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

Morphable 3D models from video

TR2001-37 May 2001

Abstract

Nonrigid 3D structure-from-motion and 2D optical flow can both be formulated as tensor factorization problems. The two problems can be made equivalent through a noisy affine transform, yielding a combined nonrigid structure-from-intensities problem that we solve via structured matrix decompositions. Often the preconditions for this factorization are violated by image noise and deficiencies of the data vis-a-vis the sample complexity of the problem. Both issues are remediated with careful use of rank constraints, norm constraints, and integration over uncertainty in the intensity values, yielding novel solutions for SVD under uncertainty, factorization under uncertainty, nonrigid factorization, and subspace optical flow. The resulting integrated algorithm can track and 3D-reconstruct nonrigid surfaces that have very little texture, for example the smooth parts of the face. Working with low-resolution low-texture "found video," these methods produce good tracking and 3D reconstruction results where prior algorithms fail. NB: Winner, IEEE CVPR 2001 Best Paper Award.

Proceedings, 2001 Conference on Computer Vision and Pattern Recognition, CVPR2001

MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Morphable 3D models from video

Matthew Brand

TR-2001-37 December 2001

Abstract

Nonrigid 3D structure-from-motion and 2D optical flow can both be formulated as tensor factorization problems. The two problems can be made equivalent through a noisy affine transform, yielding a combined nonrigid structure-from-intensities problem that we solve via structured matrix decompositions. Often the preconditions for this factorization are violated by image noise and deficiencies of the data vis-a-vis the sample complexity of the problem. Both issues are remediated with careful use of rank constraints, norm constraints, and integration over uncertainty in the intensity values, yielding novel solutions for SVD under uncertainty, factorization under uncertainty, nonrigid factorization, and subspace optical flow. The resulting integrated algorithm can track and 3D-reconstruct nonrigid surfaces that have very little texture, for example the smooth parts of the face. Working with low-resolution low-texture "found video," these methods produce good tracking and 3D reconstruction results where prior algorithms fail.

1st draft circulated October 2000; submitted to CVPR May 2001; accepted September 2001; published December 2001.

Morphable 3D models from video

Matthew Brand
Mitsubishi Electric Research Labs, Cambridge, MA 02139 USA

Abstract

Nonrigid 3D structure-from-motion and 2D optical flow can both be formulated as tensor factorization problems. The two problems can be made equivalent through a noisy affine transform, yielding a combined nonrigid structure-from-intensities problem that we solve via structured matrix decompositions. Often the preconditions for this factorization are violated by image noise and deficiencies of the data vis-a-vis the sample complexity of the problem. Both issues are remediated with careful use of rank constraints, norm constraints, and integration over uncertainty in the intensity values, yielding novel solutions for SVD under uncertainty, factorization under uncertainty, nonrigid factorization, and subspace optical flow. The resulting integrated algorithm can track and 3D-reconstruct nonrigid surfaces that have very little texture, for example the smooth parts of the face. Working with low-resolution low-texture "found video," these methods produce good tracking and 3D reconstruction results where prior algorithms fail.

1. Introduction

The problem of acquiring 3D morphable models of nonrigid objects has attracted intense interest in computer vision since the advent of deformable and eigen-models in the 1980s. Current solutions address special cases of the problem that are well-constrained by additional information. For example, when depth estimates are available from multiple cameras or laser range-finders; when the poses or articulations are fixed or chosen from a maximally informative set; when the surface is decorated with special textures or markers to make inter-frame correspondences obvious; or when structured light is used to reveal its contours. These methods require various combinations of high-quality high-resolution sources, calibrated cameras, special lighting, and careful posing. A second class of solutions relaxes image constraints but depends on having a precomputed class of possible models [1] or motions (as used in [3] for tracking).

In this paper we consider a relatively unconstrained case: Single-camera video in which the surface is freely moving and articulating. There are no shape or motion priors. We only require that the surface be at least sparsely textured, and that lighting changes, if any, be slow relative to the object's physical motion. The texture can be partially degenerate everywhere the image is sampled, as long as it is not all degenerate in the same direction. We consider low-quality sources that are difficult to constrain, for example, pre-existing footage or home movies of young children. In this paper we will work with faces and video but the methods

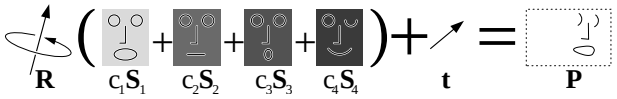
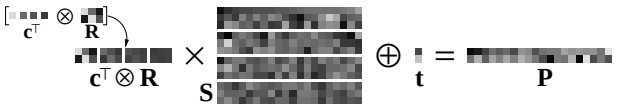


Figure 1: Image formation: Morph bases ($\mathbf{S}$) are summed according to weights ($\mathbf{c}$), rotated ($\mathbf{R}$), and translated ($\mathbf{t}$) to give the image projection ($\mathbf{P}$). To infer $\mathbf{S}, \mathbf{c}, \mathbf{R}, \mathbf{t}$ from $\mathbf{P}$ is it convenient to re-order these operations as in eqn. (1), depicted here with matrix images:



are general and apply to any flexible object observed in 2D (image) or 3D (volumetric) sequences.

Our result is a factorization algorithm for 3D nonrigid structure and motion from video that finds 2D correspondences in the course of enforcing 3D geometric invariants. Taking the Tomasi & Kanade [8] rigid-body factorization as a starting point, we reconsider the uncertainty formulation introduced by Irani & Anandan [6], the subspace formulation for optical flow introduced by Irani [5], and the nonrigid extension proposed by Bregler, Hertzmann, & Biermann [3]. Noting their common theme—geometric invariants expressed as rank constraints—we generalize and integrate the constraints from these three subproblems. Our solutions are substantially different from those of [6, 5, 3], reflecting our identification of new constraints, new solution methods, and corrections to errors in the prior literature.

2. Notation

We use matrix tensor operators and highly recommend [7] as an introduction and [4] for usage examples. a is a scalar, $\mathbf{a}$ is a vector, $\mathbf{A}$ is a matrix; $[\Rightarrow_i \mathbf{A}_i], [\Downarrow_i \mathbf{A}_i], [\searrow_i \mathbf{A}_i]$ are horizontal, vertical, and diagonal concatenations, respectively. $\mathbf{I}$ is the identity matrix; $\mathbf{0}$ and $\mathbf{1}$ are the zero and one matrices. Matrix dimensions are indicated in subscripts (e.g., $\mathbf{A}_{r\times c}$) or determined by conformance. $\mathbf{A}^\top$ denotes transpose; vector-transpose $\mathbf{A}^{(i)}$ transposes matrix $\mathbf{A}$ with each vertical group of i elements treated as a unit; block-transpose $\mathbf{A}^{(i,j)}$ does the same treating each block of $i \times j$ elements as a unit. $\otimes$ denotes Kronecker (tensor) product; $\odot$ denotes Hadamard (element-wise) product; $\oplus$ denotes tiled addition, e.g., $\mathbf{A}_{6\times 2} \oplus \mathbf{B}_{2\times 2} = \mathbf{A}_{6\times 2} + (\mathbf{1}_{3\times 1} \otimes \mathbf{B}_{2\times 2})$. vec $\mathbf{A}$ vectorizes $\mathbf{A}$ by stacking its columns and $\text{vec}_i\, \mathbf{A}_{r\times c} = (\text{vec}\, \mathbf{A})^{(i)}$ folds $(\text{vec}\, \mathbf{A})_{rc\times 1}$ into a matrix having rc/i columns of i elements each. $\mathbf{A}/\mathbf{B}$ and $\mathbf{B}\backslash\mathbf{A}$ denote right and left division; $\mathbf{A}^\dagger$ denotes Moore-Penrose pseudo-inverse.

3. Setting

We begin with a simple model of image formation, depicted in figure 1. Observed shape is a weighted sum of morph bases, rotated in 3D, projected onto the image plane, and translated in that plane. We write the projection in frame f as

$$\mathbf{P}_f = (\mathbf{c}_f^\top \otimes \mathbf{R}_f)\mathbf{S} \oplus \mathbf{t}_f \tag{1}$$

The rows of $\mathbf{S}$ contain the x, y, z ordinates of the K morph bases for N points: $\mathbf{S}_{3K\times N} \doteq [\mathbf{s}_{1x}, \mathbf{s}_{1y}, \mathbf{s}_{1z}, \mathbf{s}_{2x}, \mathbf{s}_{2y}, \mathbf{s}_{2z}, ..., \mathbf{s}_{Kx}, \mathbf{s}_{Ky}, \mathbf{s}_{Kz}]^\top$. Without loss of generality, we assume that the row sums are zero (written $\mathbf{S}\mathbf{1}_{3K\times 1} = \mathbf{0}$). By convention, the first morph basis gives a scalable mean shape and subsequent morphs deform it. These are combined according to the vector $\mathbf{c}_{K\times 1}$ of morph weights, which fixes both expression and scale. The orthonormal matrix $\mathbf{R}_{D\times 3}$ effects a 3D rotation and 2D projection (for $D = 2$), then $\mathbf{t}_{D\times 1}$ translates the projection in the image plane. This is a weak perspective model, an approximation to full perspective projection that works well when the depth variation within the object is small relative to the object's distance from the camera—typically the case for consumer camera videography.

For $F \gg K$ frames we define $\mathbf{M}_f \doteq \mathbf{c}_f^\top \otimes \mathbf{R}_f$, $\mathbf{C}_{K\times F} \doteq [\Rightarrow_f \mathbf{c}_f]$, and $\mathbf{T}_{DF\times 1} \doteq [\Downarrow_f \mathbf{t}_f]$, with projections

$$\mathbf{P}_{DF\times N} \doteq [\Downarrow_f \mathbf{P}_f] = \mathbf{M}\mathbf{S} \oplus \mathbf{T}, \quad \text{where} \tag{2}$$

$$\mathbf{M}_{DF\times 3K} \doteq [\Downarrow_f \mathbf{M}_f] = [\searrow_f \hat{\mathbf{R}}_f](\mathbf{C}^\top \otimes \mathbf{I}_3). \tag{3}$$

This is depicted in figure 2. Much of this paper will be devoted to the special structure of the motion matrix $\mathbf{M}$.

Our first goal is to infer $\mathbf{S}, \mathbf{R}, \mathbf{C}$, and $\mathbf{T}$ from the inter-frame correspondences in $\mathbf{P}$. Often these correspondences are unavailable or very hard to compute; in §6 we will leverage our analysis into an algorithm that estimates all variables including $\mathbf{P}$ directly from video.

Assuming for now that that all points are observed in all frames, the translations $\hat{\mathbf{T}}$ can be estimated as the row-means of $\mathbf{P}$ and then removed from $\mathbf{P}$ so that all rows in $\mathbf{P} \ominus \hat{\mathbf{T}}$ are zero-mean. Then $\mathbf{P} \ominus \hat{\mathbf{T}}$ can factor into pseudo-motion matrix $\tilde{\mathbf{M}}$ and pseudo-shape/morph basis matrix $\tilde{\mathbf{S}}$. $\tilde{\mathbf{M}}$ in turn can decompose into pseudo-rotations and pseudo-morph weights. There are infinitely many such factorizations and we must solve for one that yields proper rotations and maximal error reduction per morph. As with many multilinear phenomena in image formation, the key to a successful factorization will be the identification and exploitation of rank and norm constraints on substructures in these matrices.

3.1. Rigid-body factorization

In the $K = 1$ case of *rigid-body motion*, the rank theorem of Tomasi & Kanade [8] asserts that a rank-3 thin singular value decomposition (SVD) $\tilde{\mathbf{M}}\tilde{\mathbf{S}} \xleftarrow{\text{SVD}_3} \mathbf{P} \ominus \mathbf{T}$ will factor motion and shape information from tracking data. The pseudo-motion matrix $\tilde{\mathbf{M}}$ of left singular vectors associated with the three largest singular values contains the 3D rotation/scale information; the matching right singular vectors form the pseudo-shape matrix $\tilde{\mathbf{S}}$. Assuming that rigid shape statistically dominates the data in $\mathbf{P}$, the remaining vectors contain information about violations of the rigid-motion assumption, e.g., nonrigidities and tracking noise.

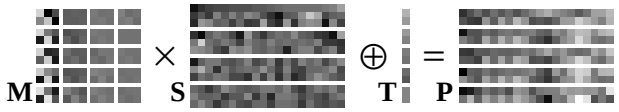


Figure 2: The forward model for multiple frames (eqn. (2)), showing the structure of the motion matrix $\mathbf{M}$ (eqn. (3)). Each block in $\mathbf{M}$ is a scaled rotation matrix; its rows have equal norm and are orthogonal. Moreover, its first row is orthogonal to any second row taken from blocks to the left and right. An SVD of $\mathbf{P} \ominus \mathbf{T}$ produces a pseudo-motion matrix $\tilde{\mathbf{M}} = \mathbf{M}\mathbf{J}$ and pseudo-shape matrix $\tilde{\mathbf{S}} = \mathbf{J}^{-1}\mathbf{S}$, where $\mathbf{J}$ is an arbitrary unknown full-rank matrix. Successful factorization thus depends on finding a correction matrix $\mathbf{J}$ that will restore the appropriate structure to $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{S}}$.

3.2. Corrective transform

The SVD determines both sides up to an invertible 3D affine transformation $\mathbf{G}_{3\times 3}$ such that $\mathbf{M}\mathbf{S} = (\tilde{\mathbf{M}}\mathbf{G}^{-1})(\mathbf{G}\tilde{\mathbf{S}}) = \tilde{\mathbf{M}}\tilde{\mathbf{S}} = \mathbf{P}$; one must solve for a $\mathbf{G}^{-1}$ that restores orthogonal structure to $\tilde{\mathbf{M}}$ in order to get proper rotations and shape. Let the row vectors $\mathbf{m}_{f_x}^\top, \mathbf{m}_{f_y}^\top \in \mathbf{M}$ be the x and y components of frame f's projection. Then the orthogonality of $\mathbf{m}_{f_x}^\top = \tilde{\mathbf{m}}_{f_x}^\top/\mathbf{G}$ and $\mathbf{m}_{f_y}^\top = \tilde{\mathbf{m}}_{f_y}^\top/\mathbf{G}$ gives the constraint $\forall_{\tilde{\mathbf{m}}_f^\top \in \tilde{\mathbf{M}}}\ \tilde{\mathbf{m}}_{f_x}^\top\mathbf{G}^{-1}\mathbf{G}^{-\top}\tilde{\mathbf{m}}_{f_x} - \tilde{\mathbf{m}}_{f_y}^\top\mathbf{G}^{-1}\mathbf{G}^{-\top}\tilde{\mathbf{m}}_{f_y} = \tilde{\mathbf{m}}_{f_x}^\top\mathbf{G}^{-1}\mathbf{G}^{-\top}\tilde{\mathbf{m}}_{f_y} = 0$. This system of constraints is linear in the six unknowns of symmetric $\mathbf{H} = \mathbf{G}^{-1}\mathbf{G}^{-\top}$, which can be obtained via standard least-squares methods from a system of linear constraints (with the added constraint $\tilde{\mathbf{m}}_{1_x}^\top\mathbf{G}^{-1}\mathbf{G}^{-\top}\tilde{\mathbf{m}}_{1_x} = c > 0$ to fix the scale of $\mathbf{G}$). Because $\mathbf{H}$ is symmetric, the constraints on it can be expressed very concisely: Define vech $\mathbf{H}$ to be the vector of the lower-triangular elements of $\mathbf{H}$, and vecs $\mathbf{H} \doteq$ vech$(\mathbf{H} + \mathbf{H}^\top - \mathbf{H} \odot \mathbf{I})$. Then $\mathbf{H}$ is the least-squares solution to the overconstrained system of linear equations $\forall_{f:\tilde{\mathbf{m}}_{f_x}^\top, \tilde{\mathbf{m}}_{f_y}^\top \in \tilde{\mathbf{M}}}$

$$(\text{vecs}(\tilde{\mathbf{m}}_{f_x}\tilde{\mathbf{m}}_{f_x}^\top - \tilde{\mathbf{m}}_{f_y}\tilde{\mathbf{m}}_{f_y}^\top))^\top \text{vech}\,\mathbf{H} = 0, \text{ (equal norm)} \tag{4}$$

$$(\text{vecs}(\tilde{\mathbf{m}}_{f_x}\tilde{\mathbf{m}}_{f_y}^\top))^\top \text{vech}\,\mathbf{H} = 0, \text{ (orthogonal)} \tag{5}$$

$$(\text{vecs}(\tilde{\mathbf{m}}_{1_x}\tilde{\mathbf{m}}_{1_x}^\top))^\top \text{vech}\,\mathbf{H} = c. \text{ (fixed scale)} \tag{6}$$

$\mathbf{G}^{-1}$ is estimated from the eigen-decomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \xleftarrow{\text{EIG}} \mathbf{H}$ as $\mathbf{G}^{-1} \leftarrow \mathbf{V}\sqrt{\mathbf{\Lambda}}$. This assumes that $\mathbf{H}$ is positive definite, which is not always the case, leading to nonpositive eigenvalues in $\mathbf{\Lambda}$ and a complex-valued or rank-deficient $\mathbf{G}^{-1}$. In this case we suggest approximating $\mathbf{G}$ from an SVD of $\mathbf{H}$, then turning to the fixpoint

$$\mathbf{G} \leftarrow \mathbf{G}([\Downarrow_f \tilde{\mathbf{M}}_f\mathbf{G}]\backslash[\Downarrow_f (\tilde{\mathbf{M}}_f\mathbf{G})^{\top\dagger}])^{1/2} \tag{7}$$

This solves for the transform that brings $\tilde{\mathbf{M}}_f\mathbf{G}$ closest to $(\tilde{\mathbf{M}}_f\mathbf{G})^{\top\dagger}$, with equality for proper rotations.

3.3. The nonrigid case

Bregler, Hertzmann & Biermann [3] recently proposed a direct extension of the above algorithm to the nonrigid case: For K morph bases one performs an SVD of $\mathbf{P} \ominus \mathbf{T}$ and retains the top $3K$ singular vectors on each side to obtain $\tilde{\mathbf{M}}_{DF\times 3K}\tilde{\mathbf{S}}_{3K\times N} \overset{\text{SVD}_{3K}}{\longleftarrow} \mathbf{P} \ominus \mathbf{T}$. The shape matrix $\tilde{\mathbf{S}}$ of right singular vectors contains K morph bases. Each set of D rows in the motion matrix $\tilde{\mathbf{M}}$ of left singular vectors is rearranged as if it were an outer product of rotation coefficients and deformation weights, then factored as such via a second round of rank-1 SVDs: $\forall_f \ (\text{vec}\,\tilde{\mathbf{R}}_f)\tilde{\mathbf{C}}_f \overset{\text{SVD}_1}{\longleftarrow} \text{vec}_{3D}\,\tilde{\mathbf{M}}_f$. The rotations and shape/deformation matrix are then affine corrected as in §3.2. This assumes that the 1st SVD leaves the singular vectors consistently signed and ordered by morph and dimension (e.g., $\tilde{\mathbf{S}} = [\mathbf{s}_{1y}, -\mathbf{s}_{1x}, \mathbf{s}_{1z}, \mathbf{s}_{2y}, -\mathbf{s}_{2x}, \mathbf{s}_{2z}, ..., \mathbf{s}_{Ky}, -\mathbf{s}_{Kx}, \mathbf{s}_{Kz}]^\top$ where $\mathbf{s}_{2x}$ is the x component of the second morph basis) whereas the SVD not only reorders but actually *mixes* these channels with an unknown affine transform $\mathbf{J}^{-1}_{3K\times 3K}$—one that maximizes concentration of variance in the top singular values. The singular vectors are also randomly signed. Fortuitously, in most human faces the first four channels of greatest variation are head height, width, depth, and vertical jaw motion ($\mathbf{s}_{1y}, \mathbf{s}_{1x}, \mathbf{s}_{1z}, \mathbf{s}_{2y}, ...$), so that shape and perhaps the first morph will be plausible, but after that the ordering of the channels is unpredictable, leading to mutual contamination of the morph and rotation estimates.

A simple example shows how the BHB factorization heuristic is vulnerable to less fortuitous datasets and SVDs: Imagine a child's toy with two beads that ride on horizontal rails. The toy has 3D shape and two independent modes of deformation. BHB factorization requires rank-9 data to determine shape and two modes of deformation, but the tracking data is only rank-5 (with a mix of channels approximated by the ordering $\mathbf{s}_{1x}, \mathbf{s}_{1y}, \mathbf{s}_{2x}, \mathbf{s}_{3x}, \mathbf{s}_{1z}$), which means that regardless of the amount of data, BHB factorization can only recover two morph bases (shape and a single deformation that combines the motions of both beads in way that may not be physically valid). The misordered singular vectors also lead to incorrect rotation estimates, which contaminate morph bases with torsions and, in the presence of noise, can create additional spurious morph bases (e.g., column 2, row 3 of figure 3). One can cyclically re-solve for each of $\hat{\mathbf{R}}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{S}}$ given the other two (solutions are given in [2]), but we found that this often converges to a mediocre local optimum.

3.4. The corrective transform problem

The crux of the problem is finding an optimal correction $\mathbf{J}^{-1}_{3K\times 3K}$ that transforms the result of the SVD into a properly structured motion matrix ($\hat{\mathbf{M}} \leftarrow \tilde{\mathbf{M}}\mathbf{J}^{-1}$). Recall from figure 2 that each $D \times 3$ block $\mathbf{M}_{f_k} \in \mathbf{M}$ is a scaled rotation whose rows $\mathbf{m}^\top_{f_{k,x}}$ and $\mathbf{m}^\top_{f_{k,y}}$ effect the x and y image projection of one morph basis; these rows have equal norm and are orthogonal. Moreover, they are orthogonal to $\mathbf{m}^\top_{f_{j,y}}$ and $\mathbf{m}^\top_{f_{j,x}}$ taken from any block to the left or right ($j \neq k$), because these blocks are all generated from the same rotation. The exact set of necessary and sufficient norm/orthogonality constraints that $\hat{\mathbf{M}}$ must satisfy are summarized by the quadratic equality $\forall_{\hat{\mathbf{M}}_f \in \hat{\mathbf{M}}}$,

$$(\underset{3}{\text{vec}}\,\hat{\mathbf{M}}_f^\top)^\top(\underset{3}{\text{vec}}\,\hat{\mathbf{M}}_f^\top) = \tfrac{1}{D}\mathbf{I}_D \otimes ((\underset{3D}{\text{vec}}\,\hat{\mathbf{M}}_f)^\top(\underset{3D}{\text{vec}}\,\hat{\mathbf{M}}_f)). \quad (8)$$

Since $\hat{\mathbf{M}} = \tilde{\mathbf{M}}\mathbf{J}^{-1}$, solution of eqn. (8) in the least-squares sense is equivalent to minimizing a system of polynomials that are quartic in the elements of $\mathbf{J}^{-1}$. In the rigid-body case, eqn. (8) is strictly quartic in $\mathbf{J}^{-1}$ and can be approached as a squared-squared-error problem via nested least-squares procedures. This is the strategy of §3.2. In the nonrigid case this strategy does not apply because eqn. (8) is both quartic and quadratic in $\mathbf{J}^{-1}$; the first least-squares procedure in §3.2—division—obliterates information about the quadratic terms that is needed by the second—eigen-decomposition. Direct solution is a very difficult problem so research has centered on finding numerically well-behaved heuristics. For example, the BHB factorization sets $\mathbf{J} \leftarrow \mathbf{I}_K \otimes \mathbf{G}$, a block-

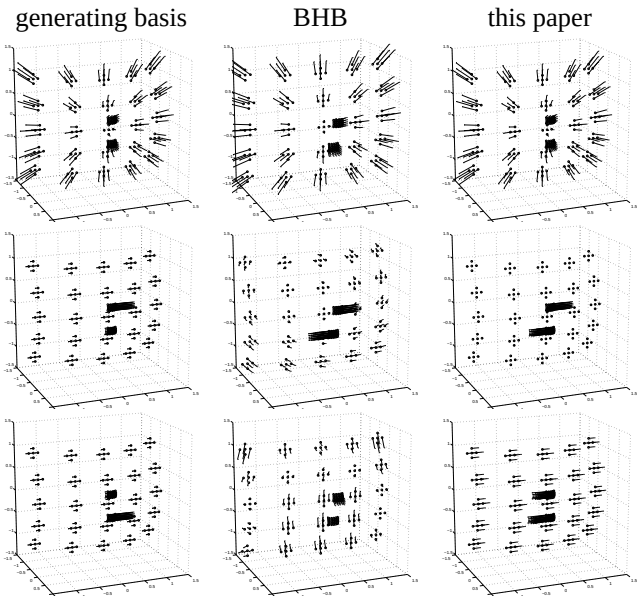


Figure 3: Reconstruction of a curved surface and two "beads" that move independently on horizontal tracks. Dots show average point locations; quivers show direction of motion for positive morph weights. COLUMN 1: The linear basis set used to generate test data: shape/scale; upper bead motion; lower bead motion. COLUMN 2: Shape and deformations recovered by BHB factorization [3] of 2D projections. One bead is misplaced in depth; there is no independent motion of the beads (except that the upper bead is allowed a spurious motion in depth); and all deformations have torsions that compensate for incorrect rotation estimates. COLUMN 3: A correct shape/deformations basis recovered from the same data by the method given below. (Adding or subtracting the deformations gives isolated motion of either bead.)

diagonal correction that assumes that the SVD correctly organizes all of the information about a morph basis in the appropriate column-triple in $\tilde{\mathbf{M}}$.

Numerical experiments with projections of 3D data whose principal components are known indicate that $\mathbf{J}$ is dense, particularly above the diagonal, meaning that the SVD mixes variation due to minor deformations into the shape and principal deformations. In fact, it is quite difficult to construct a dataset for which $\mathbf{J}$ has anything vaguely close to block-diagonal structure—even with vast amounts of noiseless synthetic data. Our experiments suggest that the scale of the deformations must drop off quadratically in order for the initial SVD to properly group their x, y, z components. Even then, it is unlikely that the components are consistently ordered within all groups.

In appendix §B we give one of a family of solutions that generalize the corrective transform (§3.2) to nonrigid motion. However, all such solutions are plagued by rank-deficiency problems because the number of unknowns grows quadratically while the rank of the constraints grows linearly: $\mathbf{J}^{-1}$ has $9K^2$ unique elements while there are $4K$ nonredundant constraints[1] per $\tilde{\mathbf{M}}_f \in \tilde{\mathbf{M}}$. Moreover, in casual video, the motions in most frames are highly redundant and contribute few new constraints. This sample-complexity problem is a property of image formation, consequently any correction algorithm based purely on the expected structure of the motion matrix will fail as the number of morph modes grows.

4. Flexible factorization

Our strategy is to bring in constraints from the shape/morph matrix $\mathbf{S}$: The deformations in $\hat{\mathbf{S}}$ should be as small as possible relative to the mean shape, so that the observed displacement of projected points away from the object-centric origin are explained mostly by the object's shape and residually by its deformations. Equivalently, whenever possible, point motions should be explained parsimoniously by rigid transforms (rotations and scale changes) rather than unparsimoniously by combinations of deformations. Otherwise all motion could be explained as deformations. Let $\hat{\mathbf{S}} = \tilde{\mathbf{S}}\mathbf{J}$ be the corrected shape/morph matrix and define $\mathbf{Z} \doteq \mathbf{I}_3 \otimes \mathrm{diag}[0, \mathbf{1}_{1\times K-1}]$. We want to minimize the Frobenius norm of $\mathbf{Z}\hat{\mathbf{S}}$, the part of the shape/morph matrix that contains deformations.

We now have two constraints—structure of the motion matrix and parsimony of the deformations. The problem is that the motion matrix gives constraints on $\mathbf{J}^{-1}$ via $\hat{\mathbf{M}} = \tilde{\mathbf{M}}\mathbf{J}^{-1}$, while the shape/morph matrix gives constraints on $\mathbf{J}$ via $\hat{\mathbf{S}} = \tilde{\mathbf{S}}\mathbf{J}$. To work around this algebraic inconvenience, we rewrite our motion constraint as $\check{\mathbf{M}}\mathbf{J} = \tilde{\mathbf{M}}$, where $\check{\mathbf{M}}$ is an initial estimate of the corrected motion matrix.

To make our initial estimate $\check{\mathbf{M}}$, one may use §B (or the BHB heuristic) and construct a properly structured motion matrix from the result. Both methods have weaknesses and we have found a third procedure which appears to be more robust for 2D data (for 3D data, §B eqn. (24) appears to be robust): First we flip signs of the left singular vectors in $\tilde{\mathbf{M}}$ to minimize the squared-error vis-a-vis the norm/orthogonality constraints of eqn. (8). Sign flipping leads to better rotation estimates and it can be done efficiently by caching intermediate results. Short-distance column-swaps can be evaluated in the same manner. We then affine-correct each column-triple in $\tilde{\mathbf{M}}$ as in §3.2 and 3D-rotate each column-triple to a common coordinate frame. We then stack all column-triples in $\tilde{\mathbf{M}}$ into $\tilde{\mathbf{M}}^{(2F,3)}$, compute a corrective transform $\mathbf{G}^{-1}$ as per §3.2, and apply it to all column-triples of $\tilde{\mathbf{M}}$. For each transform to $\tilde{\mathbf{M}}$ a compensatory inverse transform is applied to $\tilde{\mathbf{S}}$. We then factor each $\tilde{\mathbf{M}}_f \in \tilde{\mathbf{M}}$ into rotation and morph weights using an *orthonormal decomposition*[2] [2] that directly factors a matrix into a rotation and a vector. We then construct a properly structured motion matrix $\check{\mathbf{M}}$, plugging the initial estimates of $\mathbf{R}$ and $\mathbf{C}$ into eqn. (3). Unlike the BHB procedure, each column-triple has a unique correction and we have orthogonalized the pseudo-motion matrix *without* information-lossy factorization into $\mathbf{R}_f$ and $\mathbf{c}_f$. However, we have only estimated elements of $\mathbf{J}^{-1}$ in a band around the diagonal; the remaining far off-diagonal elements will be recovered in the next paragraph.

Combining the constraints from the motion and shape matrices, we obtain the objective function

$$\min_{\mathbf{J}} \mathrm{tr}((\check{\mathbf{M}}\mathbf{J} - \tilde{\mathbf{M}})^\top(\check{\mathbf{M}}\mathbf{J} - \tilde{\mathbf{M}})) + \mathrm{tr}(\tilde{\mathbf{S}}^\top\mathbf{J}^\top\mathbf{Z}\mathbf{J}\tilde{\mathbf{S}}). \quad (9)$$

This seeks the operator $\mathbf{J}$ that brings out the expected structure in $\tilde{\mathbf{M}}$ with the smallest possible deformations in $\mathbf{S}$. This error is minimized by the solution to the system of equations $\tilde{\mathbf{M}}\mathbf{J} = \check{\mathbf{M}}$ and $\mathbf{Z}\mathbf{J}\tilde{\mathbf{S}} = \mathbf{0}$. $\mathbf{J}$ is obtained from the sparse division

$$\hat{\mathbf{J}} \leftarrow \underset{3K}{\mathrm{vec}}\left(\begin{bmatrix}\mathbf{I}_{3K}\otimes\check{\mathbf{M}}\\ \tilde{\mathbf{S}}^\top\otimes\mathbf{Z}\end{bmatrix}\Big\backslash\begin{bmatrix}\mathrm{vec}\,\tilde{\mathbf{M}}\\ \mathbf{0}_{3KN\times 1}\end{bmatrix}\right) \quad (10)$$

from which we calculate $\hat{\mathbf{S}} \leftarrow \mathbf{J}\tilde{\mathbf{S}}$ and $\hat{\mathbf{M}} \leftarrow \tilde{\mathbf{M}}/\mathbf{J}$ or simply keep $\hat{\mathbf{R}}$ and re-estimate $\hat{\mathbf{C}}$. Since eqn. (10) uses information in both sides of the SVD, it is well constrained. In practice, we find that the upper triangle and several subdiagonals of $\mathbf{J}$ are usually dense, indicating that information about any one deformation is indeed spread over several columns of $\check{\mathbf{M}}$.

Eqn. (10) is a regularization that enables good factorizations from very small datasets. It could be used iteratively with refactorizations of $\hat{\mathbf{M}}$, though we do not.

5. Using image gradients

The above algorithm can be recast entirely in terms of image gradients, which are linearly related to motion in the

[1] One norm and one orthogonality constraint per block; two orthogonality constraints from from the first block to each block to its right.

[2] Calculated as $\mathbf{A} \leftarrow ((\mathrm{vec}_{3D}\,\tilde{\mathbf{M}}_f)\mathbf{1})^{(D)}$, $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \xleftarrow{\text{EIG}} (\mathbf{A}\mathbf{A}^\top)_{D\times D}$, $\hat{\mathbf{R}} \leftarrow \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}^\top\mathbf{A}$, $\hat{\mathbf{c}} \leftarrow ((\mathrm{vec}\,\hat{\mathbf{R}})^\top\mathbf{M})^\top/D$. See [2] for derivation.

setting of optical flow: Consider a small region R in image I_0 that shifts to a new location in image I_1. Assuming it views a constantly illuminated Lambertian surface, its optical flow $\mathbf{f}_{D\times 1}$ may be estimated (to first-order) from spatial image gradient $\nabla_{\mathbf{p}} \doteq \mathrm{d}I_0(\mathbf{p})/\mathrm{d}\mathbf{p}$ as $\hat{\mathbf{f}} \leftarrow \mathbf{X}\backslash\mathbf{y}$ where the spatial variation within frame I_0 is $\mathbf{X}_{D\times D} \doteq \int_R \nabla_{\mathbf{p}} \nabla_{\mathbf{p}}^\top \,\mathrm{d}\mathbf{p}$ and the temporal variation between I_0 and I_1 is $\mathbf{y}_{D\times 1} \doteq \int_R (I_0(\mathbf{p}) - I_1(\mathbf{p})) \cdot \nabla_{\mathbf{p}} \,\mathrm{d}\mathbf{p}$. Good estimates of $\mathbf{X}$ are usually available but $\mathbf{y}$ is sensitive to noise in the image intensities. Assuming this noise is gaussian distributed, $\mathbf{X}$ has special significance as the inverse covariance matrix of the flow estimate $\mathbf{f}$—its eigenvectors give the directions in which $\mathbf{f}$ is most and least certain.

We will represent N local flows to each of F images simultaneously in the stacked matrices $\mathbf{F}_{DN\times F}, \mathbf{Y}_{DN\times F}$ and diagonally stacked $\mathbf{X}_{DN\times DN}$. $\mathbf{X}$ describes spatial variation around landmarks in a reference frame; each column $\mathbf{Y}_f \in \mathbf{Y}$ describes temporal variation between the reference frame I_0 and target frame I_f. Without additional constraints, $\mathbf{Y} = \mathbf{X}\mathbf{F}$. The covariance of the uncertainty in $\mathbf{F}$ is $\mathbf{\Sigma}_{\mathbf{F}} \doteq \mathbf{X}^{-1}$; conversely $\mathbf{\Sigma}_{\mathbf{Y}} \doteq \mathbf{X}\mathbf{\Sigma}_{\mathbf{F}}\mathbf{X}^\top = \mathbf{X}$.

We will now show how all of the operations of the previous section can be applied to $\mathbf{X}$ and $\mathbf{Y}_f$. First we eigen-decompose $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \overset{\text{EIG}}{\longleftarrow} \mathbf{\Sigma}_{\mathbf{Y}} = \mathbf{X}$ and use $\mathbf{Q} \doteq \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top$ for certainty-warped operations on $\mathbf{Y}$. $\mathbf{Q}$ warps a problem having an elliptical (mahalanobis) error metric to one having a spherical (Frobenius) norm, so that minimal mahalonobis-error solutions can be obtained from least-squares procedures such as matrix division and SVD[3]. We use this to estimate pure translations:

$$\hat{\mathbf{T}}_{DF\times 1} \leftarrow \mathrm{vec}[((\mathbf{Q}\mathbf{X}(\mathbf{1}_{n\times 1} \otimes \mathbf{I}_D))^\dagger (\mathbf{Q}\mathbf{Y}))_{D\times F}] \quad (11)$$

This is a certainty-warped calculation of the mean displacements. (The pseudoinverse is quickly computed using QR-decomposition and inversion of the resulting upper-triangular $D \times D$ matrix.) We now remove translation and incorporate position into the temporal intensity variations, obtaining

$$\mathbf{Y}' \doteq \mathbf{Y} + \mathbf{X}(\mathbf{P}_0 \ominus \overline{\mathbf{P}}_0 \ominus \hat{\mathbf{T}}) \quad (12)$$

where $\mathbf{P}_0$ are the locations of reference texture patches in the reference frame and $\overline{\mathbf{P}}_0$ is their centroid. $\mathbf{Y}'$ is now a function of rotations and deformations only, satisfying

$$\tilde{\mathbf{P}} \doteq (\mathbf{X}\backslash\mathbf{Y}')^{(D)} = \mathbf{M}\mathbf{S}. \quad (13)$$

Appendix §A details how to factor the zero-meaned correspondence estimates $\tilde{\mathbf{P}}$ w.r.t. their uncertainty (covariance $\mathbf{\Sigma}_{\mathbf{X}\backslash\mathbf{Y}'} = \mathbf{\Sigma}_{\mathbf{F}} = \mathbf{X}^{-1}$) into $\tilde{\mathbf{M}}, \tilde{\mathbf{S}}$; appendix §A.1 shows how to do the same factoring *directly from intensity variations* $\mathbf{Y}'$ w.r.t. their uncertainty $\mathbf{\Sigma}_{\mathbf{Y}'} = \mathbf{X}\mathbf{\Sigma}_{\mathbf{F}}\mathbf{X}^\top = \mathbf{X}$. The flexible factorization of §4 applies directly to the results.

6. Nonrigid 3D subspace flow

The fact that nonrigid motion is a low-rank multilinear process has an unusually useful implication: *It is possible to simultaneously track a* 3D *nonrigid surface and acquire its* 3D *shape/morph basis simply by manipulating the rank of the flow calculations.* The rigid-body equivalent of this assertion was first noted by Irani [5], whose rank-reduced flow algorithm was based on the premise that flow and associated temporal image gradients from a reference frame to adjoining frames are bilinear products of two matrices whose low rank can be deduced from the camera and scene type. Our forward model similarly implies that rank-reduction of $\mathbf{P} \oplus \mathbf{P}_0 \ominus \overline{\mathbf{P}}_0 \ominus \mathbf{T}$ to rank $3K$ will force the motion data to be consistent with the subspace of plausible nonrigid 3D models. Moreover, since temporal intensity gradients are locally linearly in motion ($\mathbf{Y} = \mathbf{X}\mathbf{F} = \mathbf{X}(\mathbf{P} \ominus \mathbf{P}_0)$), uncertainty-informed rank-reduction of the temporal intensity variation matrix will similarly constrain the flow to lie in the same subspace. The key is to manipulate $\mathbf{Y}'$ (eqn. (12)) so that the rank constraints implied by eqn. (13) are applicable. This is accomplished by the intensity-based factorization in §A.1; we also give a more efficient alternate procedure:

We begin by computing $\mathbf{X}$ from image patches within a reference frame I_0 and $\mathbf{Y}'$ from comparisons of those patches to similarly located patches in all other frames. Because $\mathbf{M}\mathbf{S}$ has rank $3K$, eqn. (13) tells us that $\mathbf{Y}'$ has maximum rank $3DK$. We eigen-decompose $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \overset{\text{EIG}}{\longleftarrow} \mathbf{\Sigma}_{\mathbf{Y}'} = \mathbf{X}$ and use $\mathbf{Q} \doteq \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top$ in a certainty-warped thin SVD

$$\mathbf{U}\mathbf{\Sigma}\mathbf{W}^\top \overset{\text{SVD}_{3DK}}{\longleftarrow} \mathbf{Q}\mathbf{Y}'. \quad (14)$$

Since $\mathbf{Q}^\top\mathbf{Q} = \mathbf{X}^{-1}$, the product $\mathbf{Q}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{W}^\top \simeq \mathbf{X}^{-1}\mathbf{Y}' \equiv (\mathbf{M}\mathbf{S})^{(D)}$ is the uncertainty-informed reduction of the inter-frame correspondences to rank-$3DK$ (modulo translations). Rearranging the product to conform with $\mathbf{M}\mathbf{S}$ licenses the final rank-reduction to rank $3K$:

$$\mathbf{U}'\mathbf{\Sigma}'\mathbf{W}'^\top \overset{\text{SVD}_{3K}}{\longleftarrow} (\mathbf{Q}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{W}^\top)^{(D)}. \quad (15)$$

Finally, we restore translations to obtain point locations:

$$\hat{\mathbf{P}}_{N\times DF} = \mathbf{U}'\mathbf{\Sigma}'\mathbf{W}'^\top \oplus \mathbf{T} \oplus \overline{\mathbf{P}}_0. \quad (16)$$

New temporal image gradients $\mathbf{Y}_{\text{new}}$ are sampled w.r.t. these correspondences, and the process repeats until convergence. This is simlar in spirit to Irani's [5] rank-reduced flow but differs in that (A) it handles nonrigid scenes and objects; (B) it properly certainty-warps the intensity variations w.r.t. their own uncertainty *prior* to SVD; and (C) the rank constraints are *exact* because they are inherited directly from the forward model. The results of eqn. (15) are useful beyond rank-reduction: We use pseudo-motions $\tilde{\mathbf{M}} \leftarrow \mathbf{U}'\sqrt{\mathbf{\Sigma}'}$ and pseudo-shape $\tilde{\mathbf{S}} \leftarrow \sqrt{\mathbf{\Sigma}'}\mathbf{W}'^\top$ to "grow" the sequence by predicting correspondences in new frames via linear extrapolation of the rows at either end of $\tilde{\mathbf{M}}$.

The factorization constrains the search for correspondences; the search provides information for the factorization. As the process grows to cover the entire sequence, the space of possible nonrigid 3D models becomes increasingly constrained. For online tracking, we obtain extra efficiencies by using an incremental SVD that reduces computational complexity and automatically resolves temporary occlusions.

[3] These "covariance-weighted" methods have a long history as "directionally weighted least squares" in matrix algebra.

7. Examples

We began with 61 contiguous frames of 29.97Hz interlaced 320×240 video captured from a rented VHS video tape[4]. The scene rapidly cuts back and forth between a restaurant patron and a waitress; we modeled the patron's face, which averages 80 pixels in height. We chose roughly 90 points on his face in a reference frame, and ran the 3D flow algorithm with 4 morph bases to find correspondences and 3D structure in the remaining frames. Note that this is quite unconstrained video—there are no markers on the face, some of the points have almost no local texture, there are lighting changes, the camera parameters are unknown, and there is motion in the background. Some points are also occluded by head turns. To see whether the algorithm could handle discontinuous video, we added four more sequences totalling 87 frames from adjoining camera cuts. The 3D flow algorithm found correct correspondences across the camera cuts and in all the remaining frames. In most frames the head faces forward with very small rotations out of the fronto-parallel plane; in the last sequence he looks down at a menu. Despite the rather spare rotational depth cues, the recovered model, shown in figure 7, has good 3D shape. We used the model to render "3D video" in which the video plane is deformed according to the recovered depths, then viewed from an angle. Figure 4 shows 3 original frames and synthetic "side views."

We also took 490 frames from an old home video of a 3-year-old telling a story. Due the the child's smooth skin and blonde coloring, there is very little texture to support feature tracking and indeed, local feature trackers typically failed within 50 frames. The 3D flow algorithm of §6 was initialized with 100 points on the face found by an interest operator in a single frame, and successfully found correspondences across the entire sequence, concluding with a corrective transform to give the 3D model used to generate the images in figure 6. Figure 5 shows the recovered motion parameters. The original Irani subspace flow algorithm [5] does not successfully track this sequence, even when modified to use the same rank constraints as our version. The image correspondences found by our algorithm were fed into the original BHB algorithm, which failed to separate jaw motions from head rotations (jaw openings have a slight negative correlation with the pitch of the head around the model centroid), producing a model with an inverted jaw (figure 6, right).

8. Summary and prospects

We have presented a linear framework for recovering 3D shape, motion, and articulations of nonrigid 3D objects from video. Factoring morphable 3D models from 2D correspondences is a quartic optimization problem, for which we presented (§B) one of a family of formally "correct" solutions based on cascaded matrix decompositions that generalize the classic rigid-body structure-from-motion factorization. All algorithms based on the forward model's geometric invariants can be defeated by properties of the SVD that are at odds with the desired factorization, so we identified an additional "parsimony" constraint and used it to develop a correction to the SVD (§4). We then gave an improved and generalized method for factorization of correspondences *or intensity variations* with respect to uncertainty in the image sequence (§5&A). This led to a solution for morphable 3D models *directly from intensities* in which interframe correspondences are found in the course of computing the factorization (§6). Reseach now focuses on a refinement scheme for full perspective and more sophisticated models of texture flow.

[4] Thanks to Rahul Bhotika for this sequence.

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. SIGGRAPH99*, 1999.

[2] M. Brand. Flexible flow for 3D nonrigid tracking and shape recovery. In *Proc. CVPR*, 2001.

[3] C. Bregler, A. Hertzmann, and H. Biermann. Recovering nonrigid 3D shape from image streams. In *Proc. CVPR*, 2000.

[4] G. Golub and A. van Loan. *Matrix Computations*. Johns Hopkins U. Press, 1996.

[5] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *Proc. ICCV*, 1999.

[6] M. Irani and P. Anandan. Factorization with uncertainty. In *Proc. ECCV*, 2000.

[7] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. Wiley, 1999.

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

A. Factoring with uncertainty

Here we derive a method for factoring uncertain nonrigid tracking data. The rigid case was first treated by Irani & Anandan [6]. We correct some small errors and use a new solution method to generalize to nonrigid motion and varied uncertainty structures. To facilitate comparison with the original paper we use I&A's variable names and convert to their matrix organization:

The *D-interleave matrix* $\mathbf{E}^{[D]}_{N\times N}$ is a permutation matrix with $\mathrm{E}_{i,\lfloor(i-1)/D\rfloor+N((i-1) \bmod D)+1} = 1$. Postmultiplication with $\mathbf{E}$ rearranges a matrix with columns representing interleaved (e.g., $x_1y_1z_1x_2y_2z_2x_3y_3z_3\ldots$) data to a grouped form (e.g., $x_1x_2x_3y_1y_2y_3z_1z_2z_3\ldots$); postmultiplication with $\mathbf{E}^\top$ does the reverse. We use $\mathbf{E}$ to rearrange the block-diagonal inverse covariance matrix $\mathbf{X}$ to form a striped matrix $\mathbf{X}' \doteq \mathbf{E}^\top\mathbf{X}\mathbf{E}$ for the calculations below ($\mathbf{X} = \mathbf{E}\mathbf{X}'\mathbf{E}^\top$ recovers the block-diagonal form). We eigen-decompose $\mathbf{\Omega\Lambda\Omega}^\top \xleftarrow{\text{EIG}} \mathbf{X}'$ and compute a right-handed certainty warp $\mathbf{Q}' \doteq \mathbf{\Omega}\sqrt{\mathbf{\Lambda}}$, that maps the directionally weighted least-squares problem implied by $\mathbf{X}'$ onto an equivalent ordinary least-squares problem.

We split the tracking data into new matrices $\mathbf{U}_{F\times N}$, $\mathbf{V}_{F\times N}$, and (optional) $\mathbf{W}_{F\times N}$ containing horizontal, vertical, and (optional) depth ordinates for N points in F

frames. We desire a factorization into pseudo-shape matrix $\tilde{\mathbf{S}}_{3K\times N}$ and pseudo-motion matrix $\tilde{\mathbf{M}}_{DF\times 3K}$ satisfying $\tilde{\mathbf{M}}\tilde{\mathbf{S}} = [\mathbf{U},\mathbf{V},\mathbf{W}]^{(F,N)} = [\mathbf{U}^\top,\mathbf{V}^\top,\mathbf{W}^\top]^\top$, with any residual having minimal mahalanobis length w.r.t. the metric defined by $\mathbf{X}$. We rearrange the tracking data into a horizontally stacked matrix $[\mathbf{U},\mathbf{V},\mathbf{W}]$ in which each row describes a frame; this places all variables whose uncertainty is correlated in the same row so that the certainty warp can be applied. The identity $\mathbf{C}_{rp\times c} = \mathbf{A}_{rp\times q}\mathbf{B} \iff \mathbf{C}^{(r,c)} = \mathbf{A}^{(r,q)}(\mathbf{I}_p \otimes \mathbf{B})$ allows us to rewrite the target factorization as

$$[\mathbf{U},\mathbf{V},\mathbf{W}]\mathbf{Q}' = \tilde{\mathbf{M}}^{(F,3K)}(\mathbf{I}_D\otimes\tilde{\mathbf{S}})\mathbf{Q}'. \tag{17}$$

We begin with a thin singular value decomposition $\tilde{\mathbf{H}}_{F\times 3DK}\tilde{\mathbf{\Delta}}_{3DK\times 3DK}\tilde{\mathbf{G}}^\top_{3DK\times DN} \overset{\text{SVD}_{3DK}}{\longleftarrow} [\mathbf{U},\mathbf{V},\mathbf{W}]\mathbf{Q}'$ to suppress noise under a mahalanobis (elliptical) error metric specified by $\mathbf{X}'$. We must *unwarp* to remove the bias introduced by $\mathbf{Q}'$, using a smaller SVD: $\tilde{\tilde{\mathbf{H}}}\tilde{\tilde{\mathbf{\Delta}}}\tilde{\tilde{\mathbf{G}}}^\top \overset{\text{SVD}_{3DK}}{\longleftarrow} \tilde{\mathbf{\Delta}}\tilde{\mathbf{G}}^\top/\mathbf{Q}'$ to obtain $\mathbf{H} \leftarrow \tilde{\mathbf{H}}\tilde{\tilde{\mathbf{H}}}\tilde{\tilde{\mathbf{\Delta}}}^{1/2}$ and $\mathbf{G} \leftarrow \tilde{\tilde{\mathbf{\Delta}}}^{1/2}\tilde{\tilde{\mathbf{G}}}^\top$. (Without unwarping, effects of $\mathbf{Q}'$ will persist into the final result as shape distortions.) Now $\mathbf{HG}$ is the best (minimal mahalonobis-error w.r.t. $\mathbf{X}'$) rank-$3DK$ approximation of $[\mathbf{U},\mathbf{V},\mathbf{W}]$. For gaussian uncertainty this maximum likelihood estimate also has maximum *marginal* likelihood, which means that we have effectively integrated out the uncertainty in the temporal intensity gradients sampled from the images.

We must make $\mathbf{H}$ and $\mathbf{G}$ consistent with the target factorization (eqn. (17)) by finding an invertible transform $\mathbf{D}_{3DK\times 3DK}$ such that $\tilde{\mathbf{M}}^{(F,3K)} = \mathbf{HD}^{-1}$ and $\mathbf{DG} = (\mathbf{I}_D\otimes\tilde{\mathbf{S}})$. Using the above identity, we note that $[\mathbf{U},\mathbf{V},\mathbf{W}]^{(F,N)} \simeq (\mathbf{HG})^{(F,N)} = \tilde{\mathbf{M}}\tilde{\mathbf{S}} = (\mathbf{HD}^{-1})^{(F,3K)}\tilde{\mathbf{S}} = (\mathbf{I}_D\otimes\mathbf{H})(\mathbf{D}^{-1})^{(3DK,3K)}\tilde{\mathbf{S}}$, which implies that the desired transform $\mathbf{D}$ and shape $\tilde{\mathbf{S}}$ can be recovered directly via the rank-$3K$ decomposition

$$\begin{aligned}\widehat{\mathbf{D}^{-1}}^{(3DK,3K)}\hat{\tilde{\mathbf{S}}} &\overset{\text{SVD}_{3K}}{\longleftarrow} (\mathbf{I}_D\otimes\mathbf{H})\backslash(\mathbf{HG})^{(F,N)} \\ &= (\mathbf{I}_D\otimes\tilde{\tilde{\mathbf{\Delta}}}^{-1/2}\tilde{\tilde{\mathbf{H}}}^\top\tilde{\mathbf{H}}^\top)(\mathbf{HG})^{(F,N)}\end{aligned} \tag{18}$$

In contrast to [6], this correctly unwarps[5] the results of the first SVD, handles dense uncertainty covariances, and gives a fully constrained solution for $\widehat{\mathbf{D}^{-1}}$.

A.1. Factorization from intensity gradients

We can factor directly from intensity variations, which eqn. (12) relates to shape and rotation changes through matrix $\mathbf{Y}' = \mathbf{X}(\mathbf{MS})^{(D)}$. Equivalently, to use the notation of §A, $\mathbf{Y}'^\top\mathbf{E} = [\mathbf{U},\mathbf{V},\mathbf{W}]\mathbf{X}'$. Because the uncertainties in $\mathbf{Y}'^\top\mathbf{E}$ and $[\mathbf{U},\mathbf{V},\mathbf{W}]$ have covariances $\mathbf{X}'$ and $\mathbf{X}'^{-1}$ respectively, their certainty-warped forms are *equivalent* and *interchangeable*. This means that the factorization of §A can be applied directly to $\mathbf{Y}'$ simply by replacing the left hand side of eqn. (17) with $\mathbf{Y}'^\top\mathbf{EQ}'\mathbf{\Lambda}^{-1}$.

[5] If not unwarping, one can use the identity $(\mathbf{I}_d\otimes\mathbf{S}_{r\times n})\mathbf{A}_{dr\times q} = \mathbf{B}_{dr\times q} \iff \mathbf{S} \leftarrow (\underset{r}{\text{vec}}\,\mathbf{B}^{(r)})/(\underset{n}{\text{vec}}\,\mathbf{A}^{(n)})$ to extract a certainty-weighted estimate of shape from $\mathbf{DG} = (\mathbf{I}_D\otimes\tilde{\mathbf{S}})\mathbf{Q}'$:

$$\hat{\tilde{\mathbf{S}}} \leftarrow (\underset{3K}{\text{vec}}(\hat{\mathbf{DG}})^{(3K)})/(\underset{N}{\text{vec}}\,\mathbf{Q}'^{(N)}) \tag{19}$$

For block-diagonal $\mathbf{X}$ this gives independent equations for each point.

B. Nonrigid corrective transform

Here we generalize the correction of §3.2 to estimate a correction matrix $\mathbf{J} \doteq \mathbf{M}\backslash\tilde{\mathbf{M}}$. We break $\mathbf{M}$ into K column-triples, each being a stack of rotation matrices scaled by morph weights. Let $\mathbf{m}^\top_{f_{k,x}}, \mathbf{m}^\top_{f_{k,y}} \in \mathbf{M}_f \in \mathbf{M}$ be the rows in column-triple k giving the x and y projections in frame f. As in §3.2, these vectors should have equal norm and be orthogonal. Morever, their projections onto vectors from other column-triples should also have equal norm (because all column-triples have the same rotations): $\forall_{f,k,j}$

$$[\mathbf{m}_{f_{k,x}}\mathbf{m}^\top_{f_{j,x}} = \mathbf{m}_{f_{k,y}}\mathbf{m}^\top_{f_{j,y}}] \text{ and } [\mathbf{m}_{f_{k,x}}\mathbf{m}^\top_{f_{j,y}} = 0]. \tag{20}$$

Since $\hat{\mathbf{m}}^\top_{f_{k,x}} = \tilde{\mathbf{m}}^\top_{f_x}(\mathbf{J}^{-1})_{\text{cols}(3k-2:3k)}$, for each value of k and j this yields a separate linear system like §3.2 eqns. (4–5) giving constraints on a matrix $\mathbf{H}_{k,j}$ (with vecs and vech replaced with vec for $k \neq j$). Each $\mathbf{H}_{k,j}$ is the outer product of two column-triples in $(\mathbf{J}^{-1})$, e.g.,

$$\mathbf{H}_{k,j} = (\mathbf{J}^{-1})_{\text{cols}(3k-2:3k)}(\mathbf{J}^{-1})^\top_{\text{cols}(3j-2:3j)} \text{ and} \tag{21}$$

$$\mathbf{H} \doteq [\Downarrow_k^K [\Rightarrow_j^K \mathbf{H}_{k,j}]] = (\mathbf{J}^{-1})^{(3K,3)}(\mathbf{J}^{-1})^{(3K,3)\top} \tag{22}$$

is symmetric and should have rank 3. Let $\mathbf{V\Lambda V}^\top \overset{\text{EIG}_3}{\longleftarrow} \mathbf{H}$ be a truncated decomposition of $\mathbf{H}$ using its three largest eigenvalues and their associated eigenvectors. Then the desired correction is $(\mathbf{J}^{-1}) = (\mathbf{V}\sqrt{\mathbf{\Lambda}})^{(3K,3)}$.

Although formally correct, this procedure is of limited use because without additional constraints on the structure of $\mathbf{J}$, the constraints on all $\mathbf{H}_{k,j}$ are highly redundant, with insufficient constraints to determine all elements in $\mathbf{H}$. In practice $\mathbf{H}_{1,1}$ contains enough constraints to support an estimate of the first three columns of $\mathbf{J}^{-1}$, from which we can calculate the first column-triple $\hat{\mathbf{M}}$ and with it a good estimate of all rotations $\hat{\mathbf{R}}$ (provided that $\mathbf{H}_{1,1}$ has exactly three strongly dominant eigenvalues). If working with 3D correspondences, such as motion capture or MRI tracking, the equality

$$\tilde{\mathbf{M}} = \mathbf{MJ} = [\searrow_f\hat{\mathbf{R}}_f](\mathbf{C}^\top\otimes\mathbf{I}_3)\mathbf{J} = [\searrow_f\hat{\mathbf{R}}_f](\mathbf{J}^{(3)}\mathbf{C})^{(3)} \tag{23}$$

leads to a direct solution for all remaining unknowns:

$$\hat{\mathbf{J}}^{(3)}\hat{\mathbf{C}} \overset{\text{SVD}_k}{\longleftarrow} ([\searrow_f\hat{\mathbf{R}}_f]^\top\tilde{\mathbf{M}})^{(3)}. \tag{24}$$

With 2D data, one can project $\tilde{\mathbf{M}}$ into the space orthogonal to $\mathbf{J}^{-1}_{\text{cols}(1:3)}$ and solve for $\mathbf{J}^{-1}_{\text{cols}(4:6)}$ that will produce a second column-triple of $\hat{\mathbf{M}}$ that is consistent with the rotations. In formulæ: We project $\tilde{\mathbf{M}}' \leftarrow \tilde{\mathbf{M}}(\mathbf{I} - \mathbf{J}^{-1}_{\text{cols}(1:3)}(\mathbf{J}^{-1}_{\text{cols}(1:3)})^\dagger)$ and solve the linear system $\forall_f \quad \tilde{\mathbf{M}}'_{f_x}\mathbf{J}'\hat{\mathbf{R}}^\top_{f_x} - \tilde{\mathbf{M}}'_{f_y}\mathbf{J}'\hat{\mathbf{R}}^\top_{f_y} = \tilde{\mathbf{M}}'_{f_x}\mathbf{J}'\hat{\mathbf{R}}^\top_{f_y} = \tilde{\mathbf{M}}'_{f_y}\mathbf{J}'\hat{\mathbf{R}}^\top_{f_x} = 0$ for $\mathbf{J}'$ using the identity $\mathbf{ABC} = (\mathbf{C}^\top\otimes\mathbf{B})\,\text{vec}\,\mathbf{A}$ to obtain $\mathbf{J}^{-1}_{\text{cols}(4:6)} \leftarrow (\mathbf{I} - \mathbf{J}^{-1}_{\text{cols}(1:3)}(\mathbf{J}^{-1}_{\text{cols}(1:3)})^\dagger)\mathbf{J}'$. We then recursively solve for the remaining column-triples. Again, the quality of the result depends on the eigenvalue structure of $\mathbf{H}_{1,1}$. We are now studying how this relates to the quantity and quality of data.

Figure 6: Original frame and three synthetic frames rotating the face, closing the mouth, and pursing the lips. At right is the base shape obtained by feeding the correspondences into the BHB factorization, which inverts the jaw. The graph shows that the flexible factorization estimates morph bases that more effectively explain the data.

Figure 4: Cropped video frames and synthetic profile views showing 3D recovered for the front half of the head. The rendering is not anti-aliased, and inherits compression and interlacing artifacts visible in the original low-res frames.

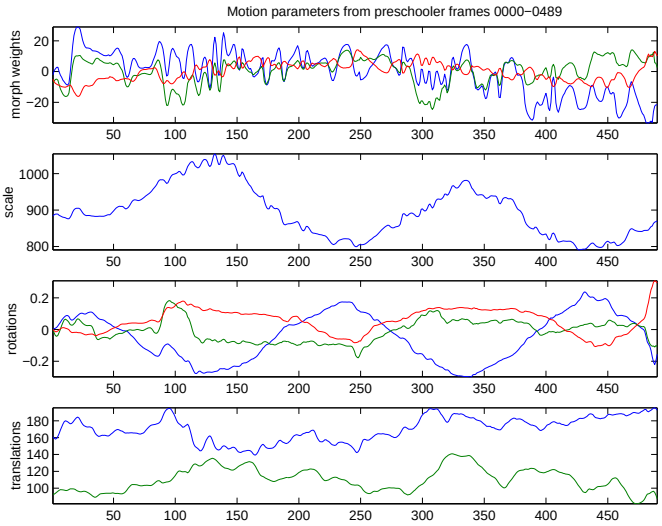


Figure 5: Morph, rotation, scale, and translation parameters recovered from the preschooler sequence. The high frequency fluctuations record mouth motions while talking.

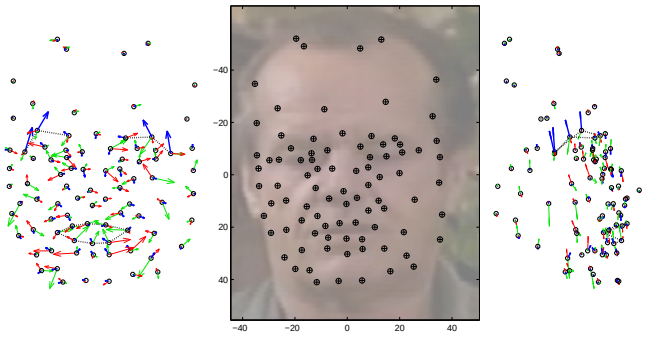


Figure 7: Front and left profile views of a $K = 4$ model recovered from 148 frames via 3D flow (see figure 4). First deformation (thick blue arrows) raises eyebrows and tightens mouth; second deformation (green arrows) opens and closes mouth; third deformation (thin red arrows) widens and narrows mouth. Dotted lines outline the mouth and eyebrows.