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

Abstract. Recovery of three dimensional (3D) shape and motion of non-static scenes from a monocular video sequence is important for applications like robot navigation and human computer interaction. If every point in the scene randomly moves, it is impossible to recover the non-rigid shapes. In practice, many non-rigid objects, *e.g.* the human face under various expressions, deform with certain structures. Their shapes can be regarded as a weighted combination of certain shape bases. Shape and motion recovery under such situations has attracted much interest. Previous work on this problem [6,4,13] utilized only orthonormality constraints on the camera rotations (***rotation constraints***). This paper proves that using only the rotation constraints results in ambiguous and invalid solutions. The ambiguity arises from the fact that the shape bases are not unique because their linear transformation is a new set of eligible bases. To eliminate the ambiguity, we propose a set of novel constraints, ***basis constraints***, which uniquely determine the shape bases. We prove that, under the weak-perspective projection model, enforcing both the basis and the rotation constraints leads to 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

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. Reliable systems exist for reconstruction of static scenes. However, most natural scenes are dynamic and non-rigid: expressive faces, people walking beside buildings, etc. Recovering the structure and motion of these non-rigid objects is a challenging task. The effects of 3D rotation and translation and non-rigid deformation are coupled together in image measurement. While it is impossible to reconstruct the shape if the scene deforms arbitrarily, in practice, many non-rigid objects, *e.g.* the human face under various expressions, deform with a class of structures.

One class of solutions model non-rigid object shapes as weighted combinations of certain shape bases that are pre-learned by off-line training [2,3,5,9]. For instance, the geometry of a face is represented as a weighted combination of

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 573–587, 2004.

shape bases that correspond to various facial deformations. Then the recovery of shape and motion is simply a model fitting problem. However, in many applications, *e.g.* reconstruction of a scene consisting of a moving car and a static building, the shape bases of the dynamic structure are difficult to obtain before reconstruction.

Several approaches have been proposed to solve the problem without a prior model [6,13,4]. Instead, they treat the model, *i.e.* shape bases, as part of the unknowns to be solved. They try to recover not only the non-rigid shape and motion, but also the shape model. This class of approaches so far has utilized only the orthonormality constraints on camera rotations (***rotation constraints***) to solve the problem. However, as shown in this paper, enforcing only the rotation constraints leads to ambiguous and invalid solutions. These approaches thus cannot guarantee the desired solution. They have to either require a priori knowledge on shape and motion, *e.g.* constant speed [10], or need non-linear optimization that involves large number of variables and hence requires a good initial estimate [13,4].

Intuitively, the above ambiguity 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 develop a factorization method that applies both metric constraints to compute the closed-form solution for the non-rigid shape, motion, and shape bases.

2 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 most closely related to our work.

The factorization method was first proposed by Tomasi and Kanade [12]. 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 [11] extended it to work under the weak perspective and paraperspective projection models. Triggs [14] generalized the factorization method to the recovery of scene geometry and camera motion under the perspective projection model. These methods work for static scenes.

Costeira and Kanade [8] extended the factorization technique to recover the structure of multiple independently moving objects. This method factorizes the

image locations of certain features to separate different objects and then individually recovers their shapes. Wolf and Shashua [16] derived a geometrical constraint, called the segmentation matrix, to reconstruct a scene containing two independently moving objects from two perspective views. Vidal and his colleagues [15] extended this approach for dynamic scenes containing multiple independently moving objects. For reconstruction of dynamic scenes consisting of both static objects and objects moving along fixed directions, Han and Kanade [10] proposed a factorization-based method that achieves a unique solution with the assumption of constant velocities. A more generalized solution to reconstructing the shapes that deform at constant velocity is presented in [17].

Bregler and his colleagues [6] first introduced the basis representation of non-rigid shapes to embed the deformation constraints into the scene structure. By analyzing the low rank of the image measurements, they proposed a factorization-based method that enforces the orthonormality constraints on camera rotations to reconstruct the non-rigid shape and motion. Torresani and his colleagues [13] extended the method in [6] to a trilinear optimization approach. At each step, two of the three types of unknowns, bases, coefficients, and rotations, are fixed and the remaining one is updated. The method in [6] is used to initialize the optimization process. Brand [4] proposed a similar non-linear optimization method that uses an extension of the method in [6] for initialization. All three methods enforce only the rotation constraints and thus cannot guarantee an optimal solution. Note that both non-linear optimization methods involve a large number of variables, *e.g.* the number of unknown coefficients equals the product of the number of images and the number of shape bases. The performance relies on the quality of the initial estimate of the unknowns.

3 Problem Statement

Given 2D locations of P feature points across F frames, $\{(u,v)^T_{fp}|f=1,...,F,p=1,...,P\}$, 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 $\{(x,y,z)^T_{fp}|f=1,...,F,p=1,...,P\}$. Throughout this paper, we assume:

- the deforming shapes can be represented as weighted combinations of shape bases;
- the 3D structure and the camera motion are non-degenerate;
- the camera projection model is the weak-perspective projection model.

We follow the representation of [3,6]. The non-rigid shapes are represented as weighted combinations 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

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

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

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

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 scalar of the weak perspective projection.

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

$$\mathbf{x}_{fp} = \begin{pmatrix} c_{f1}R_f & ... & c_{fK}R_f \end{pmatrix} \cdot \begin{pmatrix} \mathbf{b}_{1p} \\ ... \\ \mathbf{b}_{Kp} \end{pmatrix} + \mathbf{t}_f \tag{3}$$

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,

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

As in [10,6], 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 under the non-degenerate cases, the rank of $\tilde{W}$ is $3K$ 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.

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 compute the *corrective transformation* matrix G?

4 Metric Constraints

To compute G, two types of metric constraints are available and should be imposed: ***rotation constraints*** and ***basis constraints***. While using only the rotation constraints [6,4] leads to ambiguous and invalid solutions, enforcing both sets of constraints results in a closed-form solution.

4.1 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 [12,11], multiple moving objects [8,10], and non-rigid deforming objects [6,13,4].

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

$$\tilde{M}_{2*i-1:2*i}Q\tilde{M}^T_{2*j-1:2*j} = \Sigma^K_{k=1}c_{ik}c_{jk}R_i * R^T_j, \quad i,j = 1,...F \tag{6}$$

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

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

where $\mathbf{I}_{2\times 2}$ is a 2×2 identity matrix. Because 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,

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

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

For F frames, we have $2F$ linear constraints on $\frac{(9K^2+3K)}{2}$ unknowns. It appears that, when we have enough images, *i.e.* $F \geq \frac{(9K^2+3K)}{2}$, there should be enough constraints to compute Q via the least-square methods. However, it is not true in general. We will show that most of these rotation constraints are redundant and they are inherently insufficient to determine Q.

4.2 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 [12,10]. 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.

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.*

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

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

Each off-diagonal block consists of 3 independent elements. Because Y is symmetric and has $K(K-1)/2$ independent off-diagonal blocks, it 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. $Z_{ij} = \lambda_{ij}\mathbf{I}_{3\times3}$, where λ_{ij} is the only variable.*

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

Theorem 1. *The general solution of the rotation constraints in Eq. (8) and Eq. (9) can be expressed as $\tilde{Q} = GHG^T$, where G is the desired corrective transformation matrix, and $H = Y + Z$, with Y a block-skew-symmetric matrix, and Z a block-scaled-identity matrix.*

Proof. The solution $\tilde{Q}$ of Eq. (8) and Eq. (9) can be represented as $G\Lambda G^T$, since G is a non-singular square matrix. 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),

$$\begin{aligned}\tilde{M}_{2*i-1:2*i}\tilde{Q}\tilde{M}^T_{2*i-1:2*i} &= M_{2*i-1:2*i}\Lambda M^T_{2*i-1:2*i} \\ &= \alpha_i\mathbf{I}_{2\times2}, \quad i = 1, ..., F \end{aligned} \tag{12}$$

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

$$R_i\Sigma^K_{k=1}(c^2_{ik}\Lambda_{kk} + \Sigma^K_{j=k+1}c_{ik}c_{ij}(\Lambda_{kj} + \Lambda^T_{kj}))R^T_i = \alpha_i\mathbf{I}_{2\times2}, \quad i = 1, ..., F \tag{13}$$

Denote the 3×3 symmetric matrix $\Sigma^K_{k=1}(c^2_{ik}\Lambda_{kk} + \Sigma^K_{j=k+1}c_{ik}c_{ij}(\Lambda_{kj} + \Lambda^T_{kj}))$ by Γ_i. Let $\tilde{\Gamma}_i$ be the homogeneous solution of Eq. (13), *i.e.* $R_i\tilde{\Gamma}_iR^T_i = \mathbf{0}_{2\times2}$. Since R_i consists of the first two rows of the i_{th} rotation matrix, let r_{i3} denote the third row. Due to orthonormality of R_i,

$$\tilde{\Gamma}_i = r^T_{i3}\delta_i + \delta^T_i r_{i3} \tag{14}$$

where δ_i is an arbitrary 1×3 vector. Apparently $\Gamma_i = \alpha_i\mathbf{I}_{3\times3}$ is a particular solution of Eq. (13). Therefore the general solution of Eq. (13) is

$$\Gamma_i = \Sigma^K_{k=1}(c^2_{ik}\Lambda_{kk} + \Sigma^K_{j=k+1}c_{ik}c_{ij}(\Lambda_{kj} + \Lambda^T_{kj})) = \alpha_i\mathbf{I}_{3\times3} + \beta_i\tilde{\Gamma}_i \tag{15}$$

where β_i is a scalar. Now let us prove $\beta_i\tilde{\Gamma}_i$ has to be zero. Because $\tilde{Q} = G\Lambda G^T$ is the general solution on all images, Eq. (15) must be satisfied for any set of the coefficients and rotations. For any two frames i and j that are formed by the same 3D shapes, *i.e.* same coefficients, but different rotations R_i and R_j, according to Eq. (15), we have

$$\alpha_i\mathbf{I}_{3\times3} + \beta_i\tilde{\Gamma}_i = \alpha_i\mathbf{I}_{3\times3} + \beta_j\tilde{\Gamma}_j \Longleftrightarrow \beta_i\tilde{\Gamma}_i - \beta_j\tilde{\Gamma}_j = \mathbf{0}_{3\times3} \Longrightarrow R_j(\beta_i\tilde{\Gamma}_i - \beta_j\tilde{\Gamma}_j)R^T_j = \mathbf{0}_{2\times2} \tag{16}$$

According to Eq. (14), we have $R_j\tilde{\Gamma}_jR^T_j = \mathbf{0}_{2\times2}$, thus

$$R_j(\beta_i\tilde{\Gamma}_i)R^T_j = \mathbf{0}_{2\times2} \tag{17}$$

Because R_j can be any rotation matrix, $\beta_i\tilde{\Gamma}_i$ has to be zero for any frame. Therefore,

$$\Sigma^K_{k=1}(c^2_{ik}\Lambda_{kk} + \Sigma^K_{j=k+1}c_{ik}c_{ij}(\Lambda_{kj} + \Lambda^T_{kj})) = \alpha_i\mathbf{I}_{3\times3} \tag{18}$$

Because Eq. (18) must be satisfied for any set of the coefficients, the solution is

$$\Lambda_{kk} = \lambda_{kk}\mathbf{I}_{3\times 3} \tag{19}$$

$$\Lambda_{kj} + \Lambda_{kj}^T = \lambda_{kj}\mathbf{I}_{3\times 3}, \quad k = 1, ..., K; \;\; j = k+1, ..., K \tag{20}$$

where λ_{kk} and λ_{kj} are arbitrary scalars. According to Eq. (19), the diagonal block Λ_{kk} is a scaled identity matrix. From Eq. (20), $\Lambda_{kj} - \frac{\lambda_{kj}}{2}\mathbf{I}_{3\times 3} = -(\Lambda_{kj} - \frac{\lambda_{kj}}{2}\mathbf{I}_{3\times 3})^T$, *i.e.* $\Lambda_{kj} - \frac{\lambda_{kj}}{2}\mathbf{I}_{3\times 3}$ is skew-symmetric. Therefore the off-diagonal block Λ_{kj} equals the summation of a scaled identity block, $\frac{\lambda_{kj}}{2}\mathbf{I}_{3\times 3}$, and a skew-symmetric block, $\Lambda_{kj} - \frac{\lambda_{kj}}{2}\mathbf{I}_{3\times 3}$. This statement 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 $\tilde{Q} = GHG^T$. □

Because 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 $2K^2 - K$ degrees of freedom. It explains why the rotation constraints are sufficient in rigid cases ($K = 1$) but lead to ambiguous solutions when the scene is non-rigid ($K > 1$). This conclusion is also confirmed by our experiments. If every solution in the space is a valid solution of Q, then even if the ambiguity exists, we can compute an arbitrary solution in the space to solve the problem. However, the space contains many invalid solutions. Specifically, since $Q = GG^T$ must be positive semi-definite, when H is not positive semi-definite, the solutions $\tilde{Q} = GHG^T$ are not valid. For example, when H only consists of a block-skew-symmetric matrix Y, the solutions $\tilde{Q} = GYG^T$ are invalid because Y is not positive semi-definite.

4.3 Basis Constraints

Are there other constraints that we can use to remove the ambiguity of the rotation constraints? For static scenes, a variety of approaches [12,11] utilize only the rotation constraints and succeed in determining the correct solution. Intuitively, the only difference between non-rigid and rigid situations is that the non-rigid shape is a weighted combination of certain shape bases. This observation suggests that the ambiguity is related to the basis representation. Can we impose constraints on the bases to eliminate the ambiguity?

The shape bases are non-unique because any non-singular linear transformation on them yields a new set of eligible bases. However, if we find K frames including independent shapes and treat those shapes as a set of bases, the bases are determined uniquely[1]. We denote those frames as the first K images in the sequence and the corresponding coefficients are

$$\begin{aligned} &c_{ii} = 1, \;\; i = 1, ..., K \\ &c_{ij} = 0, \;\; i \neq j, \; i = 1, ..., K, \; j = 1, ..., K \end{aligned} \tag{21}$$

[1] We can find K frames in which the shapes are independent, by examining the singular values of their image projections.

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

$$\tilde{M} g_k = \begin{pmatrix} c_{1k} R_1 \\ ... \\ c_{Fk} R_F \end{pmatrix} \quad k = 1, ..., K \tag{22}$$

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

$$\tilde{M}_{2*i-1:2*i} Q_k \tilde{M}^T_{2*j-1:2*j} = c_{ik} c_{jk} R_i R_j^T \tag{23}$$

Thus Q_k satisfies the rotation constraints in Eq. (8) and Eq. (9). Besides, combining Eq. (21) and Eq. (23), we obtain another $4(K-1)F$ basis constraints on Q_k:

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

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

$$\tilde{M}_{2i-1} Q_k \tilde{M}^T_{2*j} = 0, \quad (i,j) \in \omega_1 \ or \ i = j = k \tag{26}$$

$$\tilde{M}_{2i} Q_k \tilde{M}^T_{2*j-1} = 0, \quad (i,j) \in \omega_1 \ or \ i = j = k \tag{27}$$

where $\omega_1 = \{(i,j) | i = 1, ..., K, \ j = 1, ..., F, \text{ and } i \neq k\}$.

5 A Closed-Form Solution

Due to Theorem 1, enforcing the rotation constraints on Q_k leads to the ambiguous solution $\tilde{Q} = GHG^T$. This section will prove that enforcing the basis constraints eliminates the ambiguity on $\tilde{Q}$ and determines a closed-form solution. Note that we assume that the 3D structure and camera motion are both non-degenerate, *i.e.* the rank of $\tilde{W}$ is $3K$.

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

$$H_{ij} = \begin{pmatrix} h_1 & h_2 & h_3 \\ -h_2 & h_1 & h_4 \\ -h_3 & -h_4 & h_1 \end{pmatrix} \tag{28}$$

Lemma 1 *Under non-degenerate situations, H_{ij} is a zero matrix if,*

$$R_i H_{ij} R_j^T = \begin{pmatrix} r_{i1} \\ r_{i2} \end{pmatrix} H_{ij} \begin{pmatrix} {r_{j1}}^T & r_{j2}^T \end{pmatrix} = \mathbf{0}_{2 \times 2} \tag{29}$$

Proof. First we prove that the rank of H_{ij} is at most 2. Due to the orthonormality constraints,

$$H_{ij} = \begin{pmatrix} r_{i3}^T & \delta_j^T \end{pmatrix} \begin{pmatrix} \delta_i \\ r_{j3} \end{pmatrix} \tag{30}$$

where $r_{i3} = r_{i1} \times r_{i2}$, $r_{j3} = r_{j1} \times r_{j2}$, δ_i and δ_j are two arbitrary 1×3 vectors. Both matrices on the right side of Eq. (30) are at most of rank 2. Thus the rank of H_{ij} is at most 2.

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.

We then prove $h_2 = h_3 = h_4 = 0$. Since $h_1 = 0$, we rewrite Eq. (29) as follows:

$$\begin{pmatrix} r_{i1} \cdot (\mathbf{h} \times r_{j1}) & r_{i1} \cdot (\mathbf{h} \times r_{j2}) \\ r_{i2} \cdot (\mathbf{h} \times r_{j1}) & r_{i2} \cdot (\mathbf{h} \times r_{j2}) \end{pmatrix} = \mathbf{0}_{2\times 2} \tag{31}$$

where $\mathbf{h} = (-h_4 \;\; h_3 \;\; -h_2)$. Eq. (31) 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.* $\mathbf{h} = (-h_4\, h_3 - h_2) = \mathbf{0}_{1\times 3}$. Therefore H_{ij} is a zero matrix. □

According to Lemma 1, we derive the following theorem,

Theorem 2. *Enforcing both basis constraints and rotation constraints results in a unique solution $\tilde{Q} = g_k {g_k}^T$, where g_k is the k_{th} three-column of G.*

Proof. Due to Theorem 1, by enforcing the rotation constraints, we achieve the solution $\tilde{Q} = GHG^T$. Thus $\tilde{M}\tilde{Q}\tilde{M}^T = MHM^T$, and

$$M_{2*i-1:2*i} H M_{2*j-1:2*j}^T = \Sigma_{k_1=1}^{K} \Sigma_{k_2=1}^{K} c_{ik_1} c_{jk_2} R_i H_{k_1 k_2} R_j^T, \quad i, j = 1, ..., F \tag{32}$$

According to Eq. (21),

$$M_{2*i-1:2*i} H M_{2*j-1:2*j}^T = R_i H_{ij} {R_j}^T, \quad i, j = 1, ..., K \tag{33}$$

Due to the basis constraints in Eq. (24) to (27),

$$R_k H_{kk} {R_k}^T = \mathbf{I}_{2\times 2} \tag{34}$$

$$R_i H_{ij} {R_j}^T = \mathbf{0}_{2\times 2}, \quad i, j = 1, ..., K, \text{ and } i \neq k,\ j \neq k \tag{35}$$

By definition, $H_{kk} = \lambda_{kk} \mathbf{I}_{3\times 3}$, where λ_{kk} is a scalar. Due to Eq. (34), $\lambda_{kk} = 1$ and $H_{kk} = \mathbf{I}_{3\times 3}$. From Lemma 1 and Eq. (35), H_{ij} is a zero matrix when $i, j = 1, ..., K$, and $i \neq k$, $j \neq k$. Thus $\tilde{Q} = GHG^T = (g_1, ..., g_K) H (g_1, ..., g_K)^T = (0, ..., 0, g_k, 0, ...0)(g_1, ..., g_K)^T = g_k g_k^T$. □

Now we have proved that, by enforcing both rotation and basis constraints, *i.e.* solving Eq. (8) to (9) and (24) to (27) by the least square methods, a closed-form solution, $\tilde{Q} = Q_k = g_k g_k^T$, $k = 1, ..., K$, is achieved. Then g_k, $k = 1, ..., K$ can be recovered by decomposing Q_k via SVD. We project $g_k' s$ 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.

6 Performance Evaluation

The performance of the closed-form solution is evaluated in a number of experiments.

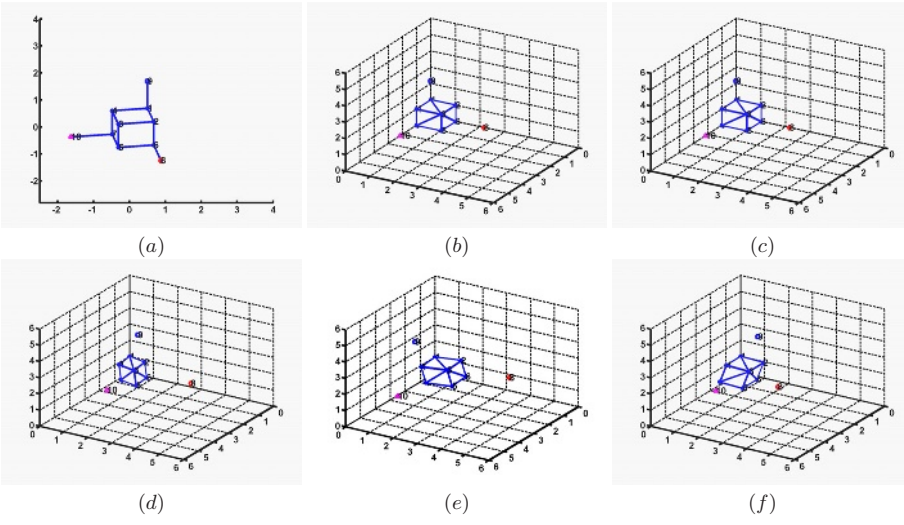

Fig. 1. A static cube and 3 points moving along straight lines. (a) Input image. (b) Ground truth 3D shape. (c) Reconstruction by the closed-form solution. (d) Reconstruction by the method in [6]. (e) Reconstruction by the method in [4] after 4000 iterations. (f) Reconstruction by the tri-linear method [13] after 4000 iterations.

6.1 Comparison with Three Previous Methods

We first compare the solution with three related methods [6,4,13] in a simple noiseless setting. Fig.1 shows a scene consisting of a static cube and 3 moving points. The measurement consists of 10 points: 7 visible vertices of the cube and 3 moving points. The 3 points move along the axes at varying speed. This setting consists 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 Fig.1.(a). The corresponding ground truth structure is demonstrated in Fig.1.(b). Fig.1.(c) to (f) show the structures reconstructed using the closed-form solution, the method in [6], the method in [4], and the tri-linear method [13], respectively. While the closed-form solution achieves the exact reconstruction with zero error, all three previous methods result in apparent errors, even for such a simple noiseless setting. Fig.2 demonstrates the reconstruction errors of the previous work on rotations, shapes, and image measurements. The errors are computed relative to the ground truth.

6.2 Quantitative Evaluation on Synthetic Data

Our approach is then 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. It is represented by the mean power ratio between each two bases, *i.e.* $mean_{i,j}\left(\frac{max(\|B_i\|,\|B_j\|)}{min(\|B_i\|,\|B_j\|)}\right)$. Larger ratio means weaker deformation, *i.e.* the

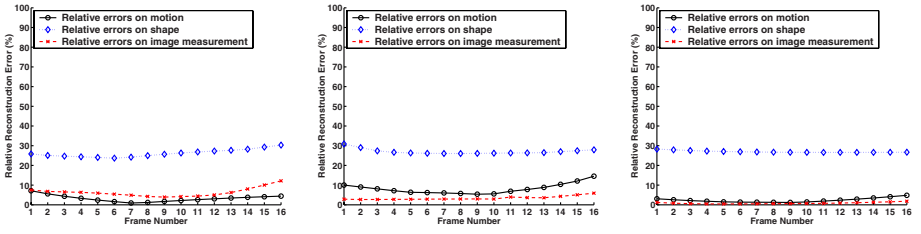

Fig. 2. The relative errors on reconstruction of a static cube and 3 points moving along straight lines. (Left) By the method in [6]. (Middle) By the method in [4] after 4000 iterations. (Right) By the trilinear method [13] after 4000 iterations. The range of the error axis is $[0\%, 100\%]$. Note that our solution achieves zero reconstruction errors.

shape is closer to rigid. The number of shape bases represents the flexibility of the shape. A bigger basis number means that the shape is more flexible. Assuming a Gaussian white noise, we represent the noise strength level by the ratio between the Frobenius norm of the noise and the measurement, *i.e.* $\frac{\|noise\|}{\|\tilde{W}\|}$. In general, when noise exists, a weaker deformation leads to better performance, because some deformation mode is more dominant and the noise relative to the dominant basis is weaker; a bigger basis number results in poorer performance, because the noise relative to each individual basis is stronger.

Fig. 3.(a) and (b) show the performance of our algorithm under various deformation strength and noise levels on a two bases setting. The power ratios are respectively 2^0, 2^1, ..., and 2^8. Four levels of Gaussian white noise are imposed. Their strength levels are 0%, 5%, 10%, and 20% respectively. We test a number of trials on each setting and compute the average reconstruction errors on the rotations and 3D shapes, relative to the ground truth. Fig.3.(c) and (d) show the performance of our method under different numbers of shape bases and noise levels. The basis number is 2, 3, ... , and 10 respectively. The bases have equal powers and thus none of them is dominant. The same noise as in the last experiment is imposed.

In both experiments, when the noise level is 0%, the closed-form solution always recovers the exact rotations and shapes with zero error. When there is noise, it achieves reasonable accuracy, *e.g.* the maximum reconstruction error is less than 15% when the noise level is 20%. As we expected, under the same noise level, the performance is better when the power ratio is larger and 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.

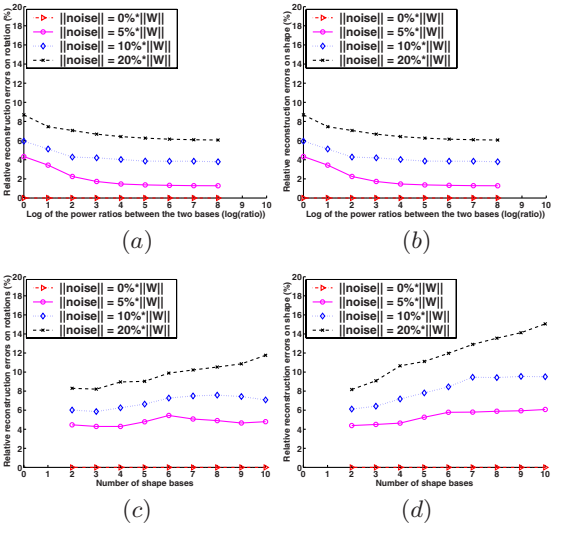

Fig. 3. $(a)\&(b)$ Reconstruction errors on rotations and shapes under different levels of noise and deformation strength. $(c)\&(d)$ Reconstruction errors on rotations and shapes under different levels of noise and various basis numbers. Each curve respectively refers to a noise level. The range of the error axis is $[0\%, 20\%]$.

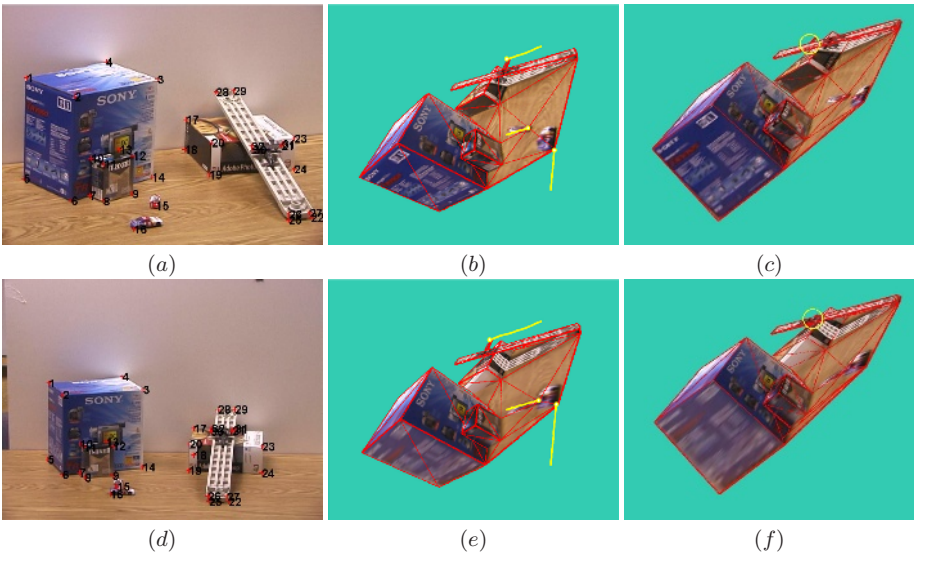

Fig. 4. Reconstruction of three moving objects in the static background. $(a)\&(d)$ Two input images with marked features. $(b)\&(e)$ Reconstruction by the closed-form solution. The yellow lines show the recovered trajectories from the beginning of the sequence until the present frames. $(c)\&(f)$ Reconstruction by the method in [4]. The yellow-circled area shows that the plane, which should be on top of the slope, is mistakenly located underneath the slope.

6.3 Qualitative Evaluation on Real Video Sequences

Finally we examine our approach qualitatively on a number of real video sequences. One example is shown in Fig.4. The sequence was taken of an indoor scene by a handheld 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 were used for reconstruction. Two of the them are shown in Fig.4.(a) and (d).

The rank of $\tilde{W}$ was estimated in such a way that after rank reduction 99% of the energy was kept. The basis number is automatically determined by $K = rank(\tilde{W})/3$. The camera rotations and dynamic scene structure are then reconstructed. To evaluate the reconstruction, we synthesize the scene appearance viewed from one side, as shown in Fig.4.(b) and (e). The wireframes show the structure and the yellow lines show the trajectories of the moving objects from the beginning of the sequence until the present frames. The reconstruction is consistent with our observation, *e.g.* the plane moved linearly on top of the slope. Fig.4.(c) and (f) show the reconstruction using the method in [4]. The shapes of the boxes are distorted and the plane is incorrectly located underneath the slope, as shown in the yellow circles. Note that occlusion was not taken into account when rendering these images, thus in the regions that should be occluded, *e.g.* the area behind the slope, the stretched texture of the occluding objects appears.

Human faces are highly non-rigid objects and 3D face shapes can be represented as weighted combinations of certain shape bases that refer to various facial expressions. They thus can be reconstructed by our approach. One example is shown in Fig.5. The sequence consists of 236 images that contain expressions like eye blinking and mouth opening. 60 feature points were tracked using an efficient Active Appearance Model (AAM) method [1]. Fig.5.(a) and (d) display two input images with marked features. Their corresponding shapes are reconstructed and shown from novel views in Fig.5.(b) and (e). Their corresponding 3D wireframe models shown in Fig.5.(c) and (f) demonstrate the recovered facial deformations such as mouth opening and eye closure. 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 demonstrates its robustness to the image noise.

7 Conclusion and Discussion

This paper proposes a closed-form solution to the problem of non-rigid shape and motion recovery from single-camera video using the least square and factorization methods. In particular, we have proven that enforcing only the rotation constraints results in ambiguous and invalid solutions. We thus introduce the basis constraints to remove this ambiguity. We have also proven that imposing both metric constraints leads to a unique reconstruction of the non-rigid shape

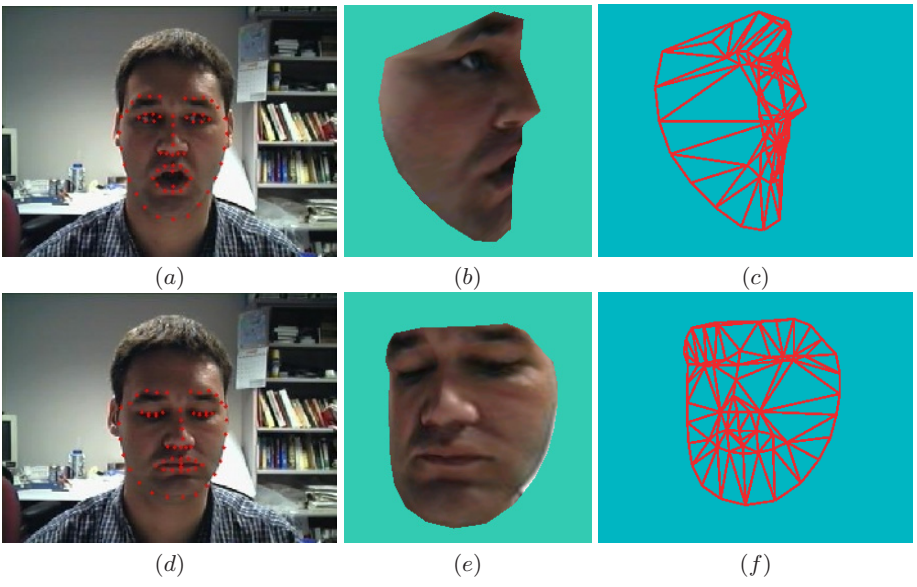

Fig. 5. Reconstruction of face shapes with expressions. $(a)\&(d)$ Input images. $(b)\&(e)$ Reconstructed face shapes seen from novel views. $(c)\&(f)$ The wireframe models demonstrate the recovered facial deformations such as mouth opening and eye closure.

and motion. The performance of our algorithm is demonstrated by experiments on both simulated data and real video data. Our algorithm has also been successfully applied to separate the local deformations from the global rotations and translations in the 3D motion capture data [7].

Currently, our approach does not consider the degenerate deformation modes of 3D shapes. 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. For example, if a scene contains only one moving object that moves along a straight line, the deformation mode referring to the linear motion is degenerate, because the corresponding basis (the motion vector) is of rank 1. It is conceivable that the ambiguity cannot be completely eliminated by the basis constraints and enforcing both metric constraints is insufficient to produce a closed-form solution in such degenerate cases. We are now exploring how to extend the current approach to recovering the non-rigid shapes that deform with degenerate modes. Another limitation of our approach is that we assume the weak perspective projection model. It would be interesting to see if the proposed approach could be extended to the full perspective projection model.

Acknowledgments. We would like to thank Simon Baker, Iain Matthews, and Mei Han for providing the image data and feature correspondence used in Section 6.3, and thank Jessica Hodgins for proofreading the paper. Jinxiang Chai was supported by the NSF through EIA0196217 and IIS0205224. Jing Xiao and

Takeo Kanade were partly supported by grant R01 MH51435 from the National Institute of Mental Health.

References

1. S. Baker, I. Matthews, " Equivalence and Efficiency of Image Alignment Algorithms," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
2. B. Bascle, A. Blake," Separability of Pose and Expression in Facial Tracing and Animation," *Proc. Int. Conf. Computer Vision*, pp. 323-328, 1998.
3. V. Blanz, T. Vetter, " A morphable model for the synthesis of 3D faces," *Proc. SIGGRAPH'99*, pp. 187-194, 1999.
4. M. Brand, " Morphable 3D Models from Video," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
5. M. Brand, R. Bhotika, " Flexible Flow for 3D Nonrigid Tracking and Shape Recovery," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
6. C. Bregler, A. Hertzmann, H. Biermann, " Recovering Non-Rigid 3D Shape from Image Streams," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2000.
7. J. Chai, J. Xiao, J. Hodgins, " Vision-based Control of 3D Facial Animation," *Eurographics/ACM Symposium on Computer Animation*, 2003.
8. J. Costeira, T. Kanade, " A multibody factorization method for independently moving-objects," *Int. Journal of Computer Vision*, 29(3):159-179, 1998.
9. S.B. Gokturk, J.Y Bouguet, R. Grzeszczuk, " A data driven model for monocular face tracking," *Proc. Int. Conf. Computer Vision*, 2001.
10. M. Han, T. Kanade, " Reconstruction of a Scene with Multiple Linearly Moving Objects," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2000.
11. C. Poelman, T. Kanade, " A paraperspective factorization method for shape and motion recovery," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(3):206-218, 1997.
12. C. Tomasi, T. Kanade, " Shape and motion from image streams under orthography: A factorization method," *Int. Journal of Computer Vision*, 9(2):137-154, 1992.
13. L. Torresani, D. Yang, G. Alexander, C. Bregler, " Tracking and Modeling Non-Rigid Objects with Rank Constraints," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
14. B. Triggs, " Factorization Methods for Projective Structure and Motion," *Proc. Int. Conf. Computer Vision and Pattern Recognition*,1996.
15. R. Vidal, S. Soatto, Y. Ma, S. Sastry, " Segmentation of Dynamic Scenes from the Multibody Fundamental Matrix," *ECCV Workshop on Vision and Modeling of Dynamic Scenes*, 2002.
16. L. Wolf, A. Shashua, " Two-body Segmentation from Two Perspective Views," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
17. L. Wolf, A. Shashua, " On Projection Matrices $P^k \rightarrow P^2, k = 3, \ldots, 6$, and their Applications in Computer Vision," *Int. Journal of Computer Vision*, 48(1):53-67, 2002.