A Closed-Form Solution to Non-Rigid Shape and Motion Recovery

Jing Xiao, Jin-xiang Chai, and Takeo Kanade

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{jxiao, jchai, tk}@cs.cmu.edu
http://www.ri.cmu.edu/people/xiao_jing.html

Abstract. Recovery of 3D shape and motion of non-static scenes from a monocular video sequence is important for applications like human computer interaction. If every point in the scene randomly deforms at each frame, it is impossible to recover the deforming shapes. In practice, many non-rigid objects, *e.g.* face under various expressions, deform regularly and their shapes are, or approximately are, weighted combination of certain shape bases. Shape and motion recovery under such situations thus has attracted many interests. Previous work on this problem [6, 4, 12] utilized only the orthonormality constraints on camera rotations (***rotation constraints***), but failed to apply another constraints on the shape bases (***basis constraints***). This paper proves that the solutions obtained using only the ***rotation constraints*** are inherently ambiguous. The ambiguity arises from the fact that, the shape bases are not unique since their linear transformation is a new set of eligible bases. To eliminate the ambiguity, we introduce the ***basis constraints*** that implicitly determine the shape bases uniquely. This paper proves that, under the weak-perspective projection model, once both the basis and the rotation constraints are imposed, we achieve a closed-form solution to the problem of non-rigid shape and motion recovery. The accuracy and robustness of our closed-form solution is evaluated quantitatively on synthetic data and qualitatively on real video sequences.

1 Introduction

The many years of work in structure from motion have led to significant successes in recovery of 3D shapes and motion estimates from 2D monocular videos to support modeling, rendering, visualization, and compression. Reliable systems exist for reconstructing the 3D geometry of static scenes. However, in real world, most biological objects and natural scenes are flexible and often dynamic: faces carrying expressions, fingers bending, etc. Recovering the structure and motion of these non-rigid objects from a single-camera video stream is a challenging task. The effects of 3D rigid motion, *i.e.* camera rotation and translation, and non-rigid motion, like deforming and stretching, are coupled together in image measurement. If every point on the objects deforms arbitrarily, it is impossible to reconstruct their shapes. In practice, many non-rigid objects, *e.g.* face under various expressions and scene consisting of static building and moving vehicles, deform regularly. Under such situations, the problem of shape and motion recovery is solvable.

One way to solve the problem is to use the application-specific models of non-rigid structure to constrain the deformation [2, 3, 5, 8]. These methods model the non-rigid object shapes as weighted combinations of certain shape bases. For instance, the geometry of a face is represented as a weighted combination of shape bases that correspond to various facial deformations. The successes of these approaches suggest the advantage of basis representation of non-rigid shapes. However, such models are usually unknown and complicated. An offline training step is thus required to learn these models. In many applications, *e.g.* reconstruction of a scene consisting of a moving car and a static building, the models of the dynamic structure are often expensive or difficult to obtain.

Several approaches [6, 12, 4] were proposed to solve the problem from another direction. These methods do not require a prior model. Instead, they treat the model, *i.e.* shape bases, as part of the unknowns to be solved. The goal of these approaches is to recover not only the non-rigid shape and motion, but also the shape model. They utilize only the orthonormality constraints on camera rotations (***rotation constraints***) to solve the problem. However, this paper proves that, enforcing only the rotation constraints leads to ambiguous and invalid solutions. Previous approaches thus cannot guarantee the desired solution. They have to either require a prior knowledge on shape and motion, *e.g.* constant speed [9], or need non-linear optimization that involves large number of variables and hence requires good initial estimate [12, 4].

Intuitively, the ambiguity of the solution obtained using only the rotation constraints arises from the non-uniqueness of the shape bases: a linear transformation of a set of shape bases is a new set of eligible bases. Once the bases are determined uniquely, the ambiguity is eliminated. Therefore, instead of imposing only the rotation constraints, we identify and introduce another set of constraints on the shape bases (***basis constraints***), which implicitly determine the bases uniquely. This paper proves that, under the weak-perspective projection model, when both the basis and rotation constraints are imposed, a closed-form solution to the problem of non-rigid shape and motion recovery is achieved. Accordingly we propose a factorization method that applies both metric constraints to compute the closed-form solution for the non-rigid shape, motion, and shape bases.

1.1 Previous Work

Recovering 3D object structure and motion from 2D image sequences has a rich history. Various approaches have been proposed for different applications. The discussion in this section will focus on the factorization techniques, which are closely related to our work.

The factorization method was first proposed by Tomasi and Kanade [11]. First it applies the rank constraint to factorize a set of feature locations tracked across the entire sequence.

Then it uses the orthonormality constraints on the rotation matrices to recover the scene structure and camera rotations in one step. This approach works under the orthographic projection model. Poelman and Kanade [10] extended it to work under the weak perspective and para-perspective projection models. Triggs [13] generalized the factorization method to the recovery of scene geometry and camera motion under the perspective projection model. These methods work only for static scenes.

For non-static scenes, Costeira and Kanade [7] proposed a factorization technique to recover the camera motion and shapes of multiple independently moving objects under the orthographic projection model. This technique factorizes the feature locations to compute a shape interaction matrix, then block-diagonalizes this matrix to segment different objects and recover their shapes and motions. Han and Kanade [9] introduced another factorization method to reconstruct a scene consisting of multiple objects, some of them static and the others moving along fixed directions and at **constant** speed. Wolf and Shashua [14] presented a more generalized solution to reconstructing the shapes that deform at **constant** velocity.

Bregler et al. [6] first introduced the shape representation as weighted combination of bases to reconstruct non-rigid shapes and motion. Without assuming constant deformation speed, they proposed the sub-block re-ordering and factorization method to determine the shape bases, combination coefficients of the bases, and camera rotations simultaneously. This approach enforces only the rotation constraints. As we proved, the solution is inherently ambiguous and not optimal. To remedy the problem, Torresani and his colleagues [12] extended Bregler's method to a trilinear optimization approach. At each step, two of the three types of unknowns, bases, coefficients, and rotations, are fixed and the rest one is updated. Bregler's method is used to initialize the optimization process. Brand [4] proposed a similar non-linear optimization method that uses an extension of Bregler's method for initialization. Both the non-linear optimization approaches still fail to impose the basis constraints, which is the essential reason that the method in [6] does not work well. Therefore they can neither guarantee the optimal solution. Note that both optimization processes involve a large number of variables, e.g. the number of coefficients to be computed equals the product of the number of images and the number of shape bases. Their performances greatly rely on the quality of the initial estimates of the large number of unknowns, which are not easy to achieve.

2 Problem Statement

Given 2D locations of P feature points across F frames, $\{(u,v)_{fp}^T | f = 1, ..., F, p = 1, ..., P\}$, our goal is to recover the motion of the non-rigid object relative to the camera, including rotations $\{R_f | f = 1, ..., F\}$ and translations $\{\mathbf{t}_f | f = 1, ..., F\}$, and its 3D deforming shapes $\{(x,y,z)_{fp}^T | f = 1, ..., F, p = 1, ..., P\}$.

We follow the representation of [3, 6]. The non-rigid shape is represented as weighted combination of K shape bases $\{B_i, i = 1, ..., K\}$. The bases are $3 \times P$ matrices controlling the deformation of P points. Then the 3D coordinate of the point p at the frame f is,

$$\mathbf{X}_{fp} = (x,y,z)_{fp}^T = \Sigma_{i=1}^K c_{fi} \mathbf{b}_{ip} \quad f = 1, ..., F, p = 1, ..., P \tag{1}$$

where $\mathbf{b}_{ip}$ is the p_{th} column of B_i and c_{if} is its combination coefficient at the frame f. The image coordinate of $\mathbf{X_{fp}}$ under the weak perspective projection model is,

$$\mathbf{x}_{fp} = (u,v)_{fp}^T = s_f(R_f \cdot \mathbf{X}_{fp} + \mathbf{t}_f) \tag{2}$$

where R_f stands for the first two rows of the f_{th} camera rotation and $\mathbf{t}_f = [t_{fx} t_{fy}]^T$ is its translation relative to the world origin. s_f is the nonzero scalar of the weak perspective projection.

Replacing $\mathbf{X}_{fp}$ using Eq. (1) and absorbing s_f into c_{fi} and $\mathbf{t}_f$,

$$\mathbf{x}_{fp} = \left[c_{f1}R_f \ ... \ c_{fK}R_f \right] \cdot \begin{bmatrix} \mathbf{b}_{1p} \\ ... \\ \mathbf{b}_{Kp} \end{bmatrix} + \mathbf{t}_f \tag{3}$$

Suppose the image coordinates of all P feature points across F frames are obtained. We form a $2F \times P$ *measurement matrix* W by stacking all image coordinates. Then $W = MB + T[11...1]$. where M is a $2F \times 3K$ scaled rotation matrix, B is a $3K \times P$ bases matrix, and T is a $2F \times 1$ translation vector,

$$M = \begin{pmatrix} c_{11}R_1 & ... & c_{1K}R_1 \\ \vdots & \vdots & \vdots \\ c_{F1}R_F & ... & c_{FK}R_F \end{pmatrix} \ , \ B = \begin{pmatrix} \mathbf{b}_{11} & ... & \mathbf{b}_{1P} \\ \vdots & \vdots & \vdots \\ \mathbf{b}_{K1} & ... & \mathbf{b}_{KP} \end{pmatrix} \ , \ T = \left[\mathbf{t}_1^T \ ... \ \mathbf{t}_F^T \right]^T \tag{4}$$

As in [9, 6], we position the world origin at the scene center and compute the translation vector by averaging the image projections of all points. We then subtract it from W and obtain the *registered* measurement matrix $\tilde{W} = MB$.

Since $\tilde{W}$ is the product of the $2F \times 3K$ scaled rotation matrix M and the $3K \times P$ shape bases matrix B, its rank is at most $min\{3K, 2F, P\}$. In practice, the frame number F and point number P are usually much larger than the basis number K. Thus the rank of $\tilde{W}$ is at most $3K$ and K is determined by $K = rank(\tilde{W})/3$. We then perform SVD on $\tilde{W}$ to get the best possible rank $3K$ approximation of $\tilde{W}$ as $\tilde{M}\tilde{B}$. This decomposition is only determined up to a non-singular $3K \times 3K$ linear transformation. The true scaled rotation matrix M and bases matrix B are of the form,

$$M = \tilde{M} \cdot G, \qquad B = G^{-1} \cdot \tilde{B} \tag{5}$$

where G is called the *corrective transformation* matrix. Once G is determined, M and B are obtained and thus the rotations, shape bases, and combination coefficients are recovered.

Since all the procedures above, except obtaining G, are standard and well-understood [3, 6], the problem of nonrigid shape and motion recovery is now reduced to: Given the measurement matrix W, how can we solve the *corrective transformation* matrix G?

3 Metric Constraints

In order to solve G, two types of metric constraints are available and should be imposed: ***rotation constraints*** and ***basis constraints***. Using only the rotation constraints [6, 4] leads to ambiguous solutions. Instead imposing both constraints results in a closed-form solution.

3.1 Rotation Constraints

The orthonormality constraints on the rotation matrices are one of the most powerful metric constraints and they have been used in reconstructing the shape and motion for static objects [11, 10], multiple moving objects [7, 9], and non-rigid deforming objects [6, 12, 4].

According to Eq. (5), $MM^T = \tilde{M}GG^T\tilde{M}^T$. Let us denote GG^T by Q. Then,

$$\tilde{M}_{2*i-1:2*i} Q \tilde{M}^T_{2*j-1:2*j} = \Sigma_{k=1}^{K} c_{ik}c_{jk}R_i * R_j^T \quad , i, j = 1, ...F \tag{6}$$

where $\tilde{M}_{2*i-1:2*i}$ represents the i_{th} two-row of $\tilde{M}$. Due to the orthonormality of the rotation matrices,

$$\tilde{M}_{2*i-1:2*i} Q \tilde{M}^T_{2*i-1:2*i} = \Sigma_{k=1}^{K} c_{ik}^2 \mathbf{I}_{2\times 2} \quad , i = 1, ..., F \tag{7}$$

where $\mathbf{I}_{2\times 2}$ is a 2×2 identity matrix. Since Q is symmetric, the number of unknowns in Q is $(9K^2 + 3K)/2$. Each diagonal block of MM^T yields two linear constraints on Q,

$$\tilde{M}_{2*i-1} Q \tilde{M}_{2*i-1}^T = \tilde{M}_{2*i} Q \tilde{M}_{2*i}^T \tag{8}$$

$$\tilde{M}_{2*i-1} Q \tilde{M}_{2*i}^T = 0 \tag{9}$$

For F frames, we have $2F$ linear constraints on $(9K^2 + 3K)/2$ unknowns. It appears that, when we have enough images, *i.e.* $F \geq (9K^2 + 3K)/2$, there will be enough constraints to solve Q via the standard least-square methods. However, this is not true in general. Many of these constraints are redundant. We will show later that no matter how many frames or feature points are given, the linear constraints from Eq. (8) and Eq. (9) are not sufficient to determine Q.

3.2 Why are Rotation Constraints not Sufficient?

When the scene is static or deforms at constant velocities, the rotation constraints are sufficient to solve the corrective transformation matrix G [11, 9, 14]. However, when the scene deforms at varying speed, no matter how many images are given or how many feature points are tracked, the solutions of the constraints in Eq. (8) and Eq. (9) are inherently ambiguous. The degree of freedom of the solution space is $2K^2 - K$.

Definition 1. *A $3K \times 3K$ symmetric matrix Y is called a block-skew-symmetric matrix, if all the diagonal 3×3 blocks are zero matrices and each off-diagonal 3×3 block is a skew symmetric matrix.*

$$Y_{ij} = \begin{pmatrix} 0 & y_{ij1} & y_{ij2} \\ -y_{ij1} & 0 & y_{ij3} \\ -y_{ij2} & -y_{ij3} & 0 \end{pmatrix} = -Y_{ij}^T = Y_{ji}^T, \quad i \neq j \tag{10}$$

$$Y_{ii} = 0_{3\times 3}, \qquad i, j = 1, ..., K \tag{11}$$

Each off-diagonal block consists of 3 independent elements. Since Y is symmetric and has $K(K-1)/2$ independent off-diagonal blocks, it totally includes $3K(K-1)/2$ independent elements.

Definition 2. *A $3K \times 3K$ symmetric matrix Z is called a block-scaled-identity matrix, if each 3×3 block is a scaled identity matrix, i.e. $Z_{ij} = \lambda_{ij} \mathbf{I}_{3\times 3}$, where λ_{ij} is the only variable.*

since Z is symmetric, the total number of variables in Z equals the number of independent blocks, $K(K+1)/2$.

Theorem 1. *Let H be the summation of Y and Z. $Q = GHG^T$ is the general solution of the rotation constraints in Eq. (8) and Eq. (9), where G is the desired corrective transformation matrix.*

Proof. Since G is a non-singular matrix, the solution Q of Eq. (8) and Eq. (9) can be represented as $Q = G\Lambda G^T$. Now we need to prove that Λ must be in the form of H, *i.e.* the summation of Y and Z.

According to Eq. (7),

$$\begin{aligned} \tilde{M}_{2*i-1:2*i} Q \tilde{M}_{2*i-1:2*i}^T &= M_{2*i-1:2*i} \Lambda M_{2*i-1:2*i}^T \\ &= \alpha_i \mathbf{I}_{2\times 2} \quad , i = 1, ..., F \end{aligned} \tag{12}$$

where α_i is an unknown scalar. Divide Λ into 3×3 blocks, Λ_{kj} (k,j=1,...,K). Combining Eq. (4) and Eq. (12),

$$R_i \Sigma_{k=1}^K (c_{ik}^2 \Lambda_{kk} + \Sigma_{j=k+1}^K 2 c_{ik} c_{ij} (\Lambda_{kj} + \Lambda_{kj}^T)) R_i^T = \alpha_i \mathbf{I}_{2\times 2} \quad , i = 1, ..., F \tag{13}$$

Denote the 3×3 symmetric matrix $\Sigma_{k=1}^{K}(c_{ik}^2 \Lambda_{kk} + \Sigma_{j=k+1}^{K} 2c_{ik}c_{ij}(\Lambda_{kj} + \Lambda_{kj}^T))$ by Γ_i. Let $\tilde{\Gamma}_i$ be the homogeneous solution of Eq. (13), *i.e.* $R_i \tilde{\Gamma}_i R_i^T = \mathbf{0}_{2\times 2}$. Note that R_i consists of only the first two rows of the i_{th} rotation matrix. Let r_{i3} denote the third row. Due to the orthonormality constraints, $\tilde{\Gamma}_i$ is determined by,

$$\Gamma_i R_i^T = [r_{i3}^T \;\; \delta_i r_{i3}^T] \tag{14}$$

where δ_i is an arbitrary scalar. Apparently $\Gamma_i = \alpha_i \mathbf{I}_{3\times 3}$ is a particular solution of Eq. (13). Therefore the general solution of Eq. (13) is,

$$\Gamma_i = \Sigma_{k=1}^{K}(c_{ik}^2 \Lambda_{kk} + \Sigma_{j=k+1}^{K} 2c_{ik}c_{ij}(\Lambda_{kj} + \Lambda_{kj}^T)) = \alpha_i \mathbf{I}_{3\times 3} + \beta_i \tilde{\Gamma}_i \tag{15}$$

where β_i is an arbitrary scalar. Since $Q = G\Lambda G^T$ is the general solution of the rotation constraints, Eq. (13) and Eq. (15) must be satisfied for any set of coefficients and rotations. If β_i for some frame i is not zero, for another frame that is formed by the same coefficients but different rotation, Eq. (15) and Eq. (14) are not satisfied. Therefore, β_i has to be zero for every frame, *i.e.*,

$$\Sigma_{k=1}^{K}(c_{ik}^2 \Lambda_{kk} + \Sigma_{j=k+1}^{K} 2c_{ik}c_{ij}(\Lambda_{kj} + \Lambda_{kj}^T)) = \alpha_i \mathbf{I}_{3\times 3} \tag{16}$$

Since Eq. (16) must be satisfied for any set of coefficients, the solution is,

$$\Lambda_{kk} = \lambda_{kk} \mathbf{I}_{3\times 3} \tag{17}$$

$$\Lambda_{kj} + \Lambda_{kj}^T = \lambda_{kj} \mathbf{I}_{3\times 3} \quad , k = 1, ..., K; \;\; j = k+1, ..., K \tag{18}$$

where λ_{kk} and λ_{kj} are arbitrary scalars. According to Eq. (17), the diagonal block Λ_{kk} is a scaled identity matrix. Since the diagonal block of Z, Z_{kk}, is a scaled identity matrix and the diagonal block of Y, Y_{kk}, is a zero matrix, $\Lambda_{kk} = Z_{kk} + Y_{kk}$. Let Λ_{kjab}, $a, b = \{1, 2, 3\}$, denote the elements of an off-diagonal block Λ_{kj}. Due to Eq. (18), the diagonal elements are $\Lambda_{kj11} = \Lambda_{kj22} = \Lambda_{kj33} = \lambda_{kj}/2$ and the off-diagonal elements satisfy $\Lambda_{kj12} = -\Lambda_{kj21}$, $\Lambda_{kj13} = -\Lambda_{kj31}$, and $\Lambda_{kj23} = -\Lambda_{kj32}$. Therefore Λ_{kj} equals the summation of a scaled identity block, Z_{kj}, and a skew-symmetric block, Y_{kj}. This concludes the proof: Λ equals H, the summation of a block-skew-symmetric matrix Y and a block-scaled-identity matrix Z, *i.e.* the general solution of the rotation constraints is $Q = GHG^T$.
$\diamondsuit$

Since H consists of $2K^2 - K$ independent elements: $3K(K-1)/2$ from Y and $K(K+1)/2$ from Z, the solution space has a degree of freedom of $2K^2 - K$. Now the question is: is every solution in the space a valid solution of Q? If so, even if the ambiguity exists, one can compute an arbitrary solution in the space to solve the problem. However, it is not the case. The space composed of two components, Y and Z, contains both valid and invalid solutions. Specifically, the solutions consisting of only Z, $\tilde{Q}_Z = GZG^T$, are valid solutions. The variety of $\tilde{Q}_Z$ refers to different linear transformations of the shape bases and any of $\tilde{Q}_Z$ can be used to recover the rotations and other unknowns. The solutions involving Y, $\tilde{Q}_Y = GYG^T$ or $G(Y+Z)G^T$, are invalid solutions. Since a valid solution $Q = GG^T$ must be positive semi-definite and a block-skew-symmetric matrix Y is not positive semi-definite, $\tilde{Q}_Y$ are invalid solutions.

3.3 Basis Constraints

For static scenes, a variety of approaches [11, 10, 13] utilize only the rotation constraints and succeed in determining the correct solution. Now we are dealing with non-static scenes with a certain assumption of the non-rigidity, *i.e.* representable by direct combination of the shape bases. Under such situations, enforcing only the rotation constraints results in a solution space that contains ambiguous and invalid solutions. Are there other constraints that we can use

to determine the desired solution in the space? Intuitively, since the only difference under non-rigid situations from under rigid situations is that the non-rigid shape deforms as direct combination of a certain number of shape bases, can we impose certain constraints on the bases and eliminate the ambiguity?

Since any non-singular linear transformation on the shape bases yields a new set of eligible bases, the bases and the corresponding combination coefficients are not unique. However, their composition, *i.e.* the non-rigid shapes, are unique. Thus the bases and coefficients depend on each other. Once one of them is determined, another is also decided. If we can obtain any K frames including independent shapes and treat the shapes as a set of bases, both the bases and coefficients are determined uniquely. Without the loss of generality, we assume the shapes in the first K frames are independent on each other[1]. The K shapes are then treated as the bases. This step determines the first K frames of coefficients as,

$$\begin{aligned} &c_{ii} = 1,\ \ i = 1, ..., K \\ &c_{ij} = 0,\ \ i \neq j,\ i = 1, ..., K,\ j = 1, ..., K \end{aligned} \tag{19}$$

For any three-column of G, $g_k, k = 1, ..., K$, according to Eq. (5),

$$\tilde{M} g_k = \begin{pmatrix} c_{1k} R_1 \\ ... \\ c_{Fk} R_F \end{pmatrix} \quad k = 1, ..., K \tag{20}$$

We denote $g_k {g_k}^T$ by Q_k. Then,

$$\tilde{M}_{2*i-1:2*i} Q_k \tilde{M}^T_{2*j-1:2*j} = c_{ik} c_{jk} R_i R_j^T \tag{21}$$

Combining Eq. (19) and Eq. (21), we obtain another $4(K-1)F$ basis constraints on Q_k:

$$\tilde{M}_{2*i-1} Q_k \tilde{M}^T_{2*j-1} = \begin{cases} 1,\ i = j = k \\ 0,\ (i,j) \in \omega_1 \end{cases} \tag{22}$$

$$\tilde{M}_{2*i} Q_k \tilde{M}^T_{2*j} = \begin{cases} 1,\ i = j = k \\ 0,\ (i,j) \in \omega_1 \end{cases} \tag{23}$$

$$\tilde{M}_{2i-1} Q_k \tilde{M}^T_{2*j} = 0, \quad (i,j) \in \omega_1 \ or\ i = j = k \tag{24}$$

$$\tilde{M}_{2i} Q_k \tilde{M}^T_{2*j-1} = 0, \quad (i,j) \in \omega_1 \ or\ i = j = k \tag{25}$$

where $\omega_1 = \{(i,j) | i = 1, ..., K,\ j = 1, ..., F\ \ and\ \ i \neq k\}$. The basis constraints eliminate the ambiguity of the rotation constraints and determine a closed-form solution to Q_k.

4 A Closed-Form Solution

Section 3.2 proves that the general solution of the rotation constraints is $GHG^T = GYG^T + GZG^T$, where G is the desired corrective transformation matrix, Y is a block-skew-symmetric matrix, and Z is a block-scaled-identity matrix. The solutions have a degree of freedom of $2K^2 - K$. This section will prove that enforcing the basis constraints eliminates the ambiguity and determines a closed-form solution.

By definition, each 3×3 block H_{ij} $(i, j = 1, ..., K)$ of H is composed of four independent entries,

$$H_{ij} = \begin{pmatrix} h_1 & h_2 & h_3 \\ -h_2 & h_1 & h_4 \\ -h_3 & -h_4 & h_1 \end{pmatrix} \tag{26}$$

[1] If the first K shapes are not independent, we can find K frames in which the shapes are independent, by examining the singular values of their image projections. We then reorder the sequence by moving these K frames to the top.

Lemma 1 *H_{ij} is a zero matrix if,*

$$R_i H_{ij} R_j^T = \begin{pmatrix} r_{i1} \\ r_{i2} \end{pmatrix} H_{ij} \begin{pmatrix} {r_{j1}}^T & r_{j2}^T \end{pmatrix} = \mathbf{0}_{2\times 2} \tag{27}$$

Proof. First we prove that the rank of H_{ij} is less than 3. Due to Eq. (27) and the orthonormality constraints,

$$H_{ij} \begin{pmatrix} {r_{j1}}^T & r_{j2}^T \end{pmatrix} = \begin{pmatrix} \alpha_1 {r_{i3}}^T & \alpha_2 r_{i3}^T \end{pmatrix} \tag{28}$$

where $r_{i3} = r_{i1} \times r_{i2}$. α_1 and α_2 are two arbitrary scalars. Therefore,

- If both α_1 and α_2 are not equal to 0, the linear system $H_{ij}\mathbf{x} = r_{i3}^T$ has at least two independent solutions r_{j1}^T/α_1 and r_{j2}^T/α_2. Hence H_{ij} is not a non-singular matrix and its rank is less than its dimension, 3.
- If either α_1 or α_2 equals 0, say α_1, the linear system $H_{ij}\mathbf{x} = \mathbf{0}_{3\times 1}$ has at least one non-zero solution r_{j1}^T. H_{ij} is thus singular and its rank is less than 3.

Next, we prove $h_1 = 0$. Since the rank of H_{ij} is less than its dimension, 3, its determinant, $h_1(\sum_{i=1}^4 {h_i}^2)$, equals 0. Therefore h_1 must be 0 and H_{ij} is a skew-symmetric matrix.

Finally, we prove $h_2 = h_3 = h_4 = 0$. Since $h_1 = 0$, we rewrite Eq. (27) as follows:

$$\begin{pmatrix} r_{i1} \cdot (\mathbf{h} \times r_{j1}) & r_{i1} \cdot (\mathbf{h} \times r_{j2}) \\ r_{i2} \cdot (\mathbf{h} \times r_{j1}) & r_{i2} \cdot (\mathbf{h} \times r_{j2}) \end{pmatrix} = \mathbf{0}_{2\times 2} \tag{29}$$

where $\mathbf{h} = (-h_4\, h_3 - h_2)$. Eq. (29) means that the vector $\mathbf{h}$ is located in the intersection of the four planes determined by $(r_{i1}, r_{j1}), (r_{i1}, r_{j2}), (r_{i2}, r_{j1})$, and (r_{i2}, r_{j2}). Under non-degenerate situations, r_{i1}, r_{i2}, r_{j1}, and r_{j2} do not lie in the same plane, hence the four planes intersect at the origin, *i.e.* $\mathbf{h} = (-h_4 \;\; h_3 \;\; -h_2) = \mathbf{0}_{1\times 3}$. This proves that H_{ij} is a zero matrix.
$\diamondsuit$

Due to Lemma 1, we derive the following theorem,

Theorem 2. *If Q_k satisfies both basis constraints and rotation constraints, Q_k equals ${g_k}{g_k}^T$, where g_k is the k_{th} three-column of G.*

Proof. Since Q_k satisfies the rotation constraints, $Q_k = GHG^T$ and $\tilde{M}Q_k\tilde{M}^T = MHM^T$. Thus,

$$M_{2*i-1:2*i} H M_{2*j-1:2*j}^T = \Sigma_{k_1=1}^K \Sigma_{k_2=1}^K c_{ik_1} c_{jk_2} R_i H_{k_1k_2} R_j^T, \quad i,j = 1,...,F \tag{30}$$

According to Eq. (19),

$$M_{2*i-1:2*i} H M_{2*j-1:2*j}^T = R_i H_{ij} {R_j}^T, \quad i,j = 1,...,K \tag{31}$$

Due to the basis constraints in Eq. (22) to (25),

$$R_k H_{kk} {R_k}^T = \mathbf{I}_{2\times 2} \tag{32}$$

$$R_i H_{ij} {R_j}^T = \mathbf{0}_{2\times 2}, \quad i,j = 1,...,K, \textit{ and } i \neq k,\ j \neq k \tag{33}$$

By definition, $H_{kk} = \lambda_{kk}\mathbf{I}_{3\times 3}$, where λ_{kk} is a scalar. Due to Eq. (32), $\lambda_{kk} = 1$ and $H_{kk} = \mathbf{I}_{3\times 3}$. From Lemma 1 and Eq. (33), H_{ij} is a zero matrix when $i,j = 1,...,K$, and $i \neq k$, $j \neq k$. Therefore $Q_k = GHG^T = [g_1,...,g_K]H[g_1,...,g_K]^T = [0,...,0,g_k,0,...0][g_1,...,g_K]^T = g_k g_k^T$.
$\diamondsuit$

According to Theorem 2, enforcing both rotation constraints and basis constraints leads to a linear closed-form solution of $Q_k = g_k g_k^T$, $k = 1,...,K$. Then g_k, $k = 1,...,K$ can be recovered via SVD. We project them to the common coordinate system and determine the corrective transformation $G = [g_1,...,g_K]$. According to Eq. (5), we recover the shape bases $B = G^{-1}\tilde{B}$, the scaled rotation matrix $M = \tilde{M}G$, and thus the rotations and coefficients.

5 Performance Evaluation

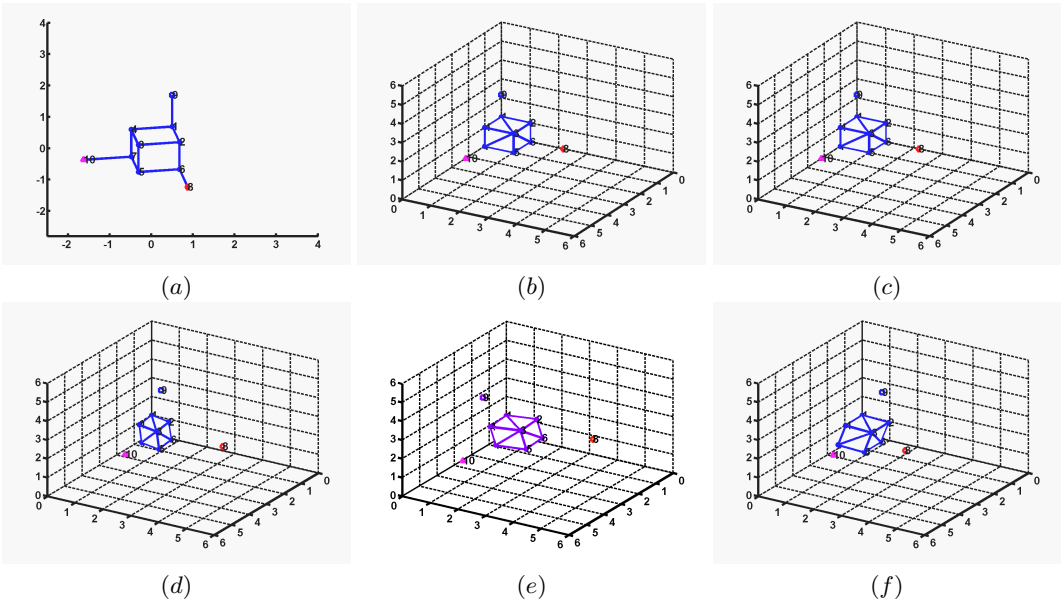


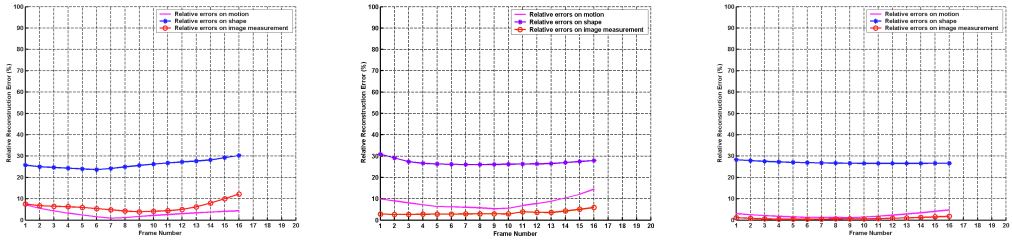
Fig. 1. A static cube and 3 points moving along fixed directions: (a) the input image; (b) the ground truth 3D shape; (c) reconstruction of the closed-form solution; (d) reconstruction of Bregler's method; (e) reconstruction of Brand's method; (f) reconstruction of the tri-linear method.

The performance of the closed-form solution is evaluated in a number of experiments. First, we compare its performance with that of previous work. Second, we evaluate its robustness and accuracy quantitatively on synthetic data. Third, we apply it on real image sequences to examine it qualitatively.

5.1 Comparison with Previous Work

Previous methods enforce only the rotation constraints and thus have limitations. [6] reorders and factorizes each two-row of $\tilde{M}$ to compute the coefficients and rotations. Then the rotation constraints are applied to compute a 3×3 corrective transformation G_s as in [11]. This process is equivalent to assume the desired G as $diag(G_s, \ldots, G_s)$. Whereas this assumption is correct for static scenes, it does not hold when the scene is non-rigid. Brand [4] extended [6] by applying the rotation constraints to compute different corrective transformations for each three-column of $\tilde{M}$ independently. It is equivalent to assume G as $diag(G_{s1}, \ldots, G_{sK})$, where the diagonal blocks are different. This assumption often does not hold, because $\tilde{M}$ can be from an arbitrary linear transformation of the true M and its three-columns usually are mixed up. The regularization term to minimize the deformation bases will not help much, since one can have arbitrarily small bases but large coefficients and achieve the same reconstruction. The tri-linear algorithm [12] does not assume certain form of G, but involves a large number of unknowns, *e.g.* the number of coefficients is FK. It enforces only the rotation constraints and there exist many local optima. Its performance depends on good quality of the initial estimate, which is not easy to achieve, especially for such a huge number of unknowns.

Let us demonstrate that the weakness of the above approaches actually results in erroneous solutions even for a simple noiseless example. Figure 1 shows a scene consisting of a static cube and 3 moving points, marked as diamonds, triangles, and squares. The measurement



(a) (b) (c)

Fig. 2. Reconstruction errors: (a) Bregler's method; (b) Brand's method; (c) the tri-linear method.

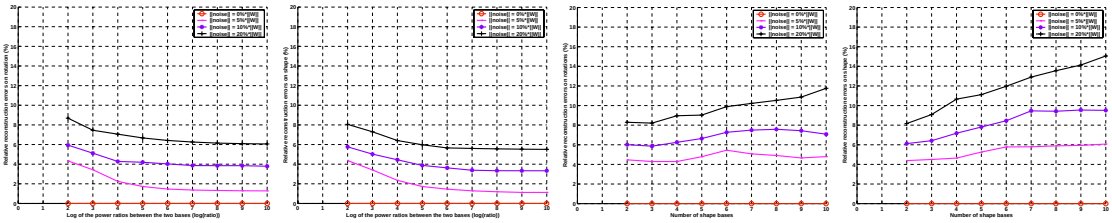
consists of 10 points: 7 visible vertices of the cube and 3 moving points. The 3 points move along the three axes respectively at non-constant speed. The scene structure is composed of $K = 2$ shape bases, one for the static cube and another for the moving points. Their image projections across 16 frames from different views are given. One of them is shown in Figure 1.(a). The corresponding ground truth scene shape is demonstrated in Figure 1.(b). Figure 1.(c) to 1.(f) show the reconstructed scene structures using the closed-form solution, Bregler's method [6], Brand's method [4] and the tri-linear method [12] both after 4000 iterations. While our closed-form solution achieves the exact reconstruction, all three previous methods result in apparent reconstruction errors, even for such a simple and noiseless setting. Figure 2 demonstrates the reconstruction errors of the previous work on rotations, shapes, and image measurements. The errors are computed relative to the ground truth.

5.2 Quantitative Evaluation on Synthetic Data

Our approach is quantitatively evaluated on the synthetic data. We evaluate the accuracy and robustness on three factors: deformation strength, number of shape bases, and noise level. The deformation strength shows how close to rigid the shape is and it is represented by the ratios of the powers (Frobenius Norm) of the bases. Larger ratio means weaker deformation, *i.e.* the shape is closer to rigid. The number of shape bases represents how flexible the shape is. Bigger basis number means more control variables on the shape need to solve for. under the noiseless situations, a good approach should provide the exact solution, no matter how strong the deformation is and how big the basis number is.

In real applications, the data are often contaminated by noise. Under such situations, a good method should be robust enough to provide reasonably accurate solutions, regardless of strong deformation or big basis number. Assuming a Gaussian white noise, we represent the noise strength level by the ratio between the standard deviation and the power of the measurement $\tilde{W}$. Under the same noise level, weaker deformation leads to better performance, since some deformation mode is more dominant and the noise relative to the dominant basis is weaker. When the powers of the bases are close to each other, bigger basis number results in poorer performance, because the noise relative to each individual basis is stronger.

Figure 3.(a) and (b) show the performance of our closed-form solution under various deformation strength and noise levels. Two bases are used. The ratios between their powers are 2^0, 2^1, ..., and 2^8. Four levels of Gaussian white noise are imposed on $\tilde{W}$. Their standard deviations are 0%, 5%, 10%, and 20% of the power of $\tilde{W}$. We test 100 trials for each setting and compute the average reconstruction errors on the rotations and 3D shapes, relative to the ground truth. Figure 3.(c) and (d) show the performance of our method under different number of shape bases and noise levels. We use 2, 3, ... , and 10 shape bases respectively. The bases have equal powers and thus none of them is dominant. The same noise as in last experiment are imposed.

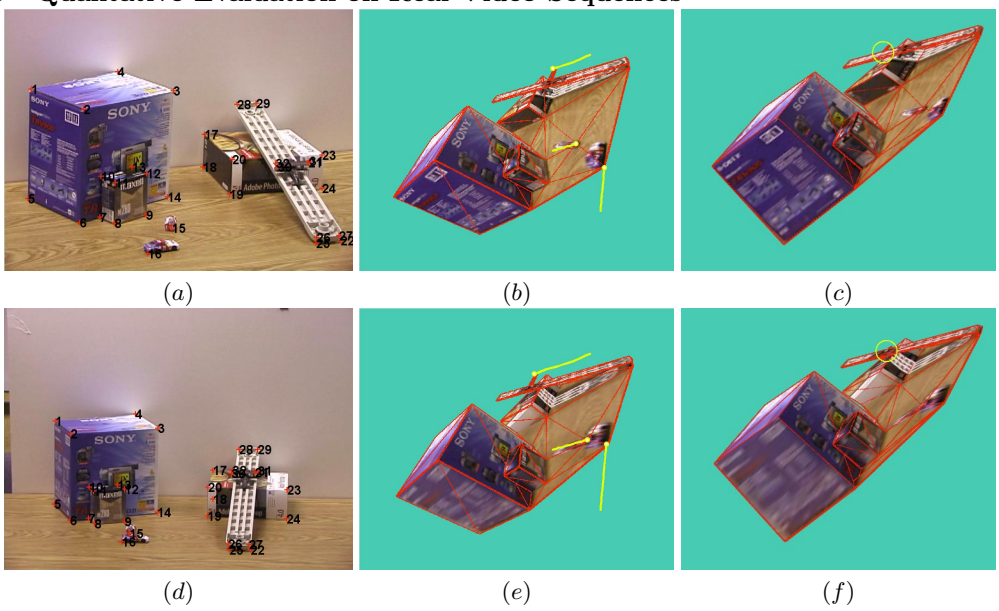


(a) (b) (c) (d)

Fig. 3. (a)&(b): Reconstruction errors on rotations and shapes under different levels of noise and deformation strength; (c)&(d): Reconstruction errors on rotations and shapes under different levels of noise and various basis numbers. Lower curve refers to weaker noise.

In both experiments, when the noise level is 0%, the closed-form solution always recovers the exact rotations and shapes. When there exists noise, it achieves reasonable accuracy, *e.g.* the maximum reconstruction error is less than 15% when the noise level is 20%. As we analyzed, under the same noise level, the performance gets better when the power ratio is larger and gets poorer when the basis number is bigger. Note that in all the experiments, the condition number of the linear system consisting of both basis constraints and rotation constraints has order of magnitude $O(10)$ to $O(10^2)$, even if the basis number is big and the deformation is strong. Our closed-form solution is thus numerically stable.

5.3 Qualitative Evaluation on Real Video Sequences



(a) (b) (c)

(d) (e) (f)

Fig. 4. Reconstruction of three moving objects in the static background. (a)&(d): two input images with marked features; (b)&(e): reconstruction by the closed-form solution; The yellow lines show the recovered moving trajectories till the present frames. (c)&(f): reconstruction by Brand's method. The yellow circle shows that the plane is mis-located.

We examined our approach qualitatively on a number of real video sequences. The first sequence was taken of an indoor scene by a handhold camera. Three objects, a car, a plane, and a toy person, moved along fixed directions and at varying speeds. The rest of the scene was static. The car and the person moved on the floor and the plane moved along a slope.

The scene structure was composed of two bases, one for the static objects and another for the moving objects. 32 feature points tracked across 18 images are used for reconstruction. Two of the images are shown in Figure 4.(a) and (d).

The rank of $\tilde{W}$ was estimated in such a way that after rank reduction at least 99% of the energy was kept. The basis number is automatically determined by $K = rank(\tilde{W})/3$. Figure 4.(b) and (e) show the images warped to a common view based on the reconstruction by the closed-form solution. The wireframes show the structure and the yellow lines show the trajectories of the moving objects till the present frames. The reconstruction is consistent with our observation, *e.g.* the plane moved linearly on top of the slope. Figure 4.(c) and (f) show the reconstruction using Brand's method [4]. The shapes of the boxes are distorted and the plane is incorrectly located underneath the slope, as shown in the yellow circles.

The second sequence was taken of a human face by a static video camera. It consisted of 236 images and contained various facial expression and head rotations. 68 feature points were manually picked in the first frame and then tracked automatically using the Active Appearance Model method [1]. Figure 5.(a) and (d) display two of the images with marked features. According to the reconstructed shapes by our method, we warp the images into a common view, as shown in Figure 5.(b) and (e). Their corresponding $3D$ wireframe models shown in Figure 5.(c) and (f) demonstrate that the non-rigid facial motions such as mouth opening and eye closure were recovered successfully. Note that the feature correspondence in these experiments was noisy, especially for those features on the sides of face. The reconstruction performance of our approach hence demonstrates its robustness to the image noise.

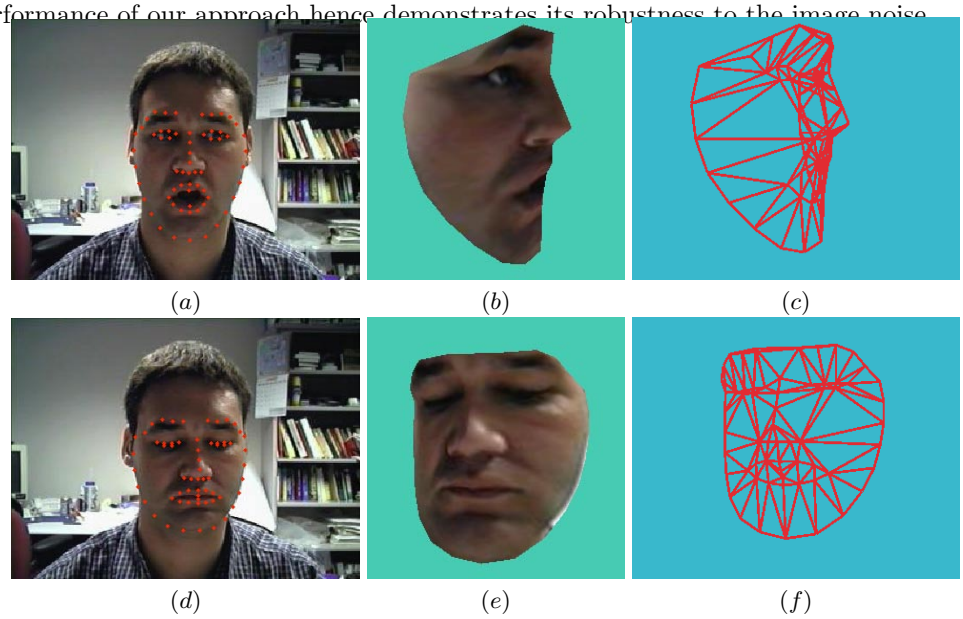


Fig. 5. Reconstruction of face shape with various expressions. (a)&(d): input images with marked features. (b)&(e): images warped to a common view based on our reconstruction. (c)&(f): The wireframe model of recovered structure. Eye closure and mouth opening are recovered.

6 Conclusion and Discussion

This paper proposes a closed-form solution to the problem of non-rigid shape and motion recovery from video, under the weak perspective projection model. It consists of three main contributions: first, we prove that enforcing only the rotation constraints results in ambiguous

and invalid solutions; second, we identify and introduce the basis constraints; Third, we prove that imposing both rotation and basis constraints leads to a closed-form solution to non-rigid shape and motion recovery.

A deformation mode is degenerate, if it limits the shape to deform in a plane, *i.e.* the rank of the corresponding basis is less than 3. Such a case occurs in practice, *e.g.* if a scene contains only one moving object that moves along a straight line, the basis referring to the linear motion is degenerate, since the motion vector is of rank 1. Under degenerate situations, the basis constraints cannot determine the degenerate bases. As a result, the ambiguity of the rotation constraints cannot be completely eliminated and thus enforcing both metric constraints is insufficient to produce a closed-form solution. The degeneracy problem can be solved using an alternating linear optimization method.

In applications such as motion capture, the acquired data are usually composition of the 3D non-rigid structures and their corresponding poses. One has to decouple the originally acquired data so as to capture the accurate 3D shapes. The proposed method can be easily extended to solve this problem.

References

1. S. Baker and I. Matthews, " Equivalence and Efficiency of Image Alignment Algorithms," *CVPR 2001*, 2001.
2. B. Bascle and A. Blake," Separability of Pose and Expression in Facial Tracing and Animation," *Proc 6th Int. Conf. Computer Vision*, pp. 323-328, 1998.
3. V. Blanz and T. Vetter, " A morphable model for the synthesis of 3D faces," *SIGGRAPH 99*, pp. 187-194, 1999.
4. M. Brand, " Morphable 3D Models from Video," *CVPR 2001*, 2001.
5. M. Brand and R. Bhotika, " Flexible Flow for 3D Nonrigid Tracking and Shape Recovery," *CVPR'01*, vol. 1, pp. 315-22, 2001.
6. C. Bregler, A. Hertzmann and H. Biermann, " Recovering Non-Rigid 3D Shape from Image Streams," *CVPR'00*, 2000.
7. J. Costeira and T. Kanade, " A multibody factorization method for independently moving-objects," *IJCV*, 29(3):159-179, 1998.
8. S.B. Gokturk, J.Y Bouguet, R. Grzeszczuk, " A data driven model for monocular face tracking," *ICCV'01*, 2001.
9. M. Han and T. Kanade, " Reconstruction of a Scene with Multiple Linearly Moving Objects," *CVPR'00*, 2000.
10. C. Poelman and T. Kanade, " A paraperspective factorization method for shape and motion recovery," *PAMI*, 19(3):206-218, 1997.
11. C. Tomasi and T. Kanade, " Shape and motion from image streams under orthography: A factorization method," *IJCV*, 9(2):137-154, 1992.
12. L. Torresani, D. Yang, G. Alexander, C. Bregler, " Tracking and Modeling Non-Rigid Objects with Rank Constraints," *CVPR'01*, 2001.
13. B.Triggs, " Factorization Methods for Projective Structure and Motion," *CVPR'96*,1996.
14. L. Wolf, A. Shashua, " On Projection Matrices $P^k \rightarrow P^2, k = 3, \ldots, 6$, and their Applications in Computer Vision," *IJCV*, 48(1):53-67, 2002.