Selective vs. Global Recovery of Rigid and Non-Rigid Motion

Sami Romdhani [*]
University of Basel
Departement Informatik
Bernoullistrasse 16
CH-4056 Basel Switzerland

Nikolaos Canterakis [†]
Albert-Ludwigs-Universitt
Institut fr Informatik
Georges-Khler-Allee, Gebude 052
D-79110 Freiburg - Germany

Thomas Vetter [‡]
University of Basel
Departement Informatik
Bernoullistrasse 16
CH-4056 Basel Switzerland

Abstract

The problem of “Structure From Motion” (SFM) aims at recovering the 3D shape and motion of flexible objects from their 2D projections on images. Assuming orthographic projection, Tomasi and Kanade introduced a closed form solution to this problem. Later, Bascle and Blake specialized this method to the single image case, assuming that the 3D object belongs to a Linear Object Class for which the model was pre-computed. We present, in this paper, an alternative to this approach which improves the accuracy of the recovery. The previous approach is based on the factorization of a matrix using the low-rank constraints of the problem. Instead of this global estimate, we advocate the use of a selective estimate which we introduce in this paper. We detail both methods and assess their performance quantitatively and qualitatively, we present an efficient implementation of our algorithm and verify the theoretical results by Monte-Carlo simulations and experiment on photographs.

1 Introduction

The problem of “Structure From Motion” (SFM) aims at recovering the 3D shape and motion of flexible objects from their 2D projections on images. In Computer Vision, objects are usually represented as 3D points that undergo rigid and non-rigid motion. Then SFM is subdivided into two problems: first estimating the motion of these points, then separating this motion into a rigid motion (i.e. camera motion) and non-rigid motion (i.e. motion of the 3D points relative to one another). This is generally a non-linear problem that must be resolved using optimization techniques. However under the assumption of orthographic projection, Tomasi and Kanade [Tomasi and Kanade 1991] have shown that the problem is reduced to a bilinear form that can be solved using low rank constraints. To model the non-rigid motion, it is usually assumed [Bascle and Blake 1998; Brand 2001; Irani and Anandan 2000; Bregler et al. 2000; Brand and Bhotika 2001; Torresani et al. 2001], that it can be decomposed into a linear combination of key-motions. Under this assumption, the object is said to belong to a *Linear Object Class* (see Section 2). Then, the recovery of the non-rigid motion reduces to the estimation of the coefficients of the linear combination. Some applications [Brand 2001; Irani and Anandan 2000; Bregler et al. 2000; Brand and Bhotika 2001; Torresani et al. 2001], aim at recovering the rigid motion, *as well as the 3D non-rigid key-motions* and the coefficients of their linear combination *for a sequence of images*. Other [Bascle and Blake 1998] assume that the non-rigid key-motions are available and estimate the non-rigid motion and the linear coefficients *of a single frame* (reviewed in Section 3.1). This is the problem we address in this paper. The approach presented here is based on the fact that several estimates of the non-rigid motion can be computed. In [Bascle and Blake 1998], the non-rigid motion is recovered from an implicit mixture of *all* these estimates. However, we show in Section 3.2 that one of these estimates is *systematically* much more accurate than all others and also than the global estimate of [Bascle and Blake 1998] (in Section 3.3). The theory is then verified on simulations in Section 5.

2 Linear Object Classes

As is generally the case in the field of rigid/non-rigid motion separation [Tomasi and Kanade 1991; Bascle and Blake 1998; Brand 2001; Irani and Anandan 2000; Bregler et al. 2000; Brand and Bhotika 2001; Torresani et al. 2001], the non-rigid motion is modeled as a linear object classes [Vetter and Poggio 1997]. The application proposed in this paper also assumes that the flexible objects under study pertain to the Linear Object Classes. For these objects, the non-rigid 3D deformations vary linearly. One example of Linear Object Classes is human faces [Blanz and Vetter 1999]. The 3D shape of an object is represented by a set of N 3D vertices, arranged into an $N \times 3$ shape matrix, $\mathbf{Q}$. This process by which a continuous shape is discretized into a finite number of vertices is called shape sampling. The shape sampling must be done consistently for different instances of objects belonging to a class. This means that the vertex number i must represent the same feature on all the examples of the Linear Object Class. The shapes produced in this manner are usually referred to as being *in correspondence*. Then, the shape of an object belongs to a linear manifold. Using a set of T example shapes, $\mathbf{Q}_i$, this manifold is computed by a Principal Component Analysis. We denote by $\mathbf{S}_0 = 1/T \sum_i^T \mathbf{Q}_i$ the average of the example shapes and by $\mathbf{S}_i$, $i = 1, \ldots, M$, the i^{th} principal component thereof (M, the number of principal components retained is bounded by $T - 1$). PCA is obtained by performing an SVD decomposition to the data matrix, $\mathbf{A}$, whose columns are the example shapes subtracted by the average shape:

$$\mathbf{A}_{3N \times T} = \big(\operatorname{vec}(\mathbf{Q}_1 - \mathbf{S}_0) \ldots \operatorname{vec}(\mathbf{Q}_T - \mathbf{S}_0)\big) = \mathbf{U}\blacksquare\mathbf{V}^{\mathrm{T}} \tag{1}$$

$$\mathbf{C}_{3N \times 3N} = \frac{1}{T}\mathbf{A} \cdot \mathbf{A}^{\mathrm{T}} = \frac{1}{T}\mathbf{U} \cdot \blacksquare^2 \cdot \mathbf{U} \tag{2}$$

$$\mathbf{S}_{i_{N \times 3}} \doteq \frac{\lambda_i}{\sqrt{T}} \mathbf{U}^{(N)}_{\cdot,i} \tag{3}$$

where $\operatorname{vec}\mathbf{Q}$ vectorizes $\mathbf{Q}$ by stacking its columns, $\mathbf{C}$ is the covariance matrix of the example heads, the columns of $\mathbf{U}$ are the principal components (PC) and $\frac{\lambda_i}{\sqrt{T}}$ are their associated standard deviations (λ_i is the i^{th} diagonal element of the diagonal matrix $\blacksquare$). $\mathbf{S}_i$ is the reshaped i^{th} PC scaled by its standard deviation. $\mathbf{U}_{\cdot,i}$ is the i^{th} column of $\mathbf{U}$, the notation $\mathbf{a}^{(n)}_{m \times 1}$ (see [Minka 2000]) folds the vector $\mathbf{a}$ into an $n \times (m/n)$ matrix. As a result, any shape can be obtained

[*] e-mail: sami.romdhani@unibas.ch

[†] e-mail:canterakis@informatik.uni-freiburg.de

[‡] e-mail: thomas.vetter@unibas.ch

by a linear combination of these M principal components (In the following $\alpha_0 = 1$):

$$\mathbf{S}_{N\times 3} = \sum_{i=0}^{M} \alpha_i \cdot \mathbf{S}_{i_{N\times 3}} \tag{4}$$

The covariance matrix of $\alpha_i, i = 1,\dots,M$ is $\mathbf{I}_M$, the identity matrix (indeed, the projection of the example shapes onto the scaled PCs are the columns of $\mathbf{V}$). Applied to faces, this equation means that the shape of the face of any individual can be obtained from a linear combination of principal components.

Assuming a weak perspective, the position of the vertices in an image is computed by the following *Imaging Equation*:

$$\mathbf{X}_{N\times 2} = f \cdot \left(\sum_{i=0}^{M} \alpha_i \cdot \mathbf{S}_i \right) \cdot \mathbf{R}_{3\times 2} + \mathbf{1}_{N\times 1} \cdot \mathbf{t}^{\mathrm{T}}_{2\times 1} \tag{5}$$

where the rows of the *correspondence matrix*, $\mathbf{X}$, hold the (x,y) image frame coordinates of the vertices, f controls the scale of the object in the image, $\mathbf{R}_{3\times 2}$ is the first two columns of a 3D rotation matrix, $\mathbf{t}$ is a 2D translation vector and $\mathbf{1}$ is a column vector full of ones. Note that there is no ambiguity between the scale factor f and the magnitude of the α's as α_0 is constrained to be equal to 1.

Given an image of an object of the class, the rigid/non-rigid transformation separation problem is to recover the parameters which explain the image. More formally, given (i) $\mathbf{X}$, the correspondences between the vertices of the model and an image of the object and given (ii) the PCA of the non-rigid deformations of the Linear Object Class ($\mathbf{S}_i$), recover (i) the imaging parameters f, $\mathbf{R}$ and $\mathbf{t}$ and (ii) the non-rigid parameters α_i.

In this paper we assume that the correspondence problem (that of assigning vertex labels to the pixels of the image) has been addressed: using a correspondence finding algorithm, correspondences were found for the N vertices (see for instance [ano n. d.] for an algorithm able to recover a large amount (in the region of 10.000) of correspondence points). Generally, this correspondence problem is not solved exactly. Hence, the recovered correspondences, $\mathbf{X}$, are noisy: $\mathbf{X} \rightarrow \mathbf{X} + \mathbf{E}$. In this paper we assume that the noise is i.i.d. (independently and identically distributed) and Gaussian with constant variance: $\mathbf{E} \sim N(0, \sigma_N)$. The Imaging equation is then modified to take the noise into account ($\mathbf{S}$ is a horizontal stacking of the $M+1$ shape matrices and α is an $M+1\times 1$ column vector of the non-rigid parameters):

$$\mathbf{X} = f \cdot \mathbf{S}_{N\times 3(M+1)} \cdot (\alpha \otimes \mathbf{R}) + \mathbf{1} \cdot \mathbf{t}^{\mathrm{T}} + \mathbf{E}_{N\times 2} \tag{6}$$

where $\otimes$ denotes the Kronecker (tensor) product [Magnus and Neudecker 1999] which multiplies an $m \times n$ and a $p \times q$ matrices into an $mp \times nq$ matrix. The requirement of this method is to have a sufficient amount of corresponding points: $N > 3(M+1)$. It is assumed that $\mathbf{S}$ has full rank: $\mathrm{rank}(\mathbf{S}) = 3(M+1)$.

3 Parameters Recovery

3.1 Global Method a.k.a. Factorization-based

In this section we outline the *global approach* introduced by Bascle and Blake [Bascle and Blake 1998], which aims at estimating the scale, $\hat{f}$, the non-rigid parameters $\hat{\alpha}$, the rotation matrix $\hat{\mathbf{R}}$ and the 2D translation $\hat{\mathbf{t}}$ from the correspondences affected by noise, Equation (6). It is assumed that the $\mathbf{S}_i$ are centered at the origin and henceforth the translation is the mean of the 2D points:

$$\hat{\mathbf{t}} = \frac{1}{N} \sum_{i}^{N} \mathbf{X}^{\mathrm{T}}_{i,\cdot} \tag{7}$$

where $\mathbf{X}_{i,\cdot}$ is the i^{th} row of $\mathbf{X}$.

Let us denote by $\mathbf{Q}$ the $3(M+1)\times 2$ matrix product of the scale, the rigid and non-rigid parameters: $\mathbf{Q} = f\alpha \otimes \mathbf{R}$. Then, omitting the error matrix, Equation (6) can be rewritten as:

$$\mathbf{S} \cdot \mathbf{Q} = \mathbf{X} - \mathbf{1} \cdot \hat{\mathbf{t}}^{\mathrm{T}} \tag{8}$$

and denoting by $\mathbf{S}^+$, the pseudo-inverse of $\mathbf{S}$, $\mathbf{S}^+ = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{T}}$ we obtain:

$$\hat{\mathbf{Q}} = \mathbf{S}^+ \cdot \left(\mathbf{X} - \mathbf{1} \cdot \hat{\mathbf{t}}^{\mathrm{T}}\right) \tag{9}$$

Then, separation of the rigid and non-rigid parameters is obtained by reshaping the matrix $\hat{\mathbf{Q}}$ into a matrix $\mathbf{P}$ and applying SVD:

$$\mathbf{P}_{6\times(M+1)} = \frac{1}{f}(\mathrm{vec}\,\hat{\mathbf{Q}}^{\mathrm{T}})^{(6)} = \mathbf{U} \cdot \blacksquare \cdot \mathbf{V}^{\mathrm{T}} \simeq \mathrm{vec}\,\hat{\mathbf{R}} \cdot \hat{\alpha}^{\mathrm{T}} \tag{10}$$

where $\blacksquare$ is a diagonal matrix that holds the singular values of $\mathbf{P}$ in descending order $(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_6)$. As this matrix is the outer-product of two vectors, in the noiseless case, $\mathbf{E} = 0$, it has rank 1, and $\mathrm{vec}\,\mathbf{R}$ and α are proportional to the first column of $\mathbf{U}$ and $\mathbf{V}$, respectively. However, generally, due to the noise, $\lambda_i, i > 1$ are not negligible and the estimated rotation matrix, $\hat{\mathbf{R}}$ and non-rigid parameters $\hat{\alpha}$ are noisy. We call these estimates, the global estimates.

Note that the ultimate step of this algorithm is to enforce the orthonormality of the rotation matrix by applying a linear transformation, see [Tomasi and Kanade 1991] for further details.

3.2 Kernel-based Selective Method

The matrix $\mathbf{Q}$ holds a series of 3×2 matrices stacked vertically. In the noiseless case, these matrices are proportional to the rotation matrix:

$$\mathbf{Q} = f\left(\alpha_0\mathbf{R}^{\mathrm{T}} \quad \alpha_1\mathbf{R}^{\mathrm{T}} \quad \dots \quad \alpha_M\mathbf{R}^{\mathrm{T}}\right)^{\mathrm{T}} \tag{11}$$

However, due to the noise, this will not be the case:

$$\hat{\mathbf{Q}} = \hat{f}\left(\hat{\alpha}_0\hat{\mathbf{R}}^{\mathrm{T}}_0 \quad \hat{\alpha}_1\hat{\mathbf{R}}^{\mathrm{T}}_1 \quad \dots \quad \hat{\alpha}_M\hat{\mathbf{R}}^{\mathrm{T}}_M\right)^{\mathrm{T}} \tag{12}$$

where the $\hat{\mathbf{R}}_i$ are different estimates of the rotation matrix. The message of this paper is that *these estimates are not impacted by the noise evenly*, and it turns out that, generally, one of these estimates is much better (i.e. less impacted by the noise) than all the others, included the global estimate of the previous section. Instead of using all equations on equal footing, we use the great redundancy of the equations available in order to find a more accurate solution.

Selective Estimation We demonstrate in the appendix that the pseudo-inverse of $\mathbf{S}$ can be expressed in terms of $\tilde{\mathbf{S}}_i$, the horizontal stacking of all shape matrices but the shape matrix i, and $\mathbf{K}_i$, the kernel of $\tilde{\mathbf{S}}_i$, whose rows are unit-length and mutually orthogonal:

$$\tilde{\mathbf{S}}_{i_{N\times 3M}} \doteq \left(\mathbf{S}_0 \quad \dots \quad \mathbf{S}_{i-1} \quad \mathbf{S}_{i+1} \quad \dots \quad \mathbf{S}_M\right) \tag{13}$$

$$\mathbf{K}_{i_{P\times N}} \cdot \tilde{\mathbf{S}}_i \doteq 0, \text{ and } \mathbf{S}^+ = \begin{pmatrix} (\mathbf{K}_0\mathbf{S}_0)^+\mathbf{K}_0 \\ \dots \\ (\mathbf{K}_M\mathbf{S}_M)^+\mathbf{K}_M \end{pmatrix} \tag{14}$$

where P, the number of kernel vectors is generally equal to $N-3M$. Hence, the i^{th} estimate of $\mathbf{R}$ is:

$$\hat{f}\hat{\alpha}_i\hat{\mathbf{R}}_i = (\mathbf{K}_i\mathbf{S}_i)^+\mathbf{K}_i \cdot \left(\mathbf{X} - \mathbf{1} \cdot \hat{\mathbf{t}}^{\mathrm{T}}\right) \tag{15}$$

This is the least square estimate of $\hat{f}\hat{\alpha}_i\hat{\mathbf{R}}_i$ obtained by mapping the correspondences $\mathbf{X}$ onto $\mathbf{K}_i$. Using the constraint that the columns of a rotation matrix are unit-length enables the determination of

$\hat{\mathbf{R}}_i$ independently of $\hat{f}\hat{\alpha}_i$ from Equation (15). Note that this method allows also the estimation of the 2D translation, $\mathbf{t}$, without using the assumption that the $\mathbf{S}_i$ are centered at the origin made in Section 3.1. Indeed, if $\mathbf{K}$ is the kernel of $\mathbf{S}$, $\mathbf{K}\cdot\mathbf{S} \doteq 0$ (in the general case, $\mathbf{K}$ has $N-3(M+1)$ rows), we have from Equation (6):

$$\hat{\mathbf{t}}^{\mathrm{T}} = (\mathbf{K}\cdot\mathbf{1})^{+}\cdot\mathbf{K}\cdot\mathbf{X} \tag{16}$$

Quality Assessment Equation (15) provides $M+1$ estimators of the rotation matrix. So, a pertinent question would be: Is one of these estimators better than the other ? By better we mean, reduces the sum of squared residuals more. In this paragraph, two criteria of the quality of an estimate $\hat{\mathbf{R}}_i$ are presented. We show qualitatively and quantitatively the fact that one estimator is *systematically* better than the others. The first quantitative criterion uses the constraint that $\mathbf{R}$ holds the first two columns of a rotation matrix; hence $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. So, one scheme to select the estimate $\hat{\mathbf{R}}_i$ which is the least impacted by noise is to use the one that, pre-multiplied by its transposed, is closest (in a Frobenius norm) to the identity matrix. As an example, Table 1, shows $\hat{\mathbf{R}}_i^{\mathrm{T}}\hat{\mathbf{R}}_i$ for $i=0,1$ and 2. In this example, the best estimate is the one for $i=0$, i.e. using the mapping onto $\mathbf{S}_0$, the mean shape. We, now, present a second quantitative criterion which is based on the minimum sum of squared residuals and does not use the constraint on $\mathbf{R}$. To obtain the i^{th} estimate, the matrix $\mathbf{X}$ must be mapped onto $\mathbf{K}_i$ (for clarity, in the following equations, the translation $\mathbf{t}$ and the noise are omitted):

i=0	i=1	i=2
$\begin{pmatrix} 1.00 & -0.01 \\ -0.01 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & -0.94 \\ -0.94 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & -0.89 \\ -0.89 & 1.00 \end{pmatrix}$

Table 1: $\hat{\mathbf{R}}_i^{\mathrm{T}}\hat{\mathbf{R}}_i$ for $i=0,1,2$ which should ideally be equal to $\mathbf{I}_2$.

$$\mathbf{K}_i\cdot\hat{\mathbf{X}} = \hat{f}\mathbf{K}_i\cdot\mathbf{S}_i\hat{\alpha}_i\hat{\mathbf{R}} \tag{17}$$

Then the fitted correspondence matrix $\hat{\mathbf{X}}$ back-mapped into its original space is obtained by plugging Equation (15) into the previous Equation:

$$\mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\hat{\mathbf{X}} = \mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\mathbf{S}_i\cdot(\mathbf{K}_i\cdot\mathbf{S}_i)^{+}\cdot\mathbf{K}_i\cdot\mathbf{X} \tag{18}$$

So the sum of squared residuals for the i^{th} estimate is:

$$\begin{aligned} \mathrm{SSR}_i &= \|\mathbf{X}-\mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\hat{\mathbf{X}}\|^2 && (19)\\ &= \|\mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\mathbf{X}-\mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\hat{\mathbf{X}}\|^2 && \\ &\quad+\|\mathbf{X}-\mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\mathbf{X}\|^2 && (20)\end{aligned}$$

The first term of Equation (20), SSR_i^1 is the SSR of the fitted correspondence matrix projected orthogonally into the row-space of $\mathbf{K}_i$ and the second term, SSR_i^2 is the SSR between the correspondence matrix and its projection into the row-space of $\mathbf{K}_i$. SSR_i^2 depends on the correlation between $\mathbf{X}$ and $\mathbf{S}_i\mathbf{R}$. The higher the correlation, the lower the SSR will be. To illustrate the case, let us assume that $\mathbf{X} \propto \mathbf{S}_k\mathbf{R}$, i.e. $\alpha_i = 0$ for $i \neq k$, then SSR_i^2 is equal to $\|\mathbf{X}\|^2$ for $i \neq k$. Moreover if the shape matrices $\mathbf{S}_i$ are mutually orthogonal, then SSR_k^2 is null. Recall that the $\mathbf{S}_i$ are obtained by PCA and, as a result, are orthogonal. Hence SSR_i^2 is minimum for $i=k$. SSR_i^1, i.e. the SSR of the fitted correspondence matrix projected into the row-space of $\mathbf{K}_i$, is equal to $\|\mathbf{K}_i^{\mathrm{T}}\cdot\mathbf{K}_i\cdot\mathbf{E}\|^2$. As we assumed that the noise was isotropic, its projection into the row-space of $\mathbf{K}_i$ does not depend on i. As a conclusion, the SSR_i is minimum for the i for which $\mathbf{S}_i\mathbf{R}$ is the most correlated with $\mathbf{X}$. For shapes which belongs to the linear object class, this correlation is given by $\lambda_i/\sqrt{T}$ (see Equation (3)), i.e. by the Frobenius norm of $\mathbf{S}_i$. PCA asserts that the correlation is maximum for $i=0$, the average shape (and decreases exponentially). As an example, the norm of $\mathbf{S}_i$ for $i=1,2$ relative to $\mathbf{S}_0$ for the 3D Morphable Face Model is 0.0232 and 0.0168. Hence, $\hat{\mathbf{R}}_0$ (see Equation (15)) is the best estimation of $\mathbf{R}$.

This result can be seen in yet another way. The signal-to-noise ratio for the i^{th} estimate is:

$$\mathrm{SNR}_i = \frac{\|(\mathbf{K}_i\cdot\mathbf{S}_i)^{+}\cdot\mathbf{K}_i\cdot\mathbf{X}\|^2}{\|(\mathbf{K}_i\cdot\mathbf{S}_i)^{+}\cdot\mathbf{K}_i\cdot\mathbf{E}\|^2} = \frac{\mathrm{SNR}^n}{\mathrm{SNR}^d} \tag{21}$$

For the aforementioned reasons, the denominator of Equation (21), SNR^d, is similar for all i, but its numerator, SNR^n, depends on the correlation between $\mathbf{X}$ and $\mathbf{S}_i\mathbf{R}$ which is maximum for $i=0$.

Non-rigid parameters Now that we have found a good estimate for $\mathbf{R}$, $\hat{\mathbf{R}}_0$, we can plug this estimate into the imaging equation and obtain a linear system in α and f. Omitting the noise, Equation (6) can be rewritten as:

$$\mathrm{vec}\left(\mathbf{X}-\mathbf{1}\cdot\mathbf{t}^{\mathrm{T}}\right) = f\cdot\left(\mathrm{vec}(\mathbf{S}\cdot(\hat{\mathbf{R}}_0\otimes\mathbf{I}_M))\right)^{(2N)}\cdot\alpha \tag{22}$$

and hence,

$$f\cdot\alpha = \left(\left(\mathrm{vec}(\mathbf{S}\cdot(\hat{\mathbf{R}}_0\otimes\mathbf{I}_M))\right)^{(2N)}\right)^{+}\cdot\mathrm{vec}\left(\mathbf{X}-\mathbf{1}\cdot\hat{\mathbf{t}}^{\mathrm{T}}\right) \tag{23}$$

and f is obtained using the constraint that $\alpha_0 = 1$.

3.3 Comparison: Selective vs. Global

We denote by $\blacksquare\mathbf{P} = \hat{\mathbf{P}}-\mathbf{P}$ the error made in estimating the matrix $\mathbf{P}$ (Equation (10)). We showed in the previous Section that the error made on the estimate $\hat{\mathbf{R}}_0$, i.e. the first column of $\hat{\mathbf{P}}$, is much lower than the one made on any other estimate $\hat{\mathbf{R}}_i$. We now investigate how this error is propagated to the global estimate computed in Section 3.1. The following developments are based on [Sun et al. 2001]. Let us define the square matrix $\mathbf{W}$ whose eigenvectors are the columns of the matrix $\mathbf{U}$ (of Equation (10)) and the eigenvalues are λ_i^2:

$$\mathbf{W} = \mathbf{P}\cdot\mathbf{P}^{\mathrm{T}} = \mathbf{U}\cdot\Lambda^2\cdot\mathbf{U}^{\mathrm{T}} \tag{24}$$

Due to the noise, there is a perturbation $\blacksquare\mathbf{P}$ and henceforth a perturbation on $\mathbf{W}$, $\blacksquare\mathbf{W} = \hat{\mathbf{P}}\cdot\hat{\mathbf{P}}^{\mathrm{T}}$, and to a first order approximation $\blacksquare\mathbf{W} \simeq \blacksquare\mathbf{P}\cdot\mathbf{P}^{\mathrm{T}}+\mathbf{P}\cdot\blacksquare\mathbf{P}^{\mathrm{T}}$. Now, we would like to compare $\blacksquare\mathbf{P}_{\cdot,1}/\|\mathbf{P}_{\cdot,1}\|$, the error made on $\mathrm{vec}(\hat{\mathbf{R}}_0)/\|\mathrm{vec}(\hat{\mathbf{R}}_0)\|$ (i.e. the selective estimate), to $\blacksquare\mathrm{U}_{\cdot,1}$, the error made on the first singular vector of $\mathbf{P}$ (i.e. the global estimate of $\mathrm{vec}\,\hat{\mathbf{R}}$). The error which has lowest norm is to be selected. It is shown in [Weng et al. 1989] that, to a first order approximation:

$$\blacksquare\mathbf{U}_{\cdot,1} \simeq \mathbf{U}\cdot\Delta\cdot\mathbf{U}^{\mathrm{T}}\cdot\blacksquare\mathbf{W}\cdot\mathbf{U}_{\cdot,1} \tag{25}$$

where $\Delta = \mathrm{diag}(0,(\lambda_2-\lambda_1)^{-1},\ldots,(\lambda_6-\lambda_1)^{-1})$. To be able to compare this error with $\blacksquare\mathbf{P}$, we need to express $\blacksquare\mathbf{W}\cdot\mathbf{U}_{\cdot,1}$ as a matrix right-multiplied by $\blacksquare\mathbf{P}$. Using Equation (8) p. 31 of [Magnus and Neudecker 1999], $\blacksquare\mathbf{W}\cdot\mathbf{U}_{\cdot,1} = (\mathbf{U}_{\cdot,1}^{\mathrm{T}}\otimes\mathbf{I}_6)\cdot\mathrm{vec}(\blacksquare\mathbf{W})$. Then using Equation (7) p. 31 and Equation (1) p. 47, of the same reference:

$$\mathrm{vec}(\blacksquare\mathbf{P}\cdot\mathbf{P}^{\mathrm{T}}) = (\mathbf{P}\otimes\mathbf{I}_6)\cdot\mathrm{vec}(\blacksquare\mathbf{P}) \tag{26}$$

$$\mathrm{vec}(\mathbf{P}\cdot\blacksquare\mathbf{P}^{\mathrm{T}}) = (\mathbf{I}_6\otimes\mathbf{P})\cdot\mathbf{K}_{6,2}\cdot\mathrm{vec}(\blacksquare\mathbf{P}) \tag{27}$$

where $\mathbf{K}_{6,2}$ is the 12×12 commutation matrix which transforms $\mathrm{vec}(\mathbf{A})$ into $\mathrm{vec}(\mathbf{A}^{\mathrm{T}})$ (see [Magnus and Neudecker 1999; Minka 2000]). Then,

$$\begin{aligned}\blacksquare\mathbf{U}_{\cdot,1} \simeq & \mathbf{U} \cdot \Delta \cdot \mathbf{U}^{\mathrm{T}} \cdot (\mathbf{U}_{\cdot,1}^{\mathrm{T}} \otimes \mathbf{I}_6) \\ & \cdot \left(\mathbf{P} \otimes \mathbf{I}_6 + (\mathbf{I}_6 \otimes \mathbf{P}) \cdot \mathbf{K}_{6,2}\right) \cdot \mathrm{vec}(\blacksquare\mathbf{P})\end{aligned} \tag{28}$$

Denoting by $\mathbf{A}$ the matrix which maps $\mathrm{vec}(\blacksquare\mathbf{P})$ into $\blacksquare\mathbf{U}_{\cdot,1}$, the selective method yields a better estimate if:

$$\|\mathbf{A} \cdot \mathrm{vec}(\blacksquare\mathbf{P})\| > \gamma \cdot \| \left[\mathbf{I}_6 \quad \mathbf{0}_{6\times 6M}\right] \cdot \mathrm{vec}(\blacksquare\mathbf{P})\| \tag{29}$$

where $\gamma = 1/\|\mathbf{P}_{\cdot,1}\|$. Note that the right side of the inequality is equal to $1/\,\mathrm{SNR}_0$. From Equation (28), it can be seen that the error made on the first singular vector of $\mathbf{P}$ is a linear combination of the error made on all $\mathbf{P}_{\cdot,1}$. Therefore, the selective estimate will in general be better. In the Monte-Carlo simulations presented in Section 5, the selective method yielded always more accurate results.

4 Efficiency

The speed bottleneck of the algorithm presented here is the computation of the kernel $\mathbf{K}_0$. Naturally, we could compute this kernel off-line if we selected a constant set of N vertices of the 3D Morphable Model on which to perform the computations. Unfortunately these N vertices may not be visible on all images, as the set of visible vertices depends on the pose of the object and hence varies from images to images. So, instead of pre-computing a single $\mathbf{K}_0$ for one set of vertices, we compute several ones for vertices visible at different poses and choose at run-time the kernel with the minimum number of hidden vertices. As, generally, this kernel would still be computed with vertices hidden for a particular image, it must be updated on-line such as the updated kernel, $\mathbf{K}^*_{V-3M\times V}$, would be the same as the one computed on the set of V visible vertices only. The rational is that it is much more efficient to update the kernel rather than to compute it from scratch. The kernel of a matrix $\mathbf{S}$ can be computed by performing a QR Factorization [Golub and van Loan 1996] (the kernel is the transpose of the $N-3M$ columns of $\mathbf{Q}$ associated to the zero diagonal elements of $\mathbf{R}$). Now, we wish to obtain the updated matrix $\mathbf{Q}^*$ which would factor the matrix $\mathbf{S}^*$ obtained by deleting the $N-V$ rows of $\mathbf{S}$ associated to the hidden vertices. There is an algorithm for doing exactly that in Section 12.5.3 of [Golub and van Loan 1996] based on Givens rotations, whose complexity is $O((N-V)\cdot N)$. Note that for a reasonable discretization of the pose sphere, $N-V$ is usually two orders of magnitude lower than N. For increased performances the Givens rotations are implemented using the Fast Scaled Givens Transformations algorithm [Anda 1995]. Then, the pseudo-inverse is computed by QR factorization implemented again using the Fast Givens Rotations whose complexity is $O(P)$, recall that P is the number of kernel vectors. So, the global efficiency of the recovery is determined by the matrix-matrix multiplication and is $O(PN)$. A MATLAB implementation of the whole algorithm (recovery of the rigid and non-rigid parameters), using $N = 1000$ vertices among which 10 are hidden, runs in 0.5s on a 2GHz Pentium computer.

5 Experiments

Using a 3D Morphable Model [Blanz and Vetter 1999] of faces, we can render photo-realistic images of faces. A 3D Morphable Model is constituted by a shape model and a texture (or color) model. It adheres to the linear object class formalism and, as a result, the shape model is that described by Equation (4). We rendered images of faces using Equation (5) by performing Monte-Carlo Simulation on the parameters $\alpha, \mathbf{R}, f$ and on the noise $\mathbf{E}$ (see Figure 5). As these images are synthetic we dispose of a large number of exact correspondence points. Additionally, we can easily measure the accuracy of the recovery as the ground truth is also known exactly.

Two 3D Morphable Models are computed, each on a training set of 100 different individuals. The first 50 principal components are retained for both models. The first model is used to generate the ground truth face images (and their 2D correspondence points), the second is used in the recovery algorithm. We adopt this scheme to ensure that the method could cope with novel instances of the object that do not *exactly* belong to the linear object class.

We verified that, when there is no noise and when the same Morphable Model is used for the ground truth and for the recovery, only $N = 3M + 3$ (i.e. 2 kernel vectors) are sufficient to recover the 3 rotation angles, the scale, the translation and the non-rigid parameters perfectly.

In this section we first test, on synthetic examples, the accuracy of the method as a function of the noise, then we measure it as a function of the number of correspondence points. Additionally we show an example of recovery of the parameters on a photograph for which the correspondence were computed by an automatic algorithm [ano n. d.].

Test Set for the Monte Carlo simulation As we tested our method on synthetic images we know the exact correspondence points given by Equation (6). We generated 50 sets of parameters. We rendered 512×512 images in which the faces were scaled to occupy the entire image. We added a Gaussian noise with zero mean and standard deviations $\sigma_N = 0, 1, 2, \ldots, 10$ (i.e. 550 experiments were conducted).

5.1 Selective vs. Global

In this section, we compare the accuracy of the global and the selective methods. The experiments were conducted using $N = 1000$ corresponding points. Table 2 presents various statistics averaged over all 550 experiments reflecting the accuracy of the selective approach. The quality of the selective and global recoveries are indicated by the comparison of $\|\blacksquare\mathbf{U}_{\cdot,1}\| = 0.747$ to the last column of the table.

i	$\|\hat{\mathbf{R}}_i^{\mathrm{T}}\hat{\mathbf{R}}_i - \mathbf{I}_2\|$	SSR_i^1	SSR_i^2	SNR_i^n	SNR_i^d	SNR_i	$\frac{\|\blacksquare\mathbf{P}_{\cdot,i}\|}{\|\mathbf{P}_{\cdot,i}\|}$
0	0.029	2.744	96	13570	518	61.9	0.011
1	1.115	2.744	99	814	682	1.38	5.022
2	1.001	2.744	99	518	463	1.27	11.07
3	0.974	2.744	99	492	451	1.21	6.255
4	1.080	2.744	99	431	393	1.29	8.822

Table 2: Accuracy of the selective approach for different *i*.

5.2 Accuracy wrt the Noise

In this section we test the accuracy of the recovery of the rotation and non-rigid parameters as a function of the noise. The experiments were conducted using $N = 1000$ corresponding points. The estimate of $\mathbf{R}$, $\hat{\mathbf{R}}_0$, is obtained using Equation (15) and normalizing its columns. The $\hat{\alpha}$ and $\hat{f}$ are recovered by Equation (23).

We show the mean and the standard deviation of the absolute value of the error of the recovery of the rotation angles and the non-rigid parameter as a function of σ_N in the Figures 1 and 2.

To measure the estimation error on the non-rigid parameters, we chose the normalized correlation, which is a good measure used for

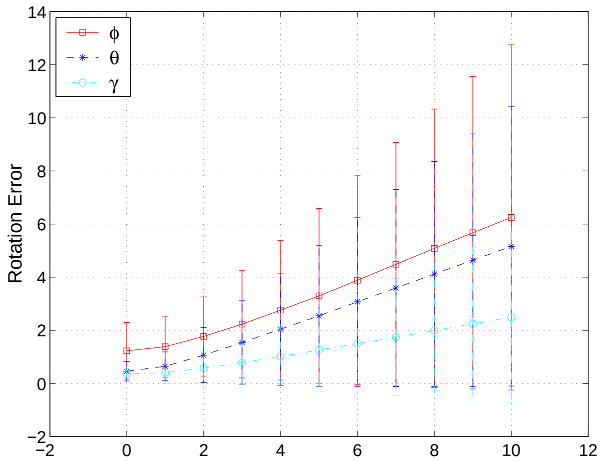


Figure 1: Absolute value of the difference between the ground truth *rotation angles* and the ones recovered by the algorithm (in degrees) as a function of the standard deviation of the correspondence noise.

nearest-neighbor identification. If α are the ground truth parameters and $\hat{\alpha}$ the recovered parameters, the normalized correlation is:

$$\frac{\alpha^{\mathrm{T}} \cdot \hat{\alpha}}{\|\alpha\| \cdot \|\hat{\alpha}\|} \tag{30}$$

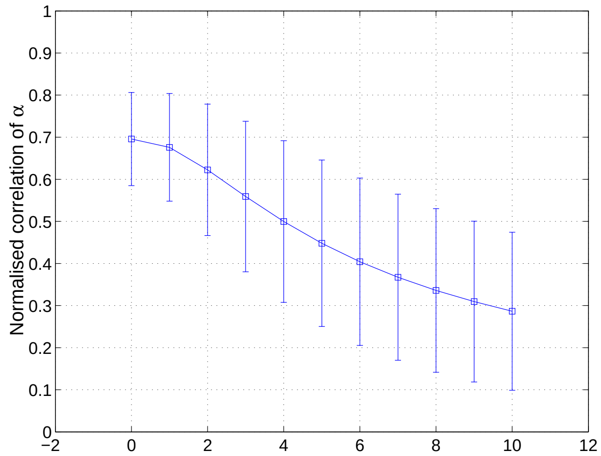


Figure 2: Normalized correlation between the ground truth α_i and the one recovered by the algorithm as a function of the standard deviation of the correspondence noise.

5.3 Accuracy wrt the Number of Corresponding Points

In this section we measure the accuracy of the algorithm with respect to the number of correspondence points available for three noise levels ($\sigma_N = 2, 5$, and 8). Figures 3 and 4 show the results for the rotation angle ϕ (azimuth) and the non-rigid parameters. In these graphs, to increase the visibility, only the means are drawn.

Figure 5 shows renderings of the ground truth along with renderings of the recovered parameters for different number of correspondence points and various noise levels.

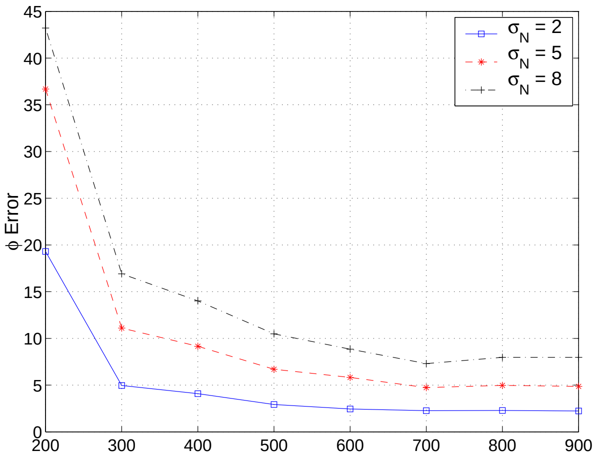


Figure 3: Error in ϕ as a function of N, the number of correspondence points used, for three values of the noise.

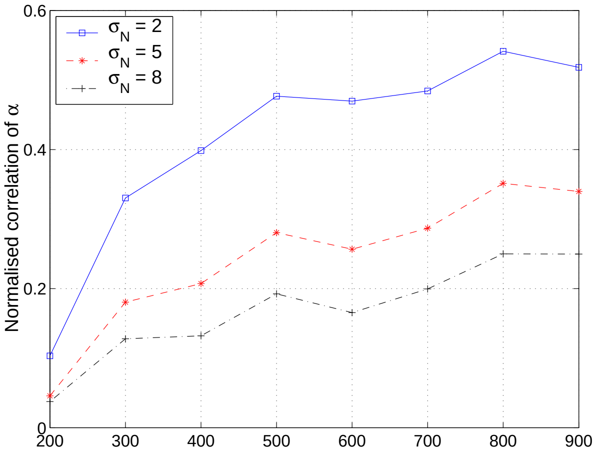


Figure 4: Normalized correlation between the ground truth α_i and the one recovered by the algorithm as a function of N, the number of correspondence points used for three values of the noise.

5.4 Experiment on a Photograph

We present on Figure 5.4 an example of recovery obtained on one of the CMU-PIE face image [Baker and Matthews 2001]. Correspondences of more than 6000 vertices were obtained using an extension of the Inverse Compositional Image Alignment algorithm [ano n. d.]. Note that the synthetic image presented is the result of not only a correspondence search (and rigid and non-rigid parameters recovery, as explained in this paper) but also a texture (or color) fitting as is common in 3D Morphable Model Fitting.

6 Conclusions

In this paper, we presented a new method which addresses the problem of Structure From Motion when a model of the non-rigid motion is available, i.e. the recovery of the rigid and non-rigid parameters of an image of a linear object class given the correspondences between model vertices and image pixels. A closed form solution of this problem existed already [Bascle and Blake 1998] which is based on the low rank constraint of the problem. We call this estimate *global*. We showed in this paper, that, up to a scale factor that can be resolved using the constraints of the problem, a series of estimates of the rigid parameters can be computed. We showed that one

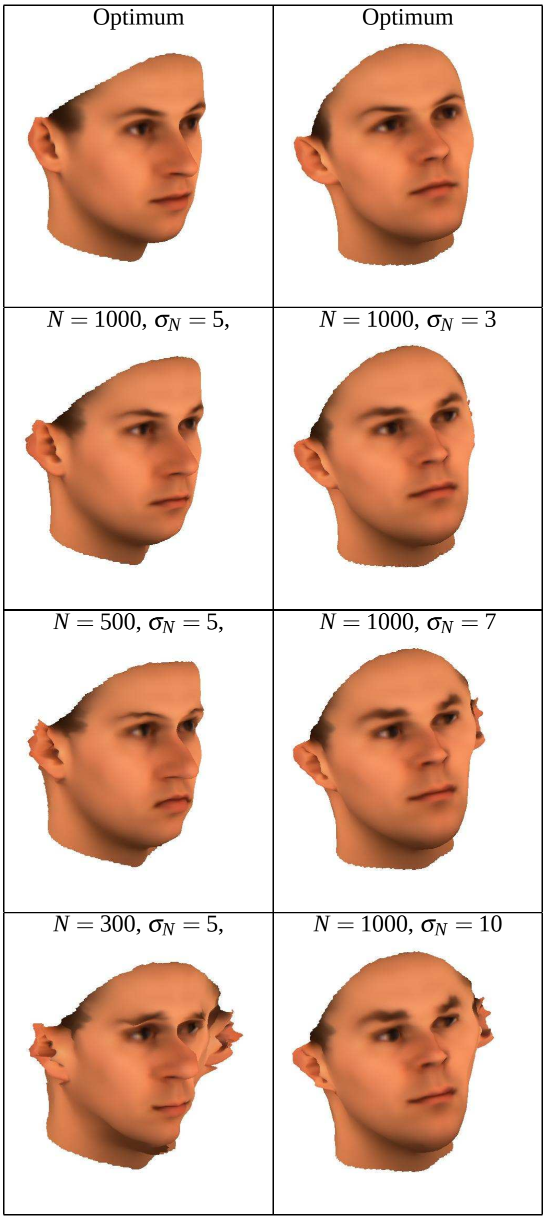
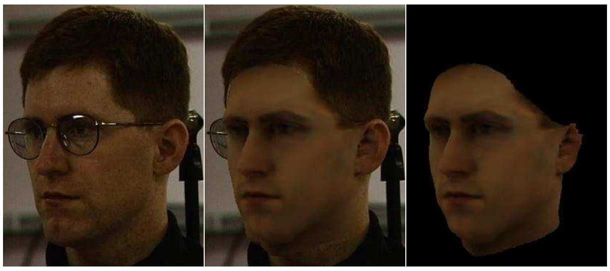

Optimum
Optimum
$N = 1000$, $\sigma_N = 5$,
$N = 1000$, $\sigma_N = 3$
$N = 500$, $\sigma_N = 5$,
$N = 1000$, $\sigma_N = 7$
$N = 300$, $\sigma_N = 5$,
$N = 1000$, $\sigma_N = 10$

Figure 5: The first row shows the synthesis of two ground truth faces and the next rows, synthesis using the parameters recovered for different number of correspondence points (first column) and level of noise (second column).



(a) Input (b) Overlaid Recov. (c) Synthetic

Figure 6: Example of one recovery on an image from the CMU-PIE data set.

of these estimates is systematically better than the other ones. Then, using this estimate, an improved estimate of the non-rigid parameters is obtained by solving a linear system of equations. We call this approach *selective*. We presented results on synthetic images and on a photograph verifying the theory. Naturally, the analytic solution provided by this algorithm might be further refined by using it as an initial estimate in an iterative minimization of Equation (5), as suggested by the Bundle Adjustment theory[Triggs et al. 1999].

A Pseudo-Inverse as a function of Kernels

Theorem 1. *Partitioning an $m \times n$, $m > n$, matrix $\mathbf{S}$ horizontally into $\begin{pmatrix} \mathbf{S_1}_{m\times k} & \mathbf{S_2}_{m\times n-k} \end{pmatrix}$ and defining by $\mathbf{K}_1$ and $\mathbf{K}_2$ the kernels of $\mathbf{S}_2$ and $\mathbf{S}_1$, respectively, (i.e. $\mathbf{K}_1 \cdot \mathbf{S}_2 \doteq \mathbf{0}$ and $\mathbf{K}_2 \cdot \mathbf{S}_1 \doteq \mathbf{0}$) the pseudo-inverse of $\mathbf{S}$ is:*

$$\mathbf{S}^+ = \begin{pmatrix} (\mathbf{K}_1 \cdot \mathbf{S}_1)^+ \cdot \mathbf{K}_1 \\ (\mathbf{K}_2 \cdot \mathbf{S}_2)^+ \cdot \mathbf{K}_2 \end{pmatrix} \doteq \mathbf{B} \tag{31}$$

Proof. The unique pseudo-inverse of a matrix satisfies the four Moore-Penrose conditions (see for example [Golub and van Loan 1996] p. 257). So we need to prove that $\mathbf{B}$ satisfies these four conditions. As we have:

$$\mathbf{B} \cdot \mathbf{S} = \begin{pmatrix} \mathbf{I}_k & \mathbf{0}_{k\times n-k} \\ \mathbf{0}_{n-k\times k} & \mathbf{I}_{n-k} \end{pmatrix} = \mathbf{I}_n \tag{32}$$

the first two and the last conditions are trivial: (i) $\mathbf{S} \cdot \mathbf{B} \cdot \mathbf{S} = \mathbf{B}$, (ii) $\mathbf{B} \cdot \mathbf{S} \cdot \mathbf{B} = \mathbf{B}$ and (iv) $(\mathbf{B} \cdot \mathbf{S})^{\mathrm{T}} = \mathbf{B} \cdot \mathbf{S}$. Proving the condition three, that $\mathbf{S} \cdot \mathbf{B}$ is symmetric is more involved. Let us define $\mathbf{K}_{m-n\times m}$ as the kernel of $\mathbf{S}$, $\mathbf{K} \cdot \mathbf{S} = \mathbf{0}$. Now let us take for granted that $\mathbf{B} \cdot \mathbf{K}^{\mathrm{T}} = \mathbf{0}$, we will prove it hereafter. Then

$$\begin{pmatrix} \mathbf{B} \\ \mathbf{K} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{S} & \mathbf{K}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times m-n} \\ \mathbf{0}_{m-n\times n} & \mathbf{I}_{m-n} \end{pmatrix} = \mathbf{I}_m \tag{33}$$

Hence,

$$\begin{pmatrix} \mathbf{B} \\ \mathbf{K} \end{pmatrix} = \begin{pmatrix} \mathbf{S} & \mathbf{K}^{\mathrm{T}} \end{pmatrix}^{-1} \quad \text{and} \quad \begin{pmatrix} \mathbf{S} & \mathbf{K}^{\mathrm{T}} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{B} \\ \mathbf{K} \end{pmatrix} = \mathbf{I}_m \tag{34}$$

It follows that $\mathbf{S} \cdot \mathbf{B} + \mathbf{K}^{\mathrm{T}}\mathbf{K} = \mathbf{I}_m$, therefore $\mathbf{S} \cdot \mathbf{B}$ is symmetric, which proves condition (iii). As the four Moore-Penrose conditions are verified by $\mathbf{B}$, $\mathbf{B}$ is equal to the unique pseudo-inverse. To conclude the proof, it remains to show that $\mathbf{B} \cdot \mathbf{K}^{\mathrm{T}} = \mathbf{0}$.

The row-space of $\mathbf{K}$ is embedded into the row-space of $\mathbf{K}_1$ and $\mathbf{K}_2$, i.e. there exists matrices $\mathbf{F}$ and $\mathbf{G}$ such that:

$$\mathbf{K} = \mathbf{F} \cdot \mathbf{K}_1 = \mathbf{G} \cdot \mathbf{K}_2 \tag{35}$$

Then, solving for $\mathbf{F}$ and $\mathbf{G}$ and plugging the solutions back into Equation (35) yields:

$$\mathbf{K} = \mathbf{K} \cdot \mathbf{K}_1^{\mathrm{T}} \cdot \mathbf{K}_1 = \mathbf{K} \cdot \mathbf{K}_2^{\mathrm{T}} \cdot \mathbf{K}_2 \tag{36}$$

Post-multiplying these expressions by $\mathbf{S}$ gives:

$$\mathbf{K} \cdot \mathbf{S} = \mathbf{K} \cdot \mathbf{K}_1^{\mathrm{T}} \cdot \mathbf{K}_1 \cdot \mathbf{S} = \mathbf{K} \cdot \mathbf{K}_2^{\mathrm{T}} \cdot \mathbf{K}_2 \cdot \mathbf{S} = \mathbf{0} \tag{37}$$

Hence,

$$\mathbf{K} \cdot \mathbf{K}_1^{\mathrm{T}} \cdot \mathbf{K}_1 \cdot \mathbf{S}_1 = \mathbf{K} \cdot \mathbf{K}_2^{\mathrm{T}} \cdot \mathbf{K}_2 \cdot \mathbf{S}_2 = \mathbf{0} \tag{38}$$

Then, utilizing the fact that $(\mathbf{K}_1 \cdot \mathbf{S}_1)^+ = \left(\mathbf{S}_1^{\mathrm{T}} \cdot \mathbf{K}_1^{\mathrm{T}} \cdot \mathbf{K}_1 \cdot \mathbf{S}_1\right)^{-1} \cdot \mathbf{S}_1^{\mathrm{T}} \cdot \mathbf{K}_1^{\mathrm{T}}$, it follows that $\mathbf{K} \cdot \mathbf{B}^{\mathrm{T}} = \mathbf{0}$, which was to be proved. □

It is trivial to extend this proof to multiple partitions.

References

ANDA, A. A. 1995. *Self-Scaling Fast Plane Rotation Algorithms*. PhD thesis, University of Minnesota.

Anonymous submission.

BAKER, S., AND MATTHEWS, I. 2001. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*.

BASCLE, B., AND BLAKE, A. 1998. Separability of pose and expression in facial tracking and animation. In *Sixth International Conference on Computer Vision*, 323–328.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3D-faces. In *SIGGRAPH 99 Conference Proceedings*, Addison Wesley, Los Angeles.

BRAND, M., AND BHOTIKA, R. 2001. Flexible flow for 3d nonrigid tracking and shape recovery. In *EEE Computer Vision and Pattern Recognition*.

BRAND, M. 2001. Morphable 3d models from video. In *IEEE Computer Vision and Pattern Recognition (CVPR), December*.

BREGLER, C., HERTZMANN, A., AND BIERMANN, H. 2000. Recovering non-rigid 3d shape from image streams. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

GOLUB, G. H., AND VAN LOAN, C. F. 1996. *Matrix Computations*. Johns Hopkins.

IRANI, M., AND ANANDAN, P. 2000. Factorization with uncertainty. In *European Conference on Computer Vision (ECCV)*.

MAGNUS, J., AND NEUDECKER, H. 1999. *Matrix Differential Calculus*. Wiley.

MINKA, T. P. 2000. Old and new matrix algebra useful for statistics. http://www.stat.cmu.edu/˜minka/papers/matrix.html, December.

SUN, Z., RAMESH, V., AND TEKALP, M. 2001. Error characterization of the factorization method. *Computer Vision and Image Understanding 82*, 110–137.

TOMASI, C., AND KANADE, T. 1991. Factoring image sequences into shape and motion. In *Proceedings of the IEEE Workshop on Visual Motion*, IEEE Comput. Soc. Press Los Alamitos, CA, USA, IEEE, 21–28.

TORRESANI, L., YANG, D. B., ALEXANDER, E. J., AND BREGLER, C. 2001. Tracking and modeling non-rigid objects with rank constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*.

TRIGGS, B., MCLAUCHLAN, P., HARTLEY, R., AND FITZGIBBON, A. 1999. Bundle adjustment: A modern synthesis. In *Vision Algorithms: Theory and Practice*.

VETTER, T., AND POGGIO, T. 1997. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 7, 733–742.

WENG, J., HUANG, T., AND AHUJA, N. 1989. Motion and structure from two perspective views: Algorithms, error analysis and error estimation. *PAMI 11*, 451–476.