 2004.

[4] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An introduction to computational algebraic geometry and commutative algebra*. Springer-Verlag, 1991.

[5] H. Li, P. Roivainen, and R. Forchheimer. 3-D motion estimation in model-based facial image coding. *IEEE Trans. PAMI*, 15(6):545–555, June 1993.

[6] F. I. Parke. *A parametric model for human faces*. PhD thesis, University of Utah, 1974.

[7] W. Press, B. Flannery, S. Teuskolky, and W. Vetterling. *Numerical Recipes*. Cambridge University Press, 1992.

[8] D. Terzopoulos and K. Waters. Physically-based facial modeling, analysis, and animation. *J. of Visualization and Computer Animation*, 1(4):73–80, 1990.

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

[10] L. Torresani and A. Hertzmann. Automatic non-rigid 3D modelling from video. In *Proc. ECCV'04*, 2004.

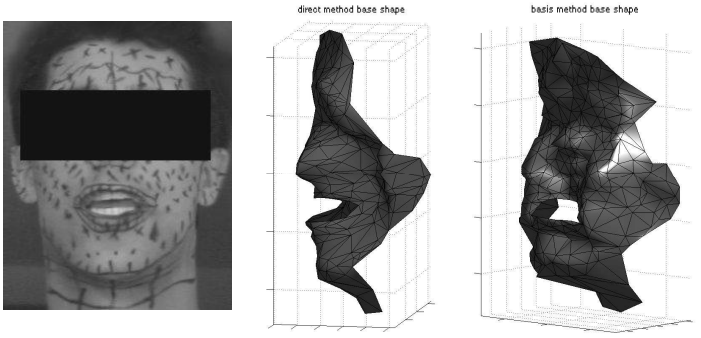

Figure 4: After optical flow tracking of $N = 352$ marked facial features in $F = 790$ frames, factorization by the direct method yields a base shape with good enough inferred depth for a profile (middle), while the basis method output depth-inverts the mouth, lower nose, and forehead (right, shown in 3/4 view because the cheeks occlude the inverted features in profile).

[11] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3D shape from 2D motion. In *Proc. NIPS 16*, 2004.

[12] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. CVPR'01*, 2001.

[13] J. Xiao, J.-X. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Proc. ECCV'04*, 2004.

A Line minimization

See section 3 for preliminaries. We seek a line optimum of $E'(\mathbf{Z}')$ at point $\mathbf{Z}' \doteq \mathbf{Z} + x \cdot \mathbf{D}$, indexed by x. Define scale-invariant error $E'(\mathbf{Z}') = \mathbf{v}^T \mathbf{Q}_A \mathbf{v}$ with $\mathbf{v} \doteq \text{vech}(\mathbf{Z}'\mathbf{Z}'^T)/\|\mathbf{Z}'\|_F^2$. Differentiating,

$$\frac{dE'(\mathbf{Z}')}{dx} = \frac{d}{dx}\mathbf{v}^T\mathbf{Q}_A\mathbf{v} = 2\mathbf{v}^T\mathbf{Q}_A\frac{d\mathbf{v}}{dx} \qquad (11)$$

with

$$\|\mathbf{Z}'\|_F^4 \frac{d\mathbf{v}}{dx} = \frac{d\text{vech}(\mathbf{Z}'\mathbf{Z}'^T)}{dx}\|\mathbf{Z}'\|_F^2 - \text{vech}(\mathbf{Z}'\mathbf{Z}'^T)\frac{d\|\mathbf{Z}'\|_F^2}{dx}$$

We drop the denominators of $\mathbf{v}$ and $\frac{d\mathbf{v}}{dx}$ because they do not affect the zeros of the derivative $\frac{dE'(\mathbf{Z}')}{dx}$. After copious algebra, we obtain scaled vectors

$$\begin{aligned}
\|\mathbf{Z}'\|_F^2 \cdot \mathbf{v} &= \text{vech}(\mathbf{D}\mathbf{D}^T)x^2 + \text{vech}(\mathbf{B})x + \text{vech}(\mathbf{Z}\mathbf{Z}^T), \\
\|\mathbf{Z}'\|_F^4 \frac{d\mathbf{v}}{dx} &= -\text{vech}(\|\mathbf{D}\|_F^2\mathbf{B} - 2\text{trace}(\mathbf{B})\mathbf{D}\mathbf{D}^T)x^2 \\
&\quad + \text{vech}(\|\mathbf{Z}\|_F^2\mathbf{A} - 2\text{trace}(\mathbf{A})\mathbf{Z}\mathbf{Z}^T)x \\
&\quad + \text{vech}(\|\mathbf{Z}\|_F^2\mathbf{B} - 2\text{trace}(\mathbf{B})\mathbf{Z}\mathbf{Z}^T), \\
&\text{with } \mathbf{A} = 2 \cdot \mathbf{D}\mathbf{D}^T \text{ and } \mathbf{B} = \mathbf{D}\mathbf{Z}^T + \mathbf{Z}\mathbf{D}^T.
\end{aligned}$$

Since both vectors are quadratic in x, the scaled derivative $\frac{dE'(\mathbf{Z}')}{dx} \propto \mathbf{v}^T\mathbf{Q}_A\frac{d\mathbf{v}}{dx} \cdot \|\mathbf{Z}'\|_F^6$ is quartic in x and thus solvable in closed form, with the negative real root(s) indexing the extrema of $E'(\mathbf{Z}')$ along the line. Since the error is invariant to rescalings of $\mathbf{Z}$, the line minimization is effectively a constrained optimization on a sphere (having 180° symmetry, because because $E(\mathbf{Z}) = E(-\mathbf{Z})$.). Geometrically, the optimization line projects to a half great-circle oriented with the gradient on the sphere; x indexes the error on that arc. W.l.o.g. we may exploit this spherical geometry by forcing $\|\mathbf{Z}\|_F = \|\mathbf{D}\|_F = 1$ and $\text{trace}(\mathbf{D}^T\mathbf{Z}) = (\text{vec}\mathbf{D})^T(\text{vec}\mathbf{Z}) = 0$ at each step; the results are identical but several terms in the gradient formulæ vanish, improving numerical precision. The gradient $\mathbf{D}$ can be further constrained with regard to previous gradients or an estimated Hessian, yielding a highly efficient procedure similar to conjugate gradient search and BFGS. Typically the number of iterations to convergence is a small multiple of the number of unknowns.

B Symbols and operators

In the following, N=#points, F=#frames, K=#modes, SO_n=special orthogonal group, GL_n=general linear group.

variable & dimension	meaning & discussion
$\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_F]$, $K \times F$	deformation weights, §2
$\mathbf{D}$, $3K \times 3$	error gradient, §3
$\mathbf{E}$, 3×3	SO_3 nuisance DOF, §2.1, 5.1
$\mathbf{F}$, $K \times K$	GL_K nuisance DOF, §2.1
$\mathbf{G} = [\mathbf{G}_{1:3}, \cdots]$, $3K \times 3K$	corrective transform, §2
$\mathbf{H}$, $3K \times 2$	diversity constraint, §5
$\mathbf{I}_n$, $n \times n$	identity matrix, §2
$\mathbf{L}$, varies	orthogonality constraints, §2
$\mathbf{M} = \tilde{\mathbf{M}}\mathbf{G}$, $2F \times 3K$	motion matrix, §2
$\mathbf{N} = \text{vec}_{9K}\mathbf{G}$, $9K \times K$	nullspace estimate, §4
$\mathbf{P} = \mathbf{M}\mathbf{S}$, $2F \times N$	observation data, §2
$\mathbf{Q}_A, \mathbf{Q}_B, \mathbf{Q}_C$, varies	error metrics, §2, 2.1, 5.1
$\mathbf{R}$, 2×3	projective rotation, §2
$\mathbf{S} = \mathbf{G}^{-1}\tilde{\mathbf{S}}$, $3K \times N$	shape basis, §2
$\mathbf{Z} = \mathbf{G}_{1:3}$, $9K \times 3$	optimization variable, §3, A

operator	Matlab & octave equivalents for $\mathbf{A} \in \mathbb{R}^{r \times c}$
$\mathbf{A} \otimes \mathbf{B}$	`kron(A,B)`
vec $\mathbf{A}$	`A(:)`
$\text{vec}_n\mathbf{A}$	`reshape(A,n,r*c/n)`
vech $\mathbf{A}$	`vec(triu(A))`
$\text{vc}(\mathbf{a}, \mathbf{b})$	`vech(a*b'+b*a'-diag(a.*b))`
$\mathbf{A}^{(n)}$	`vec`$_{n*c}$`(permute(reshape(A,n,r/n,c),[1,3,2]))`