Factorization for Non-Rigid and Articulated Structure using Metric Projections

Marco Paladini
Queen Mary, University of London
paladini@dcs.qmul.ac.uk

Alessio Del Bue
ISR - Instituto Superior Técnico
adb@isr.ist.utl.pt

Marko Stošić
ISR - Instituto Superior Técnico
mstosic@isr.ist.utl.pt

Marija Dodig
CELC, Universidade de Lisboa
dodig@cii.fc.ul.pt

João Xavier
ISR - Instituto Superior Técnico
jxavier@isr.ist.utl.pt

Lourdes Agapito
Queen Mary, University of London
lourdes@dcs.qmul.ac.uk

Abstract

This paper describes a new algorithm for recovering the 3D shape and motion of deformable and articulated objects purely from uncalibrated 2D image measurements using an iterative factorization approach. Most solutions to non-rigid and articulated structure from motion require metric constraints to be enforced on the motion matrix to solve for the transformation that upgrades the solution to metric space. While in the case of rigid structure the metric upgrade step is simple since the motion constraints are linear, deformability in the shape introduces non-linearities. In this paper we propose an alternating least-squares approach associated with a globally optimal projection step onto the manifold of metric constraints. An important advantage of this new algorithm is its ability to handle missing data which becomes crucial when dealing with real video sequences with self-occlusions. We show successful results of our algorithms on synthetic and real sequences of both deformable and articulated data.

1. Introduction and Previous Work

Structure from motion (SfM) can be defined as the problem of combined inference of the motion of a camera and the 3D geometry of the scene it views solely from a sequence of images. The fundamental assumption which has allowed robust solutions to be achieved is that of scene rigidity. This assumption was recently relaxed to extend structure from motion algorithms to the case of deformable objects. Bregler *et al*.'s key insight [4] was to use a low-rank shape model to represent the deforming shape as a linear combination of k basis shapes which encode the main modes of deformation. Based on this model, they proposed a non-rigid factorization method for an affine camera model that exploited the rank constraint on the measurement matrix and enforced orthonormality constraints on camera rotations to recover the motion and the non-rigid 3D shape.

Although the low-rank shape model has proved a successful representation, the non-rigid structure from motion problem is inherently under-constrained. Recent approaches have focused on overcoming the problems caused by inherent ambiguities and degeneracies by proposing the use of generic priors or different optimization schemes. Aanaes *et al*. [1] impose the prior knowledge that the reconstructed shape does not vary much from frame to frame while Del Bue *et al*. [5] impose the constraint that some of the points on the object are rigid. Both approaches use bundle adjustment to refine all the parameters of the model together. Bartoli *et al*. [2] on the other hand, use a coarse to fine shape model where new deformation modes are added iteratively to capture as much of the variance left unexplained by previous modes as possible. Torresani *et al*. [12] also argue that simple linear subspace shape models are extremely sensitive to noise and missing data so priors should be used to constrain the shape. They propose to place a Gaussian prior distribution on the deformation weights and then generalise the model to represent linear dynamics in the deformations.

One advantage of the linear subspace model is that it has allowed closed form solutions to be proposed, for the cases of both affine [14] and perspective [15, 8] viewing conditions. However, they are known to be very sensitive to noise [3, 12] and to the selection of the basis constraints. Moreover, none of the closed form solutions proposed so far can deal with missing data.

Articulated motion has also been recently formulated using a structure from motion approach [13, 16] modelling the articulated motion space as a set of intersecting motion subspaces — the intersection of two motion subspaces implies the existence of a link between the parts. Articulation constraints can then be imposed during factorization to recover the location of joints and axes. Tresadern and Reid [13] go further and compute the metric upgrade, but only recover a linear approximation of the correcting transformation. Both approaches [13, 16] require full data and therefore cannot deal with missing tracks.

1.1. Contributions

In this paper we present a new unified approach to perform the metric upgrade in the cases of articulated and deformable structure viewed by an orthographic camera. In the non-rigid case our approach is most closely related to Torresani *et al.*'s alternating least-squares solution [12]. While they do enforce the exact metric constraints through an exponential map parameterization of the rotation matrices, the update of the camera matrix is only an approximation — the camera matrix cannot be updated in closed form and instead they perform a single Gauss-Newton step.

Similarly to Torresani *et al.* we also propose an iterative alternating scheme to solve the non-rigid structure from motion problem. However, in contrast to their approach, our metric upgrade step solves an unconstrained least-squares problem and optimally projects the solution onto the nonlinear motion manifold. The notion of *motion manifolds* has been recently introduced in the case of rigid shapes by Marques and Costeira [9]. Our work extends and generalises it to the case of deformable and articulated shapes. In particular, we impose that the rotation matrices lie on the $\mathtt{V}_{2,3}$ Stiefel manifold[1]. This constraint results in a non-convex problem which can then be solved by semidefinite tight relaxation in the case of deformable shape. In the articulated case, we efficiently compute the joints given the non-linear constraints on the motion of the two bodies. The result is an algorithm where the recovered motion matrices have the exact orthogonality constraints imposed. One of the main advantages of our approach is that it can be extended naturally to deal with missing data in a similar way to [9].

2. Factorization for Structure from Motion

Consider the set of 2D image trajectories obtained when the points lying on the surface of a 3D object are viewed by a moving camera. Defining the non-homogeneous coordinates of a point j in frame i as the vector $\mathbf{w}_{ij} = (u_{ij} \;\; v_{ij})^T$ we may write the measurement matrix $\mathtt{W}$ that gathers the coordinates of all the points in all the views as:

$$\mathtt{W} = \begin{bmatrix} \mathbf{w}_{11} & \dots & \mathbf{w}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{f1} & \dots & \mathbf{w}_{fp} \end{bmatrix} = \begin{bmatrix} \mathtt{W}_1 \\ \vdots \\ \mathtt{W}_f \end{bmatrix} \tag{1}$$

where f is the number of frames and p the number of points.

The measurement matrix can be factorized into the product of two low-rank matrices as $\mathtt{W} = \mathtt{M}_{2f\times r} \;\; \mathtt{S}_{r\times p}$, where $\mathtt{M}$ and $\mathtt{S}$ correspond to the motion and shape subspaces respectively. As a result, the rank of $\mathtt{W}$ is constrained to be $\mathrm{rank}\{\mathtt{W}\} \leq r$ where $r \ll \min\{2f, p\}$. The rank of these subspaces is dictated by the properties of the camera projection and the nature of the shape of the object being observed (rigid, deformable, articulated, *etc*.). This rank constraint forms the basis of the factorization method for the estimation of 3D structure and motion.

Matrices $\mathtt{M}$ and $\mathtt{S}$ can be expressed as $\mathtt{M} = \left[\mathtt{M}_1^T \cdots \mathtt{M}_f^T\right]^T$ and $\mathtt{S} = [\mathbf{S}_1 \cdots \mathbf{S}_p]$ where $\mathtt{M}_i$ is the $2 \times r$ camera matrix that projects the 3D shape onto the image frame i and $\mathbf{S}_j$ encodes the 3D coordinates of point j.

2.1. Rigid Shape

In the case of a rigid object viewed by an orthographic camera, if we assume the measurements in $\mathtt{W}$ are registered to the image centroid, the camera motion matrices $\mathtt{M}_i$ and the 3D points $\mathbf{S}_j$ can be expressed as: $\mathtt{M}_i = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \end{bmatrix} = \mathtt{R}_i$ and $\mathbf{S}_j = \begin{bmatrix} X_j & Y_j & Z_j \end{bmatrix}^T$ where $\mathtt{R}_i$ is a 2×3 matrix that lies on the Stiefel manifold since it contains the first two rows of a rotation matrix (i.e. $\mathtt{R}_i\mathtt{R}_i^T = \mathtt{I}_{2\times 2}$) and $\mathbf{S}_j$ is a 3-vector containing the metric coordinates of the 3D point. Therefore the rank of the measurement matrix is $r \leq 3$. The rigid *motion manifold* corresponds to the manifold of matrices with pairwise orthogonal rows (i.e. the Stiefel manifold $\mathtt{V}_{2,3}$).

2.2. Deformable Shape Model

In the case of deformable objects the observed 3D points change as a function of time. In this paper we use the low-rank shape model defined by Bregler *et al.* [4] in which the 3D points deform as a linear combination of a fixed set of k rigid shape bases according to time varying coefficients. In this way, $\mathtt{S}_i = \sum_{d=1}^{k} l_{id}\mathtt{B}_d$ where the matrix $\mathtt{S}_i = [\mathbf{S}_{i1}, \cdots \mathbf{S}_{ip}]$ is the 3D shape of the object at frame i, the $3 \times p$ matrices $\mathtt{B}_d$ are the shape bases and l_{id} are the coefficient weights. If we assume an orthographic projection model the coordinates of the 2D image points observed at each frame i are then given by:

$$\mathtt{W}_i = \mathtt{R}_i \left(\sum_{d=1}^{k} l_{id}\mathtt{B}_d \right) + \mathtt{T}_i \tag{2}$$

where $\mathtt{R}_i$ is a 2×3 Stiefel matrix and the $2 \times p$ matrix $\mathtt{T}_i$ aligns the image coordinates to the image centroid. When the image coordinates are registered to the centroid of the object and we consider all the frames in the sequence, we may write the measurement matrix as:

$$\mathtt{W} = \begin{bmatrix} l_{11}\mathtt{R}_1 & \dots & l_{1k}\mathtt{R}_1 \\ \vdots & \ddots & \vdots \\ l_{f1}\mathtt{R}_f & \dots & l_{fk}\mathtt{R}_f \end{bmatrix} \begin{bmatrix} \mathtt{B}_1 \\ \vdots \\ \mathtt{B}_k \end{bmatrix} = \begin{bmatrix} \mathtt{M}_1 \\ \vdots \\ \mathtt{M}_f \end{bmatrix} \begin{bmatrix} \mathtt{B}_1 \\ \vdots \\ \mathtt{B}_k \end{bmatrix} = \mathtt{MS} \tag{3}$$

Since $\mathtt{M}$ is a $2f \times 3k$ matrix and $\mathtt{S}$ is a $3k \times p$ matrix in the case of deformable structure the rank of $\mathtt{W}$ is constrained

[1] The Stiefel manifold $\mathtt{V}_{k,m}$ may be viewed as the collection of all $m \times k$ matrices whose columns form an orthonormal set. More precisely, the (real) Stiefel manifold $\mathtt{V}_{k,m}$ is the collection of all ordered sets of k orthonormal vectors in Euclidean space $\mathbb{R}^m$.

to be at most $3k$. The motion matrices now have the form $\mathtt{M}_i = [\mathtt{M}_{i1} \ldots \mathtt{M}_{ik}] = [l_{i1}\mathtt{R}_i \ldots l_{ik}\mathtt{R}_i]$. Therefore, in the deformable *motion manifold* the motion matrices have a distinct repetitive structure and every 2×3 $\mathtt{M}_{ik}$ sub-block is a Stiefel matrix multiplied by a scalar.

2.3. Articulated Shape Model

In the case of articulated structure, the relative motions of the segments that form an articulated body are dependent and this results in a drop in the dimensionality of the measurement matrix $\mathtt{W} = \left[\begin{array}{c|c} \mathtt{W}^{(1)} & \mathtt{W}^{(2)} \end{array}\right]$ that contains the 2D image points of the two segments. In the case of a *universal joint* the two shapes share a common translation (i.e. the distance between the centers of mass of the shapes is constant) while in the case of a *hinge joint* the shapes also share a common rotation axis [13, 16]. Naturally, this approach requires that an initial segmentation stage has taken place to assign the trajectories in $\mathtt{W}$ to the respective shapes for which Yan and Pollefeys [16] have recently provided a solution. In the case of the *hinge joint* the motion matrices $\mathtt{M}_i$ that lie on the articulated *motion manifold* can be written as:

$$\mathtt{M}_i = \left[\begin{array}{ccc} \mathbf{u}_i & \mathtt{A}_i & \mathtt{B}_i \end{array}\right] \tag{4}$$

where $\mathbf{u}$ is the common rotation axis for both objects, $\mathtt{A}_i$ and $\mathtt{B}_i$ are 2×2 matrices such that $\left[\begin{array}{c|c} \mathbf{u}_i & \mathtt{A}_i \end{array}\right]$ and $\left[\begin{array}{c|c} \mathbf{u}_i & \mathtt{B}_i \end{array}\right]$ are the 2×3 Stiefel matrices associated with the first and second shape respectively. The metric constraints in the case of a hinge can therefore be expressed as:

$$\begin{aligned} [\mathbf{u}_i \; \mathtt{A}_i] \left[\begin{array}{c} \mathbf{u}_i^T \\ \mathtt{A}_i^T \end{array}\right] &= \mathtt{I}_{2\times2} \\ [\mathbf{u}_i \; \mathtt{B}_i] \left[\begin{array}{c} \mathbf{u}_i^T \\ \mathtt{B}_i^T \end{array}\right] &= \mathtt{I}_{2\times2} \end{aligned} \tag{5}$$

where, without loss of generality, we have implicitly assumed that the axis of rotation is aligned with the x-axis of the first object. Thus we can write $\mathtt{S}$ as:

$$\mathtt{S} = \left[\begin{array}{cccccc} x_1^{(1)} & \cdots & x_{p_1}^{(1)} & x_1^{(2)} & \cdots & x_{p_2}^{(2)} \\ y_1^{(1)} & \cdots & y_{p_1}^{(1)} & 0 & \cdots & 0 \\ z_1^{(1)} & \cdots & z_{p_1}^{(1)} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & y_1^{(2)} & \cdots & y_{p_2}^{(2)} \\ 0 & \cdots & 0 & z_1^{(2)} & \cdots & z_{p_2}^{(2)} \end{array}\right] \tag{6}$$

where now $\mathtt{S}$ is a $5 \times p$ matrix and $p = p_1 + p_2$ (we assume the shapes have been registered to the respective object centroids). Therefore, in the case of a hinge joint the rank of the measurement matrix is at most 5.

3. Metric Upgrade

The classic approach in factorization is to exploit the rank constraint to factorize the measurement matrix into a motion matrix $\tilde{\mathtt{M}}$ and a shape matrix $\tilde{\mathtt{S}}$ by truncating the SVD of $\mathtt{W}$ to the rank r specific to the problem. However, this factorization is not unique since any invertible $r \times r$ matrix $\mathtt{Q}$ can be inserted, leading to the alternative factorization: $\mathtt{W} = (\tilde{\mathtt{M}}\mathtt{Q})(\mathtt{Q}^{-1}\tilde{\mathtt{S}})$. The problem is to find the transformation matrix $\mathtt{Q}$ that removes the affine ambiguity, upgrading the reconstruction to metric and constraining the motion matrices to lie on the appropriate *motion manifold*.

While in the rigid case the matrix $\mathtt{Q}$ can be explicitly computed linearly by imposing orthogonality constraints on the rows of the motion matrix [11], in the non-rigid and articulated cases the metric constraints on the motion matrices are non-linear. Although some closed-form solutions have been recently proposed [15, 14, 8] these algorithms perform badly in the presence of noise and cannot cope with missing data. Iterative solutions provide a viable alternative in the presence of noise and missing data. In this paper we solve the factorization of $\mathtt{W}$ as an alternating least-squares problem where at each step t the motion $\mathtt{M}^{(t)}$ and shape $\mathtt{S}^{(t)}$ matrices are optimized separately keeping the other one fixed as shown in Algorithm 1.

Algorithm 1 Iterative metric upgrade via alternation for deformable and articulated shape

```
Require: An initial estimate $\mathtt{M}^{(0)}$.
Ensure: A factorization of $\mathtt{W}$ that satisfies the given metric constraints.
1: Project each frame of $\mathtt{M}^{(t)}$ onto the motion manifold of the motion matrices (See Section 3.1 for the deformable case and Section 3.2 for the articulated case).
2: Estimate $\mathtt{S}^{(t)}$ from the projected $\mathtt{M}^{(t)}$ as: $\mathtt{S}^{(t)} = \mathtt{M}^{(t)\dagger}\mathtt{W}$.
3: Estimate $\mathtt{M}^{(t+1)}$ such that: $\mathtt{M}^{(t+1)} = \mathtt{W}\mathtt{S}^{(t)\dagger}$.
4: Repeat until convergence.
```

Crucially, Step 1 of the algorithm computes the projection of the affine motion components onto the *motion manifold* in which the exact metric constraints are satisfied. Steps 2 and 3 alternate the estimation of $\mathtt{M}^{(t)}$and $\mathtt{S}^{(t)}$ assuming the other one known.

Previous approaches have also used iterative methods to perform the metric upgrade in the case of non-rigid structure including the alternating least-squares method described in [12]. However, even though Torresani *et al.*'s method imposes exact metric constraints on the camera matrices, the update of the camera matrix is only an approximation and is not optimal. While other papers have chosen to use priors on the shape to constrain the solution and obtain the metric upgrade [2, 12], in this paper we provide a metric upgrade step that solves an unconstrained least-squares problem and optimally projects the solution onto the nonlinear motion manifold. In the case of articulated structure, we solve globally for both the motion components related to the bodies and the joint axis. We now describe how these projections are computed.

3.1. Metric Projection: Deformable Case

The projection is carried out on each $2 \times 3k$ sub-matrix $\mathtt{M}_i$ as defined in Section 2 and it corresponds to solving the following minimization problem:

$$\min_{\mathtt{R}_i, l_{i1}\dots l_{ik}} \left\|\mathtt{M}_i - [l_{i1}\mathtt{R}_i|\dots|l_{ik}\mathtt{R}_i]\right\|_F^2 \tag{7}$$

with the added constraint that $\mathtt{R}_i$ be a 2×3 Stiefel matrix. This is equivalent to minimizing separately all the 2×3 blocks of $\mathtt{M}_i$ giving:

$$\min_{\mathtt{R}_i} \sum_{d=1}^{k} \min_{l_{i1}\dots l_{ik}} \left\|\mathtt{M}_{id} - l_{id}\mathtt{R}_i\right\|_F^2 \tag{8}$$

which is equivalent to:

$$\min_{\mathtt{R}_i, l_{i1}\dots l_{ik}} \sum_{d=1}^{k} \left\|\mathtt{M}_{id}\right\|^2 + l_{id}^2 \left\|\mathtt{R}_i\right\|^2 - 2 l_{id} \operatorname{Tr}[\mathtt{M}_{id}^T \mathtt{R}_i]. \tag{9}$$

We can then decouple the problem by minimizing first for l_{id} given $\mathtt{R}_i$, that is, solving for the zeros of the derivative of eq. (8). The configuration weight l_{id} is minimized at:

$$l_{id} = \frac{\operatorname{Tr}[\mathtt{M}_{id}^T \mathtt{R}_i]}{\left\|\mathtt{R}_i\right\|^2} = \frac{1}{2} \operatorname{Tr}[\mathtt{M}_{id}^T \mathtt{R}_i] \tag{10}$$

putting this value back in eq. (8) and following with the simplification, the minimization can be written as:

$$\min_{\mathtt{R}_i} \quad \mathbf{r}_i^T \left[-\sum_{d=1}^{k} \mathbf{m}_{id}\mathbf{m}_{id}^T \right] \mathbf{r}_i \tag{11}$$

where $\mathbf{r}_i = vec(\mathtt{R}_i^T)$ with $\mathtt{R}_i\mathtt{R}_i^T = \mathtt{I}_{2\times 2}$ and $\mathbf{m}_{id} = vec(\mathtt{M}_{id}^T)$. This quadratic minimization problem presents non-convex constraints given by $\mathtt{R}_i$. Appendix A shows that it is possible to obtain a tight convex relaxation which can be efficiently solved using SeDuMi [10]. Further details can also be found in [7]. The computed Stiefel matrix $\mathtt{R}_i$ is then used to recover the weights l_{id}, obtaining a full non-rigid motion matrix that satisfies the metric constraints. This allows us to solve iteratively for the motion and shape as described in Algorithm 1.

Initialization. Algorithm 1 requires an initial estimate of the motion matrix $\mathtt{M}_i$ at each frame. This in turn requires initial estimates for the rotation matrices $\bar{\mathtt{R}}_i$ and the configuration weights $\bar{l}_{id}$. The rigid motion $\bar{\mathtt{R}}_i$ and the first basis shape $\bar{\mathtt{S}}_1$ are initialized from a rank 3 rigid factorization of the measurement matrix. The second component of the shape bases is estimated from the residual $\mathtt{W}_r = \mathtt{W} - \bar{\mathtt{M}}\bar{\mathtt{S}}_1$. A new rank 3 factorization is performed on $\mathtt{W}_r$ and the new configuration weights l_{i2} can be estimated solving for $l_{i2}\bar{\mathtt{R}}_i = \mathtt{M}_{i2}$ keeping the rotations fixed. This process is repeated to obtain all k deformation modes.

3.2. Metric Projection: Articulated Case

Projection onto the *motion manifold* of the universal joint can be simply solved by performing two separate rigid factorizations for each of the parts of the articulated object followed by estimation of the joint location as presented in [13]. The hinge joint is far more interesting given the non-linear relations between the motion subspaces. Here the problem is to find the closest matrix that satisfies the metric constraints given a rotation axis between two objects. Following eq. (4) the projection problem for the hinge *motion manifold* can be written as the following minimization:

$$\min_{\mathbf{u},\mathtt{A},\mathtt{B}} J(\mathbf{u},\mathtt{A},\mathtt{B}) = \left\|\mathbf{u}-\mathbf{x}\right\|^2 + \left\|\mathtt{A}-\mathtt{Y}\right\|^2 + \left\|\mathtt{B}-\mathtt{Z}\right\|^2, \tag{12}$$

subject to the constraints defined in eq. (5). Here $\mathbf{x}$, $\mathtt{Y}$ and $\mathtt{Z}$ are obtained directly from the affine motion matrix $\tilde{\mathtt{M}}_i = [\mathbf{x}|\mathtt{Y}|\mathtt{Z}]$, recovered through SVD. Our aim is now to reformulate the minimization of $J(\mathbf{u},\mathtt{A},\mathtt{B})$ only as a function of the common axis $\mathbf{u}$ such that:

$$\min_{\mathbf{u},\mathtt{A},\mathtt{B}} J(\mathbf{u},\mathtt{A},\mathtt{B}) = \min_{\mathbf{u}} J(\mathbf{u}) \tag{13}$$

This is possible as we will show that, once the optimal $\mathbf{u}$ is estimated, it is straightforward to obtain $\mathtt{A}$ and $\mathtt{B}$ in closed form. The equivalent cost function $J(\mathbf{u})$ can be written as:

$$\min_{\mathbf{u}} J(\mathbf{u}) = \min_{\mathbf{u}} \left\{ \left\|\mathbf{u}-\mathbf{x}\right\|^2 + \phi_Y(\mathbf{u}) + \phi_Z(\mathbf{u}) \right\}. \tag{14}$$

Thus now we will show how to transform the minimization of $\left\|\mathtt{A}-\mathtt{Y}\right\|^2$ into the minimization of $\phi_Y(\mathbf{u})$ (the same reasoning can be replicated for $\phi_Z(\mathbf{u})$). First, we use the polar decomposition to change variables as $\mathtt{A} = \mathtt{PQ}$ where $\mathtt{P} \succeq 0$ (i.e. $\mathtt{P}$ is a semidefinite matrix) and $\mathtt{Q}$ is orthogonal (both $\mathtt{P}$ and $\mathtt{Q}$ are 2×2). Moreover, given the metric constraints in eq. (5), it follows that $\mathtt{P}^2 = \mathtt{I} - \mathbf{u}\mathbf{u}^T$. Thus, the matrix $\mathtt{I} - \mathbf{u}\mathbf{u}^T$ must be positive definite, restricting the vector $\mathbf{u}$ to be inside the unitary circle. Then, for a chosen $\mathbf{u}$ we can write $\phi_Y(\mathbf{u})$ as:

$$\begin{aligned}
\phi_Y(\mathbf{u}) &= \min_{\mathtt{QQ}^T=\mathtt{I}} \left\|(\mathtt{I}-\mathbf{u}\mathbf{u}^T)^{1/2}\mathtt{Q}-\mathtt{Y}\right\|^2 \\
&= \min_{\mathtt{QQ}^T=\mathtt{I}} \left\{ \left\|(\mathtt{I}-\mathbf{u}\mathbf{u}^T)^{1/2}\right\|^2 + \left\|\mathtt{Y}\right\|^2 \right. \\
&\qquad \left. - 2\operatorname{Tr}\left(\mathtt{Y}^T\left(\mathtt{I}-\mathbf{u}\mathbf{u}^T\right)^{1/2}\mathtt{Q}\right) \right\}.
\end{aligned}$$

Minimizing this cost function over the orthogonal matrix $\mathtt{Q}$ equals to maximizing the trace in the previous expression.

Using the property:

$$\max_{\mathtt{QQ}^T=\mathtt{I}} \left\{\operatorname{Tr}(\mathtt{XQ})\right\} = \sigma_1(\mathtt{X}) + \sigma_2(\mathtt{X}) + \cdots + \sigma_n(\mathtt{X}) = \left\|\mathtt{X}\right\|_{\mathrm{N}} \tag{15}$$

where $\left\|\mathtt{X}\right\|_{\mathrm{N}}$ denotes the *nuclear norm* of $\mathtt{X}$ (i.e. the sum of its singular values), we can write that:

$$\phi_Y(\mathbf{u}) = 2 - \left\|\mathbf{u}\right\|^2 + \left\|\mathtt{Y}\right\|^2 - 2\left\|\left(\mathtt{I}-\mathbf{u}\mathbf{u}^T\right)^{1/2}\mathtt{Y}\right\|_{\mathrm{N}} \tag{16}$$

The same reasoning can be replicated for $\phi_Z(\mathbf{u})$ giving the final optimization problem to be solved as:

$$\min_{\|\mathbf{u}\| \leq 1} \quad -\|\mathbf{u}\|^2 - 2\mathbf{u}^T\mathbf{x} - 2\left\|\left(\mathtt{I} - \mathbf{u}\mathbf{u}^T\right)^{1/2}\mathtt{Y}\right\|_{\mathrm{N}} - 2\left\|\left(\mathtt{I} - \mathbf{u}\mathbf{u}^T\right)^{1/2}\mathtt{Z}\right\|_{\mathrm{N}} \tag{17}$$

Once the optimal $\mathbf{u}^*$ is found we substitute back in order to recover the solution for $\mathtt{A}$ (and similarly for $\mathtt{B}$). First we obtain $\mathtt{Q}$ from the SVD of $\mathtt{Y}^T(\mathtt{I} - \mathbf{u}^*\mathbf{u}^{*T})^{1/2} \mapsto \mathtt{UDV}^T$ leading to $\mathtt{Q} = \mathtt{VU}^T$. The matrix $\mathtt{P}$ is simply given knowing that $\mathtt{P}^2 = \mathtt{I} - \mathbf{u}^*\mathbf{u}^{*T}$. This will result in the matrix that exactly satisfies the metric structure of a hinge joint. The optimization of the cost function in eq. (17) is not trivial since the cost function is nonconvex and nonsmooth. However the domain in which the function resides is very constrained (i.e. the unitary circle) and the value of eq. (17) for an arbitrary $\mathbf{u}$ can be computed efficiently without the need of calculating the nuclear norm at each sample (see Appendix B for details). With such constraints, it is realistic to search for the minimum using a simple brute force procedure which can efficiently compute the function samples in a small amount of time.

Initialization. We first consider the two bodies separately and perform a rigid factorization for each shape. Given this factorization, we can then obtain an initial closed form solution for the metric upgrade in the case of a hinge using Tresadern and Reid's [13] linear approximation.

4. Reconstruction with Missing Data

In this section we show a modification of our algorithm to deal with the case of incomplete measurement matrices. The algorithm can be seen as an extension to the case of deformable and articulated structure of Marques and Costeira's algorithm for rigid scenes [9]. The key idea is that the metric projection can be used to estimate the missing entries in the matrix $\mathtt{W}$ since they are projected to the correct *motion manifold* at each iteration. The steps of this method are summarised in Algorithm 2.

Algorithm 2 3D reconstruction with missing data

```
Require: An initial estimate $\mathtt{W}^{(0)}$ of the missing data in $\mathtt{W}$.
Ensure: A factorization of $\mathtt{W}$ that satisfies the given metric constraints.
Remove the 2D centroid $\mathtt{T}^{(t)}$ from $\mathtt{W}^{(t)}$, i.e. $\hat{\mathtt{W}}^{(t)} = \mathtt{W}^{(t)} - \mathtt{T}^{(t)}$.
Factorize $\hat{\mathtt{W}}^{(t)} = \mathtt{M}^{(t)}\mathtt{S}^{(t)}$ using Steps 1, 2 and 3 of Algorithm 1.
Estimate the missing data entries of $\mathtt{W}$ as $\mathtt{W}^{(t+1)} = \mathtt{M}^{(t)}\mathtt{S}^{(t)} + \mathtt{T}^{(t)}$
Repeat until convergence.
```

The algorithm requires an initial estimate of the missing entries in the measurement matrix $\mathtt{W}$. For this purpose, we have used the Marques and Costeira's rigid factorization algorithm [9]. In the case of articulated structure we apply the algorithm independently to each of the bodies.

The iterations are stopped when the distance $||\mathtt{W}^{(t+1)} - \mathtt{W}^{(t)}||_F$ falls below a user-defined threshold, that is, when the new estimate does not modify the previous values much.

5. Experiments

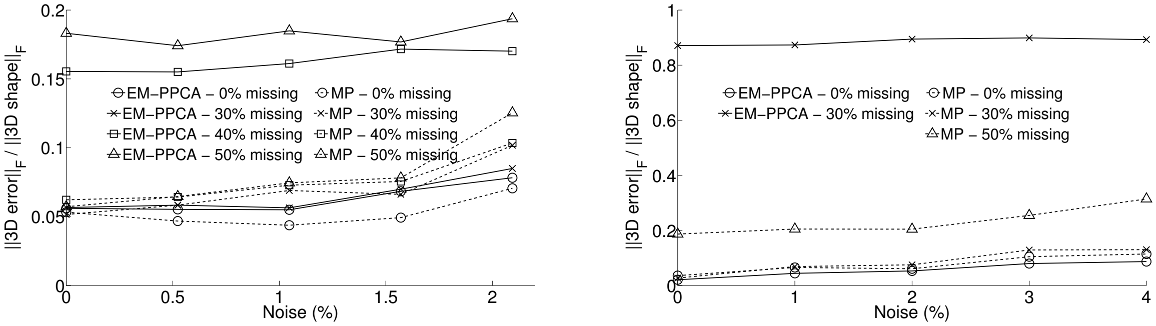
5.1. Synthetic Experiments

Deformable Structure – Motion capture data

In our synthetic experiments we used two different 3D motion capture sequences, both showing faces. The first sequence, *Face1*, was captured in our own laboratory using a VICON system tracking a subject wearing 37 markers on the face. The 3D points were then projected synthetically onto an image sequence 74 frames long using an orthographic camera model. The second sequence, *Face2*, used the motion capture data made available by Torresani *et al.* and described in [12]. The subject wore 40 markers and the original sequence was sub-sampled down to 106 frames.

To test the performance of our algorithm we computed the 3D error, measured in the camera coordinate system, as the sum of the squared differences between the estimated 3D shapes and the ground truth divided by the norm of the shape. We evaluated the performance of the algorithm with respect to noise in the image measurements and varying levels of missing data. Zero mean additive Gaussian noise was applied with standard deviation $\sigma = \mathrm{n} \times \mathrm{s}/100$ where n is the noise percentage and s is defined as $\max(\mathtt{W})$ in pixels. We ran experiments for noise levels of up to 2% for the *Face1* sequence and up to 4% for the *Face2* sequence. Missing data ratios of 0%, 30%, 40% and 50% were generated randomly for each test. In all experiments the number of basis shapes was fixed to $k = 5$. The trials for each level of noise were averaged 10 times.

In Figure 1 we compare the results of our algorithm (MP) with Torresani *et al.*'s algorithm [12] (EM-PPCA). While in the case of full data the performance of both algorithms is comparable for both sequences, our proposed algorithm clearly outperforms EM-PPCA in all cases of missing data (excluding 30% missing data in *Face1* sequence for which they had comparable performance). We also tested the performance of the bundle adjustment approach described in [6]. While the results with complete data were comparable to EM-PPCA and MP, for missing data levels of 30% the 3D error was 120% (MP gave 6% error). Figure 2(A) shows front and side views of the ground truth (squares) and reconstructed 3D shapes (crosses) for three frames of the *Face1* sequence in the absence of noise and missing data. Figure 2(B) shows the 3D reconstructions achieved in the case of 30% missing data. Missing data points are highlighted using red circles.



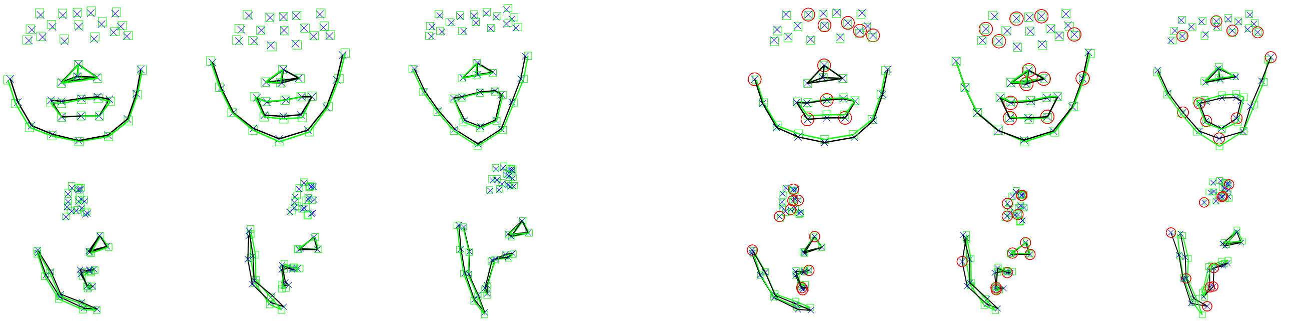
Face1 sequence

Face2 sequence

Figure 1. Synthetic results for 5 different levels of noise and different ratios of missing data for *Face1* (left) and *Face2* (right) sequences. The graphs compare the results obtained using Torresani *et al.*'s EM-PPCA algorithm and our Metric Projection algorithm (MP).

(A) Complete data

(B) Missing data (30%)



Frame 10 Frame 45 Frame 70

Frame 10 Frame 45 Frame 70

Figure 2. (A) Front and side views of the ground truth (squares) and reconstructed 3D shapes (crosses) for three frames of the *Face1* sequence in the absence of noise and missing data. (B) 3D reconstructions achieved in the case of 30% missing data (marked with a ○).

Articulated Structure

In the articulated case our synthetic data simulated two 3D boxes joined by a hinge and projecting this 3D shape into synthesized images via orthographic projection. The sequences contained global rotation and translations as well opening and closing of the hinge. Each box contains 231 points, and the sequence is 63 frames long. We tested the algorithm in the case of full data for noise levels ranging from 0% to 4%. Figure 3 shows the absolute error in the recovered relative angle between the two boxes (averaged over all frames) and the 3D error of recovered 3D structure. The plots in Figure 3 show comparative results between the performance of Tresadern and Reid's [13] algorithm (TR) and our new approach (MP). Slightly superior results are achieved with our algorithm.

5.2. Real Sequences

Deformable Shapes

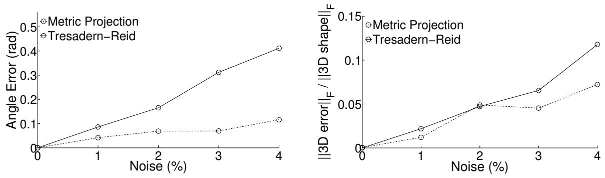


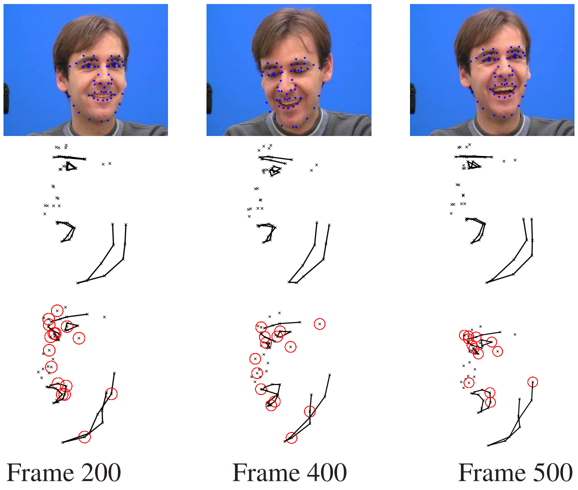
Figure 3. Left: Error on relative rotation angle between the two boxes in the synthetic experiment compared with Tresadern and Reid's linear approach. Right: 3D error of recovered structure

In these experiments we used the Talking Face video[2] taken from a video of a person engaged in conversation. We selected 700 frames from the 5000 frame sequence. An Active Appearance Model (AAM) was used to track 68 features on the face. Figure 4 shows three frames of the original images and a view of the resulting 3D reconstruction in the cases of complete 2D data and 30% missing data. The number of basis shapes was chosen to be 6 in this case. Figure 4 shows good 3D reconstructions are achieved in both cases.

Articulated Shape

We tested our algorithm on a sequence of 815 frames of two boxes linked by a hinge joint. The number of tracked points

[2] www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html



Frame 200 Frame 400 Frame 500

Figure 4. The Franck sequence (first row) used for our real experiment. Second row shows the 3D reconstruction using full data, third row is the resulting 3D shape with 30% missing data in the input tracks. Missing points not visible in the corresponding frame are highlighted with a red circle.

on the upper box was 21 and 47 on the lower box. Figure 5 shows two frames of the image sequence showing the tracked points and the recovered joint axis projected onto the images. The 3D reconstruction of the articulated structure together with the common hinge axis is also shown in Figure 5. In this case there was no missing data.

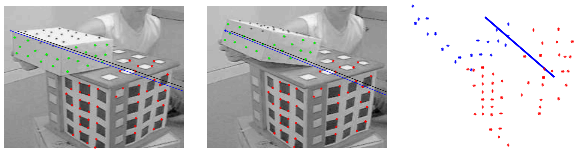


Figure 5. Two images from the articulated sequence. The black line represent the hinge location computed with the algorithm of [13] while the blue line is the solution given by our method. The last figure shows the final 3D reconstruction of our approach.

6. Conclusions

We have described a new alternating least-squares approach associated with a globally optimal projection step onto the manifold of metric constraints. At each step of the minimization we project the motion matrices onto the correct deformable or articulated metric *motion manifolds* respectively. This constraint results in a non-convex problem which can then be solved by semidefinite tight relaxation in the case of deformable shape. In the articulated case, we efficiently compute the joints given the non-linear constraints on the motion of the two bodies. We show results for both synthetic and real video sequences and in the presence of missing data.

7. Acknowledgements

This research has received funding from the European Research Council under ERC Starting Grant agreement 204871-HUMANIS. The Portuguese research team was partially supported by FCT, under ISR/IST pluri-anual funding (POSC program, FEDER), grant SIPM-PTDC/EEA-ACR/73749/2006, grant MODI-PTDC/EEA-ACR/72201/2006, the project ISFL-1-1431 and scholarship SFRH/BPD/26607/2006 .

References

[1] H. Aanæs and F. Kahl. Estimation of deformable structure and motion. In *Workshop on Vision and Modelling of Dynamic Scenes, Copenhagen, Denmark*, 2002.

[2] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-Fine Low-Rank Structure-from-Motion. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska*, 2008.

[3] M. Brand. A direct method for 3D factorization of nonrigid motion observed in 2D. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, pages 122–128, 2005.

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 690–696, June 2000.

[5] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, New York, NY*, 2006.

[6] A. Del Bue, F. Smeraldi, and L. Agapito. Non-rigid structure from motion using ranklet–based tracking and non-linear optimization. *Image and Vision Computing*, 25(3):297–310, March 2007.

[7] M. Dodig, M. Stoŝić, and J. Xavier. On minimizing a quadratic function on stiefel manifolds. Technical report, Instituto de Sistemas e Robotica, 2009. Available at http://users.isr.ist.utl.pt/~jxavier/ctech.pdf.

[8] R. Hartley and R. Vidal. Perspective nonrigid shape and motion recovery. In *Proc. European Conference on Computer Vision*, 2008.

[9] M. Marques and J. Costeira. Estimating 3d shape from degenerate sequences with missing data. *Computer Vision and Image Understanding*, 113(2):261–272, 2009.

[10] J. Sturm. Using SeDuMi 1.02, A Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1):625–653, 1999.

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

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

[13] P. Tresadern and I. Reid. Articulated structure from motion by factorization. In *Proc. IEEE Conference on Computer*

Vision and Pattern Recognition, San Diego, California, volume 2, pages 1110–1115, June 2005.

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

[15] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *Proc. 10th International Conference on Computer Vision, Beijing, China*, October 2005.

[16] J. Yan and M. Pollefeys. A factorization-based approach for articulated non-rigid shape, motion and kinematic chain recovery from video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5), May 2008.

Appendix A: Optimization, deformable case

For $\mathtt{E} \in \mathbb{R}^{6\times 6}$, our aim is to compute

$$\min_{\mathbf{q}=vec(\mathtt{Q})} \mathbf{q}^T \mathtt{E}\mathbf{q}, \tag{18}$$

where $Q \in \mathbb{R}^{3\times 2}$ runs through Stiefel matrices, i.e. $\mathtt{Q}^T\mathtt{Q} = \mathtt{I}_{2\times 2}$. We rewrite (18) as

$$\min_{\mathbf{q}=vec(\mathtt{Q})} \mathrm{Tr}(\mathtt{E}\mathbf{q}\mathbf{q}^T) = \min_{\mathtt{X}\in S} \mathrm{Tr}(\mathtt{EX}), \tag{19}$$

where S is the set of all real symmetric 6×6 matrices $\mathtt{X} = \begin{bmatrix} \mathtt{A} & \mathtt{B} \\ \mathtt{B}^T & \mathtt{C} \end{bmatrix}$, with $\mathtt{A} \in \mathbb{R}^{3\times 3}$, satisfying

$$\mathtt{X} \succcurlyeq 0, \tag{20}$$

$$\mathrm{Tr}(\mathtt{A}) = \mathrm{Tr}(\mathtt{C}) = 1, \quad \mathrm{Tr}(\mathtt{B}) = 0, \tag{21}$$

$$\mathrm{rank}\,\mathtt{X} = 1. \tag{22}$$

This problem, has a nonconvex constraint ($\mathrm{rank}\,\mathtt{X} = 1$). Since the cost function is linear we have

$$\min_{\mathtt{X}\in S} \mathrm{Tr}(\mathtt{EX}) = \min_{\mathtt{X}\in \mathrm{co}(S)} \mathrm{Tr}(\mathtt{EX}), \tag{23}$$

where $\mathrm{co}(S)$ is the convex hull of the set S. Here, we compute the convex hull (tight convex relaxation) $\mathrm{co}(S)$ as all the real symmetric 6×6 matrices $\mathtt{X}$ that satisfy

$$\mathtt{X} \succcurlyeq 0, \tag{24}$$

$$\mathrm{Tr}(\mathtt{A}) = \mathrm{Tr}(\mathtt{C}) = 1, \quad \mathrm{Tr}(\mathtt{B}) = 0, \tag{25}$$

$$\begin{bmatrix} \mathtt{I}_{3\times 3} - \mathtt{A} - \mathtt{C} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succcurlyeq 0, \tag{26}$$

with $\mathbf{w}$ given by

$$\mathbf{w} = \begin{bmatrix} b_{23} - b_{32} \\ b_{31} - b_{13} \\ b_{12} - b_{21} \end{bmatrix} \tag{27}$$

where $\mathtt{B} = [b_{ij}]$. Moreover, this set is defined only by linear matrix inequalities (LMI). Hence, we have that our problem (18) is equivalent to finding the minimum of a linear function ($\mathrm{Tr}(\mathtt{EX})$) on a convex set ($\mathrm{co}(S)$), which is given only by LMI (24)-(26). Thus, the optimization problem in the right-hand side of (23) is a Semi-Definite Program (SDP). By using SeDuMi [10], we quickly obtain the optimal matrix $\mathtt{X}$ for (23). In 100% of experiments that we ran, the optimal matrix X was always of rank 1. By factorizing $\mathtt{X} = \mathbf{q}\mathbf{q}^T$, we obtain the optimal Stiefel matrix as $\mathtt{Q} = vec^{-1}(\mathbf{q})$. For more details the reader can refer to [7]

Appendix B: Optimization, articulated case

We show here that it is possible to compute the *nuclear norm* of a 2×2 matrix $\mathtt{X}$ without explicitly using a SVD at each location in the unit sphere:

$$\begin{aligned} \|\mathtt{X}\|_{\mathtt{N}} &= \sigma_1(\mathtt{X}) + \sigma_2(\mathtt{X}) \\ &= \left(\sigma_1^2(\mathtt{X}) + \sigma_2^2(\mathtt{X}) + 2\sigma_1(\mathtt{X})\sigma_2(\mathtt{X})\right)^{1/2} \\ &= \left(\|\mathtt{X}\|^2 + 2\,|\det(\mathtt{X})|\right)^{1/2}. \end{aligned} \tag{28}$$

Now we can use this result to efficiently evaluate the cost function at any given $\mathbf{u}$ with $\|\mathbf{u}\| \leq 1$. For $\mathbf{u} \neq 0$, we can write $\mathbf{u} = \mathtt{R}\mathbf{v}$ where $\mathtt{R} = \|u\|$ and $\mathbf{v} = \mathbf{u}/\,\|\mathbf{u}\|$. Then, we have the eigenvalue decomposition

$$\mathtt{I} - \mathbf{u}\mathbf{u}^T = \mathtt{V} \begin{bmatrix} 1 - \mathtt{R}^2 & 0 \\ 0 & 1 \end{bmatrix} \mathtt{V}^T$$

where $\mathtt{V} = \begin{bmatrix} \mathbf{v} & \mathtt{J}\mathbf{v} \end{bmatrix}$ $\quad \mathtt{J} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. The matrix $\mathtt{V}$ is orthogonal. Consequently we can write:

$$\left(\mathtt{I} - \mathbf{u}\mathbf{u}^T\right)^{1/2} = \mathtt{V} \begin{bmatrix} \sqrt{1 - \mathtt{R}^2} & 0 \\ 0 & 1 \end{bmatrix} \mathtt{V}^T$$

Let $\mathtt{Y} = \mathtt{V}_Y \Sigma_Y \mathtt{W}_Y^T$ be an SVD for $\mathtt{Y}$. We have:

$$\begin{aligned} &\left\| \left(\mathtt{I} - \mathbf{u}\mathbf{u}^T\right)^{1/2} \mathtt{Y} \right\|_{\mathtt{N}} = \left\| \mathtt{V} \begin{bmatrix} \sqrt{1 - \mathtt{R}^2} & 0 \\ 0 & 1 \end{bmatrix} \mathtt{V}^T \mathtt{V}_Y \Sigma_Y \mathtt{W}_Y \right\|_{\mathtt{N}} \\ &= \left\| \begin{bmatrix} \sqrt{1 - \mathtt{R}^2} & 0 \\ 0 & 1 \end{bmatrix} \mathtt{V}^T \mathtt{V}_Y \Sigma_Y \right\|_{\mathtt{N}}. \end{aligned} \tag{29}$$

Now, let $\mathbf{v} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$ and let, without loss of generality,

$$\mathtt{V}_Y = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \qquad \Sigma_Y = \begin{bmatrix} \boldsymbol{\sigma}_1 & 0 \\ 0 & \boldsymbol{\sigma}_2 \end{bmatrix}.$$

Then,

$$\mathtt{V}^T \mathtt{V}_Y = \begin{bmatrix} \cos(\theta - \alpha) & \sin(\theta - \alpha) \\ -\sin(\theta - \alpha) & \cos(\theta - \alpha) \end{bmatrix}.$$

Substituting into eq. (29), it follows that

$$\begin{aligned} &\left\| \begin{bmatrix} \sqrt{1 - \mathtt{R}^2} & 0 \\ 0 & 1 \end{bmatrix} \mathtt{V}^T \mathtt{V}_Y \Sigma_Y \right\|^2 = \\ &= (1 - \mathtt{R}^2)\big(\sigma_2^2 + (\sigma_1^2 - \sigma_2^2) C_\alpha(\theta)^2\big) + \sigma_1^2 - (\sigma_1^2 - \sigma_2^2) C_\alpha(\theta)^2 \\ &= \sigma_1^2 + \sigma_2^2 - \mathtt{R}^2\sigma_2^2 - \mathtt{R}^2(\sigma_1^2 - \sigma_2^2) C_\alpha(\theta)^2 \end{aligned}$$

where $C_\alpha(\theta) = \cos(\theta - \alpha)$. Additionally we can write:

$$\left| \det \begin{bmatrix} \sqrt{1 - \mathtt{R}^2} & 0 \\ 0 & 1 \end{bmatrix} \mathtt{V}^T \mathtt{V}_Y \Sigma_Y \right| = \sigma_1\sigma_2\sqrt{1 - \mathtt{R}^2}$$

Thus,

$$\begin{aligned} &\left\| \left(\mathtt{I} - \mathbf{u}\mathbf{u}^T\right)^{1/2} Y \right\|_{\mathtt{N}} = \\ &= \left[\sigma_1^2 + \sigma_2^2 - \mathtt{R}^2\sigma_2^2 - \mathtt{R}^2(\sigma_1^2 - \sigma_2^2) C_\alpha(\theta)^2 + 2\sigma_1\sigma_2\sqrt{1 - \mathtt{R}^2}\right]^{1/2} \end{aligned}$$

The same reasoning can be then replicated for the nuclear norm given $\mathtt{Z}$ thus avoiding the costly computations in eq. (28). If $\mathtt{V}_Y = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix}$ note that $C_\alpha(\theta) = \mathbf{v}^T \mathbf{e}_1$.