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

Jing Xiao, Jin-xiang Chai, Takeo Kanade

The Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
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, 14] 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. An arbitrary linear transformation of the bases produces another 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. Many reliable methods have been proposed for reconstruction of static scenes [13, 11, 15]. However, most biological objects and natural scenes vary their shapes: expressive faces, people walking beside buildings, etc. Recovering the structure and motion of these non-rigid objects is a challenging task. The effects of rigid motion, *i.e.* 3D rotation and translation, and non-rigid shape deformation, *e.g.* stretching, are coupled together in the image measurements. 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 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, 14, 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 [14, 4].

Intuitively, the above ambiguity arises from the non-uniqueness of the shape bases: an arbitrary linear transformation of a set of shape bases yields 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 linear closed-form solution to the problem of non-rigid shape and motion recovery is achieved. Accordingly we develop a factorization method that applies both the 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 originally proposed by Tomasi and Kanade [13]. 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 para-perspective projection models. Triggs [15] 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 [17] 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 [16] 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 [18].

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 [14] 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 the three methods enforce only the rotation constraints and thus cannot guarantee an optimal solution. Note that both the 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 $\tilde{W} = MB$.

Under the non-degenerate cases, the $2F \times 3K$ scaled rotation matrix M and the $3K \times P$ shape bases matrix B are both of full rank, respectively $min\{2F, 3K\}$ and $min\{3K, P\}$. Their product, $\tilde{W}$, is of rank $min\{3K, 2F, P\}$. In practice, the frame number F and point number P

are usually much larger than the basis number K such that $2F > 3K$ and $P > 3K$. Thus the rank of $\tilde{W}$ is $3K$ and K is determined by $K = \frac{rank(\tilde{W})}{3}$. We then factorize $\tilde{W}$ into the product of a $2F \times 3K$ matrix $\tilde{M}$ and a $3K \times P$ matrix $\tilde{B}$, using Singular Value Decomposition (SVD). 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 the non-singular $3K \times 3K$ matrix 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

G is made up of K triple-columns, denoted as g_k, $k = 1, \ldots, K$. Each of them is a $3K \times 3$ matrix. They are independent on each other because G is non-singular. According to Eq. (4,5), g_k satisfies,

$$\tilde{M} g_k = \begin{pmatrix} c_{1k} R_1 \\ \ldots \\ c_{Fk} R_F \end{pmatrix} \tag{6}$$

Then,

$$\tilde{M} g_k g_k^T \tilde{M}^T = \begin{pmatrix} c_{1k}^2 R_1 R_1^T & c_{1k} c_{2k} R_1 R_2^T & \ldots & c_{1k} c_{Fk} R_1 R_F^T \\ c_{1k} c_{2k} R_2 R_1^T & c_{2k}^2 R_2 R_2^T & \ldots & c_{2k} c_{Fk} R_2 R_F^T \\ \vdots & \vdots & \ddots & \vdots \\ c_{1k} c_{Fk} R_F R_1^T & c_{2k} c_{Fk} R_F R_2^T & \ldots & c_{Fk}^2 R_F R_F^T \end{pmatrix} \tag{7}$$

We denote $g_k {g_k}^T$ by Q_k, a $3K \times 3K$ symmetric matrix. Once Q_k is determined, g_k can be computed uniquely using SVD. To compute Q_k, 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 linear closed-form solution for Q_k.

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 [13, 11], multiple moving objects [8, 10], and non-rigid deforming objects [6, 14, 4].

According to Eq. (7), we have,

$$\tilde{M}_{2i-1:2i} Q_k \tilde{M}_{2j-1:2j}^T = c_{ik} c_{jk} R_i R_j^T, \quad i, j = 1, \ldots F \tag{8}$$

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

$$\tilde{M}_{2i-1:2i} Q_k \tilde{M}_{2i-1:2i}^T = c_{ik}^2 \mathbf{I}_{2\times 2}, \quad i = 1, ..., F \tag{9}$$

where $\mathbf{I}_{2\times 2}$ is a 2×2 identity matrix. The two diagonal elements of Eq. (9) yield one linear constraints on Q_k, since c_{ik} is unknown. The two off-diagonal constraints are identical, because Q_k is symmetric. For all F frames, we obtain $2F$ linear constraints as follows,

$$\tilde{M}_{2i-1} Q_k \tilde{M}_{2i-1}^T - \tilde{M}_{2i} Q_k \tilde{M}_{2i}^T = 0, \quad i = 1, ..., F \tag{10}$$

$$\tilde{M}_{2i-1} Q_k \tilde{M}_{2i}^T = 0, \quad i = 1, ..., F \tag{11}$$

Since Q_k is symmetric, it contains $\frac{(9K^2+3K)}{2}$ independent unknowns. It appears that, when enough images are given, *i.e.* $2F \geq \frac{(9K^2+3K)}{2}$, the rotation constraints in Eq. (10,11) should be sufficient to determine Q_k via the linear least-square method. However, it is not true in general. We will show that most of the rotation constraints are redundant and they are inherently insufficient to resolve Q_k.

4.2 Why Are Rotation Constraints Not Sufficient?

Under specific assumptions like static scene or constant speed of deformation, the orthonormality constraints are sufficient to reconstruct the $3D$ shapes and camera rotations [13, 10]. In general cases, however, no matter how many images or feature points are given, the solution of the rotation constraints in Eq. (10,11) is 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.*

$$Y_{ij} = \begin{pmatrix} 0 & y_{ij1} & y_{ij2} \\ -y_{ij1} & 0 & y_{ij3} \\ -y_{ij2} & -y_{ij3} & 0 \end{pmatrix} = -Y_{ij}^T = Y_{ji}^T, \quad i \neq j \tag{12}$$

$$Y_{ii} = 0_{3\times 3}, \quad i, j = 1, ..., K \tag{13}$$

Each off-diagonal block consists of 3 independent elements. Because Y is symmetric and has $\frac{K(K-1)}{2}$ independent off-diagonal blocks, it includes $\frac{3K(K-1)}{2}$ 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.*

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

Theorem 1. *The general solution of the rotation constraints in Eq. (10,11) is GHG^T, where G is the desired corrective transformation matrix, and H is the summation of an arbitrary block-skew-symmetric matrix and an arbitrary block-scaled-identity matrix.*

Proof. Let us denote $\tilde{Q}$ as the general solution of Eq. (10,11). Since G is a non-singular square matrix, $\tilde{Q}$ can be represented as GHG^T, where $H = G^{-1}\tilde{Q}G^{-T}$. We then prove that H must be the summation of a block-skew-symmetric matrix and a block-scaled-identity matrix.

According to Eq. (5,9),

$$\begin{aligned} c_{ik}^2\mathbf{I}_{2\times 2} &= \tilde{M}_{2i-1:2i}\tilde{Q}\tilde{M}_{2i-1:2i}^T \\ &= \tilde{M}_{2i-1:2i}GHG^T\tilde{M}_{2i-1:2i}^T \\ &= M_{2i-1:2i}HM_{2i-1:2i}^T, \quad i = 1, ..., F \end{aligned} \tag{14}$$

H consists of K^2 3×3 blocks, denoted as H_{mn}, m,n=1,. . .,K. Combining Eq. (4) and (14), we have,

$$R_i \varSigma_{m=1}^K (c_{im}^2 H_{mm} + \varSigma_{n=m+1}^K c_{im}c_{in}(H_{mn} + H_{mn}^T))R_i^T = c_{ik}^2\mathbf{I}_{2\times 2}, \quad i = 1, ..., F \tag{15}$$

Denote the 3×3 symmetric matrix $\varSigma_{m=1}^K(c_{im}^2 H_{mm} + \varSigma_{n=m+1}^K c_{im}c_{in}(H_{mn} + H_{mn}^T))$ by $\varGamma_i$. Then Eq. (15) becomes $R_i\varGamma_i R_i^T = c_{ik}^2\mathbf{I}_{2\times2}$. Let $\tilde{\varGamma}_i$ be its homogeneous solution, *i.e.* $R_i\tilde{\varGamma}_i R_i^T = \mathbf{0}_{2\times2}$. Because the two rows of the 2×3 matrix R_i are orthonormal, we have,

$$\tilde{\varGamma}_i = r_{i3}^T\delta_i + \delta_i^T r_{i3} \tag{16}$$

where r_{i3} is a unitary 1×3 vector that are orthogonal to both rows of R_i. δ_i is an arbitrary 1×3 vector. Apparently $\varGamma_i = c_{ik}^2\mathbf{I}_{3\times3}$ is a particular solution of $R_i\varGamma_i R_i^T = c_{ik}^2\mathbf{I}_{2\times2}$. Thus the general solution of Eq. (15) is,

$$\varSigma_{m=1}^K(c_{im}^2 H_{mm} + \varSigma_{n=m+1}^K c_{im}c_{in}(H_{mn} + H_{mn}^T)) = \varGamma_i = c_{ik}^2\mathbf{I}_{3\times3} + \alpha_i\tilde{\varGamma}_i \tag{17}$$

where α_i is an arbitrary scalar.

As a general solution, $\tilde{Q}$ should work for arbitrary image projection of the non-rigid shape, *i.e.* Eq. (17) has to be satisfied for arbitrary coefficients and rotations. Suppose two images j and l contain the projections of the same 3D shapes as that in image i, but from different views. All the three images refer to the same coefficients but different rotations respectively. According to the left side of Eq. (17), $\varGamma_i$, $\varGamma_j$, and $\varGamma_l$ are independent on the rotations and thus equivalent. For image i and j, we have,

$$\begin{aligned} c_{ik}^2\mathbf{I}_{3\times3} + \alpha_i\tilde{\varGamma}_i = c_{ik}^2\mathbf{I}_{3\times3} + \alpha_j\tilde{\varGamma}_j &\Longleftrightarrow \alpha_i\tilde{\varGamma}_i - \alpha_j\tilde{\varGamma}_j = \mathbf{0}_{3\times3} \\ &\Longrightarrow R_j(\alpha_i\tilde{\varGamma}_i - \alpha_j\tilde{\varGamma}_j)R_j^T = \mathbf{0}_{2\times2} \end{aligned} \tag{18}$$

$\tilde{\varGamma}_j$ is the solution of $R_j\tilde{\varGamma}_j R_j^T = \mathbf{0}_{2\times2}$. Thus $\alpha_i R_j\tilde{\varGamma}_i R_j^T = \mathbf{0}_{2\times2}$. Similarly, we have $\alpha_i R_l\tilde{\varGamma}_i R_l^T = \mathbf{0}_{2\times2}$. Then,

$$\alpha_i(R_j\tilde{\varGamma}_i R_j^T - R_l\tilde{\varGamma}_i R_l^T) = \mathbf{0}_{2\times2} \tag{19}$$

Because R_j and R_l are different, α_i has to be zero. We then rewrite Eq. (17) as follows,

$$\varSigma_{m=1}^K(c_{im}^2 H_{mm} + \varSigma_{n=m+1}^K c_{im}c_{in}(H_{mn} + H_{mn}^T)) = c_{ik}^2\mathbf{I}_{3\times3} \tag{20}$$

Denote $(H_{mn} + H_{mn}^T)$ by $\varTheta_{mn}$. Because the right side of Eq. (20) is a scaled identity matrix, for each off-diagonal element h_{mm}^o and θ_{mn}^o, we achieve the following linear equation,

$$\varSigma_{m=1}^K(c_{im}^2 h_{mm}^o + \varSigma_{n=m+1}^K c_{im}c_{in}\theta_{mn}^o)) = 0 \tag{21}$$

As a general solution, Eq. (21) has to be satisfied for arbitrary coefficient sets. Given K^2 independent sets of coefficients, Eq. (21) leads to a non-singular linear equation set on h^o_{mm} and θ^o_{mn}. The right sides of the equations are all zeros. Thus the solution is a zero vector, *i.e.* the off-diagonal elements of H_{mm} and Θ_{mn} are all zeros. Similarly, we can derive the constraint as Eq. (21) on the difference between any two diagonal elements. Therefore the difference is zero, *i.e.* the diagonal elements are all equivalent. We thus have,

$$H_{mm} = \lambda_{mm}\mathbf{I}_{3\times3}, \quad m = 1, ..., K \tag{22}$$

$$H_{mn} + H^T_{mn} = \lambda_{mn}\mathbf{I}_{3\times3}, \quad m = 1, ..., K, \;\; n = m+1, ..., K \tag{23}$$

where λ_{mm} and λ_{mn} are arbitrary scalars. According to Eq. (22), the diagonal block H_{mm} is a scaled identity matrix. Due to Eq. (23), $H_{mn} - \frac{\lambda_{mn}}{2}\mathbf{I}_{3\times3} = -(H_{mn} - \frac{\lambda_{mn}}{2}\mathbf{I}_{3\times3})^T$, *i.e.* $H_{mn} - \frac{\lambda_{mn}}{2}\mathbf{I}_{3\times3}$ is skew-symmetric. Thus the off-diagonal block H_{mn} equals the summation of a scaled identity block, $\frac{\lambda_{mn}}{2}\mathbf{I}_{3\times3}$, and a skew-symmetric block, $H_{mn} - \frac{\lambda_{mn}}{2}\mathbf{I}_{3\times3}$. Since λ_{mm} and λ_{mn} are arbitrary, the entire matrix H is the summation of an arbitrary block-skew-symmetric matrix and an arbitrary block-scaled-identity matrix.
□

Let Y denote the block-skew-symmetric matrix and Z denote the block-scaled-identity matrix in H. Since Y and Z respectively contain $\frac{3K(K-1)}{2}$ and $\frac{K(K+1)}{2}$ independent elements, H include $2K^2 - K$ free elements, *i.e.* the solution of the rotation constraints has $2K^2 - K$ degrees of freedom. In rigid cases, *i.e.* $K = 1$, the solution is unique, as suggested in [13]. For non-rigid objects, *i.e.* $K > 1$, the rotation constraints result in an ambiguous solution space. This space contains invalid solutions. Specifically, because the desired $Q_k = g_k {g_k}^T$ is positive semi-definite, the solution GHG^T is not valid when H is not positive semi-definite. The solution space includes many instances that refer to non-positive-semi-definite H. For example, when the block-scaled-identity matrix Z is zero, H equals the block-skew-symmetric matrix Y, which is not positive semi-definite.

4.3 Basis Constraints

The only difference between non-rigid and rigid situations is that the non-rigid shape is a weighted combination of certain shape bases. The rotation constraints are sufficient for recovering the rigid shapes, but they cannot determine a unique set of shape bases in the non-rigid cases. Instead any non-singular linear transformation applied on the bases leads to another set of eligible bases. Intuitively, the basis non-uniqueness results in the solution ambiguity of the rotation constraints. We thus introduce the basis constraints that determine a unique bases set and resolve the ambiguity.

Because the deformable shapes lie in a K-bases linear space, any K independent shapes in the space form an eligible bases set. We thus select K frames that contain independent $3D$ shapes, and specify those shapes as a set of bases. The K frames of image measurements form a $2K \times P$ sub-matrix of $\tilde{W}$. Its condition number measures the independence of the K involved shapes. A smaller condition number refers to stronger independence. We thus specify the $3D$ shapes in the set of K frames, for which the condition number is the smallest, as the bases. Note that so far we have not recovered the bases, but decided in which frames they are located. This step implicitly determines a unique set of bases.

We denote the selected frames as the first K images in the sequence. The corresponding coefficients are,

$$\begin{aligned} c_{ii} &= 1,\ \ i = 1, ..., K \\ c_{ij} &= 0,\ \ i, j = 1, ..., K,\ i \neq j \end{aligned} \tag{24}$$

Let Ω denote the set, $\{(i, j)|i = 1, ..., K;\ i \neq k;\ j = 1, ..., F\}$. According to Eq. (8,24), we have,

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

$$\tilde{M}_{2i} Q_k \tilde{M}_{2j}^T = \begin{cases} 1,\ i = j = k \\ 0,\ (i, j) \in \Omega \end{cases} \tag{26}$$

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

$$\tilde{M}_{2i} Q_k \tilde{M}_{2j-1}^T = 0, \quad (i, j) \in \Omega\ or\ i = j = k \tag{28}$$

These $4F(K - 1)$ linear constraints are called the basis constraints.

5 A Closed-Form Solution

Due to Theorem 1, enforcing the rotation constraints on Q_k leads to the ambiguous solution GHG^T. H consists of K^2 3×3 blocks, H_{mn}, m,n=1,. . .,K. H_{mn} contains four independent entries as follows,

$$H_{mn} = \begin{pmatrix} h_1 & h_2 & h_3 \\ -h_2 & h_1 & h_4 \\ -h_3 & -h_4 & h_1 \end{pmatrix} \tag{29}$$

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

$$R_i H_{mn} R_j^T = \mathbf{0}_{2\times 2} \tag{30}$$

where R_i and R_j are 2×3 rotation matrices.

Proof. First we prove that the rank of H_{mn} is at most 2. Due to the orthonormality of rotation matrices, from Eq. (30), we have,

$$H_{mn} = r_{i3}^T \delta_i + \delta_j^T r_{j3} = \begin{pmatrix} r_{i3}^T & \delta_j^T \end{pmatrix} \begin{pmatrix} \delta_i \\ r_{j3} \end{pmatrix} \tag{31}$$

where r_{i3} and r_{j3} respectively are the cross products of the two rows of R_i and those of R_j. δ_i and δ_j are arbitrary 1×3 vectors. Because both matrices on the right side of Eq. (31) are at most of rank 2, the rank of H_{mn} is at most 2.

Next, we prove $h_1 = 0$. Since H_{mn} is 3×3 matrix of rank 2, its determinant, $h_1(\sum_{i=1}^4 {h_i}^2)$, equals 0. Therefore $h_1 = 0$, *i.e.* H_{mn} is a skew-symmetric matrix.

Finally we prove $h_2 = h_3 = h_4 = 0$. Denote the rows of R_i and R_j as r_{i1}, r_{i2}, r_{j1}, and r_{j2} respectively. Since $h_1 = 0$, we can rewrite Eq. (30) as follows,

$$\begin{pmatrix} r_{i1} \cdot (\mathbf{h} \times r_{j1}) & r_{i1} \cdot (\mathbf{h} \times r_{j2}) \\ r_{i2} \cdot (\mathbf{h} \times r_{j1}) & r_{i2} \cdot (\mathbf{h} \times r_{j2}) \end{pmatrix} = \mathbf{0}_{2\times 2} \tag{32}$$

where $\mathbf{h} = (-h_4 \, h_3 \, -h_2)$. Eq. (32) 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_{mn} is a zero matrix. □

Using Lemma 1, we derive the following theorem,

Theorem 2. *Enforcing both the basis constraints and the rotation constraints results in a closed-form solution of* Q_k.

Proof. According to Theorem 1, using the rotation constraints we achieve the ambiguous solution of Q_k, GHG^T. Due to the basis constraints, replacing Q_k in Eq. (25∼28) with GHG^T,

$$\tilde{M}_{2k-1:2k} GHG^T \tilde{M}^T_{2k-1:2k} = M_{2k-1:2k} H M^T_{2k-1:2k} = \mathbf{I}_{2\times 2} \quad (33)$$

$$\tilde{M}_{2i-1:2i} GHG^T \tilde{M}^T_{2j-1:2j} = M_{2i-1:2i} H M^T_{2j-1:2j} = \mathbf{0}_{2\times 2}, \quad i, j = 1, ..., K, \; i \neq k \; or \; j \neq k \quad (34)$$

From Eq. (4), we have,

$$M_{2i-1:2i} H M^T_{2j-1:2j} = \Sigma^K_{m=1} \Sigma^K_{n=1} c_{im} c_{jn} R_i H_{mn} R^T_j, \quad i, j = 1, ..., F \quad (35)$$

where H_{mn} is the 3×3 block of H. According to Eq. (24),

$$M_{2i-1:2i} H M^T_{2j-1:2j} = R_i H_{ij} {R_j}^T, \quad i, j = 1, ..., K \quad (36)$$

Combining Eq. (33,34) and (36), we have,

$$R_k H_{kk} {R_k}^T = \mathbf{I}_{2\times 2} \quad (37)$$

$$R_i H_{ij} {R_j}^T = \mathbf{0}_{2\times 2}, \quad i, j = 1, ..., K, \; i \neq k \; or \; j \neq k \quad (38)$$

By definition, the k_{th} diagonal block $H_{kk} = \lambda_{kk} \mathbf{I}_{3\times 3}$. Due to Eq. (37), $\lambda_{kk} = 1$ and $H_{kk} = \mathbf{I}_{3\times 3}$. According to Lemma 1, all the other blocks, H_{ij} in Eq. (38), are zero matrices. Thus,

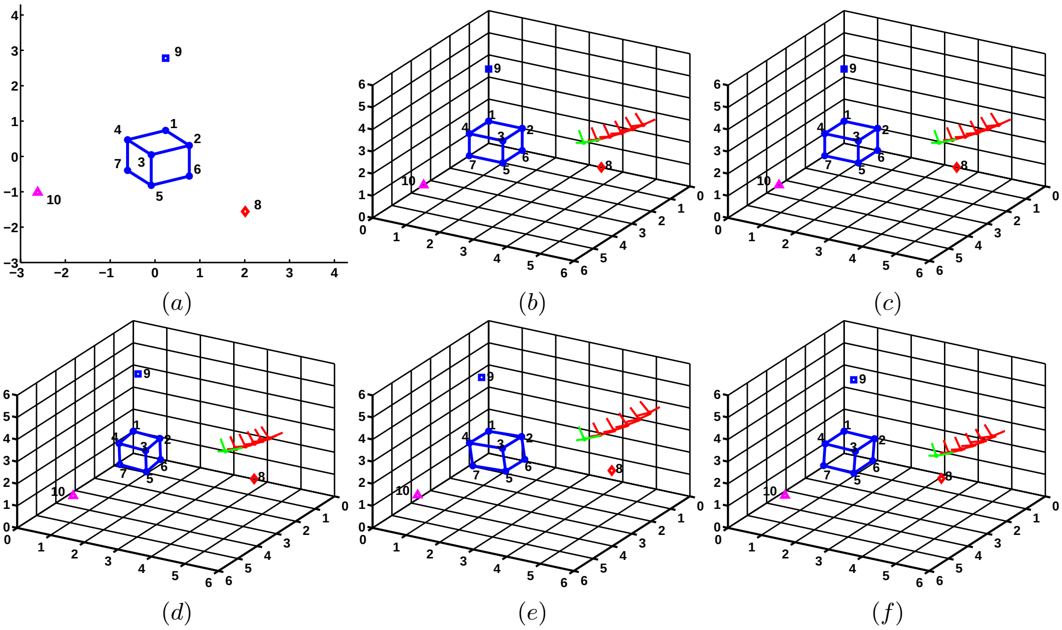
$$\begin{aligned} GHG^T &= (g_1 \dots g_K) H (g_1 \dots g_K)^T \\ &= (0 \dots g_k \dots 0)(g_1 \dots g_K)^T \\ &= g_k g_k^T = Q_k \end{aligned} \quad (39)$$

i.e. a closed-form solution of the desired Q_k has been achieved.
□

According to Theorem 2, we compute $Q_k = g_k g_k^T$, $k = 1, ..., K$, by solving the linear equations, Eq. (10∼11,25∼28), via the least square methods. We then recover g_k by decomposing Q_k via SVD. The decomposition of Q_k is up to an arbitrary 3×3 orthonormal transformation Φ, since $(g_k \Phi)(g_k \Phi)^T$ also equals Q_k. This ambiguity arises from the fact that $g_1, \dots, g_K$ are estimated independently under different coordinate systems. To resolve the ambiguity, we need to transform $g_1, \ldots, g_K$ to be under a single reference coordinate system.

Due to Eq. (6), $M_{2i-1:2i}g_k = c_{ik}R_i, i = 1, \ldots, F$. Because the rotation matrix R_i is orthonormal, *i.e.* $\|R_i\| = 1$, we have $R_i = \pm\frac{M_{2i-1:2i}g_k}{\|M_{2i-1:2i}g_k\|}$. The sign of R_i determines which orientations are in front of the camera. It can be either positive or negative, determined by the reference coordinate system. Since $g_1, \ldots, g_K$ are estimated independently, they lead to respective rotation sets, each two of which are different up to a 3×3 orthonormal transformation. We choose one set of the rotations to specify the reference coordinate system. Then the signs of the other sets of rotations are determined in such a way that these rotations are consistent with the corresponding references. Finally the orthogonal Procrustes method [12] is applied to compute the orthonormal transformations from the rotation sets to the reference. The transformed $g_1, \ldots, g_K$ form the desired corrective transformation G. The coefficients are then computed by Eq. (6), and the shape bases are recovered by Eq. (5). Their combinations reconstruct the non-rigid $3D$ shapes.

6 Performance Evaluation



(a) (b) (c)

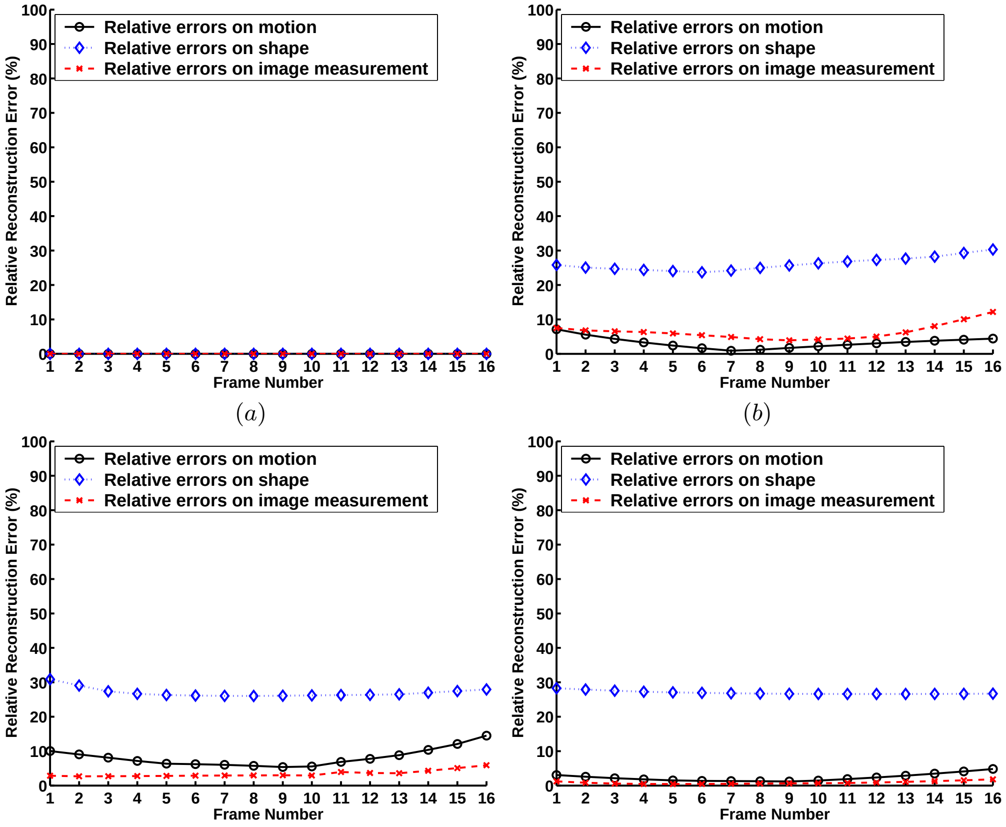
(d) (e) (f)

Fig. 1. A static cube and 3 points moving along straight lines. (a) Input image. (b) Ground truth 3D shape and camera trajectory. (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 [14] after 4000 iterations.

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

6.1 Comparison with Three Previous Methods

We first compared the solution with three related methods [6, 4, 14] in a simple noiseless setting. Fig.1 shows a scene consisting of a static cube and 3 moving points. The measurement included 10 points: 7 visible vertices of the cube and 3 moving points. The 3 points moved along the three axes simultaneously at varying speed. This setting involved $K = 2$ shape bases, one for the static cube and another for the linear motions. While the points were moving, the camera was rotating around the scene. A sequence of 16 frames were captured. One of them is shown in Fig.1.(a). Fig.1.(b) demonstrates the ground truth shape in this frame and the ground truth camera trajectory from the first frame till this frame. The three orthogonal green bars show the present camera pose and the red bars display the camera poses in the previous frames. Fig.1.(c) to (f) show the structures and camera trajectories reconstructed using the closed-form solution, the method in [6], the method in [4], and the tri-linear method [14], respectively. While the closed-form solution achieved the exact reconstruction with zero error, all the three previous methods resulted in apparent errors, even for such a simple noiseless setting.



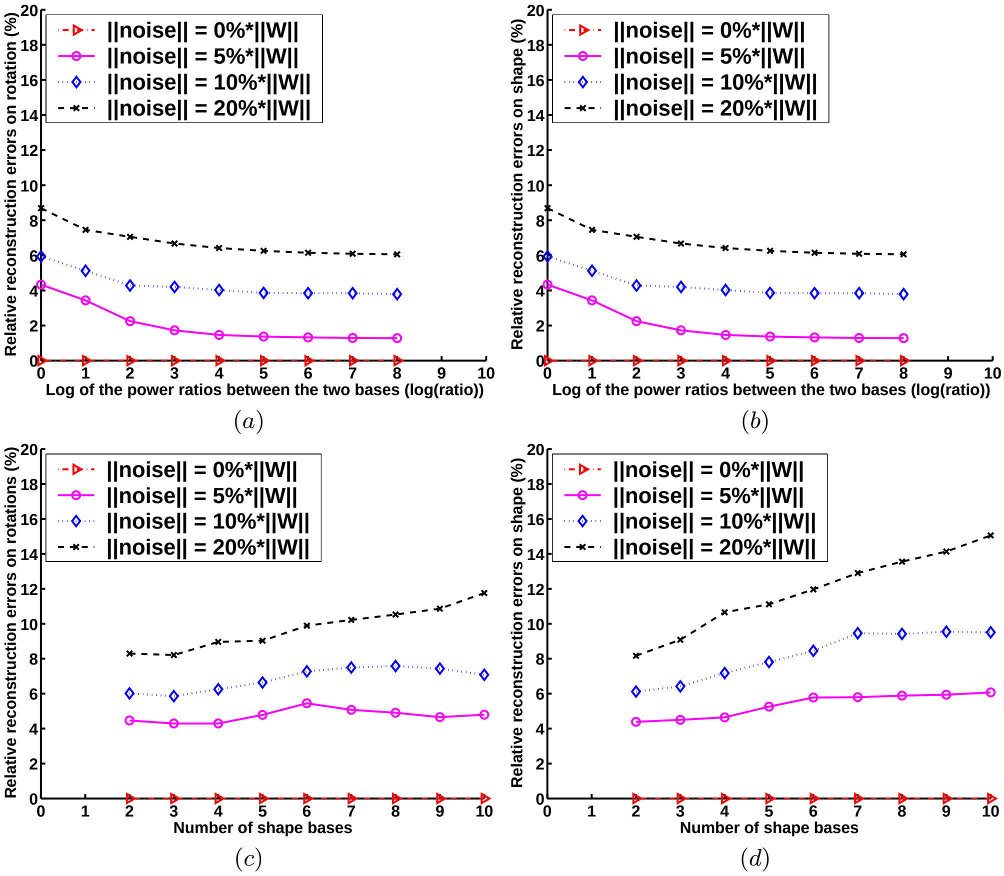
(a) (b)

(c) (d)

Fig. 2. The relative errors on reconstruction of a static cube and 3 points moving along straight lines. (a) By the closed-form solution. (b) By the method in [6]. (c) By the method in [4] after 4000 iterations. (d) By the trilinear method [14] after 4000 iterations. The scaling of the error axis is $[0\%, 100\%]$. Note that our method achieved zero reconstruction errors.

Fig.2 demonstrates the reconstruction errors of the four methods on camera rotations, shapes, and image measurements. The error was computed as the percentage relative to the ground truth, $\frac{\|Reconstruction-Truth\|}{\|Truth\|}$. Note that because the space of rotations is a manifold, a better error measurement for rotations is the Riemannian distance, $d(R_i, R_j) = acos(\frac{[trace(R_i R_j^T)-1]}{2})$. However it is measured in degrees. For consistency, we used the relative percentage for all the three reconstruction errors.

6.2 Quantitative Evaluation on Synthetic Data



(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. Each curve respectively refers to a noise level. The scaling of the error axis is $[0\%, 20\%]$.

Our approach was then quantitatively evaluated on the synthetic data. We evaluated 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 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 were respectively 2^0, 2^1, ..., and 2^8. Four levels of Gaussian white noise were imposed. Their strength levels were 0%, 5%, 10%, and 20% respectively. We tested a number of trials on each setting and computed the average reconstruction errors on the rotations and 3D shapes. The errors were measured by the relative percentage as in Section 6.1. Fig.3.(c) and (d) show the performance of our method under different numbers of shape bases and noise levels. The basis number was 2, 3, ... , and 10 respectively. The bases had equal powers and thus none of them was dominant. The same noise as in the last experiment was imposed.

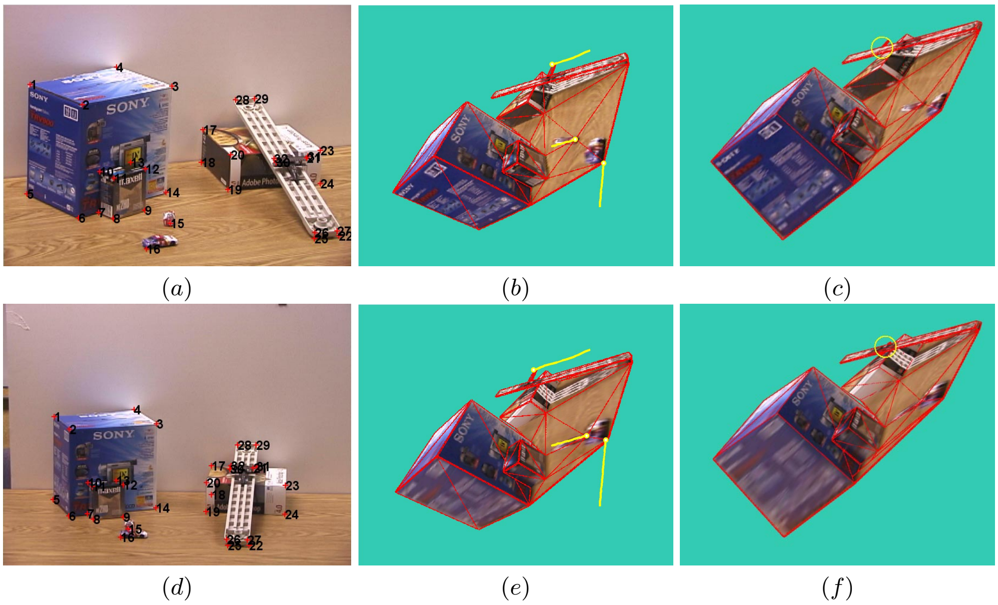
In both experiments, when the noise level was 0%, the closed-form solution recovered the exact rotations and shapes with zero error. When there was noise, it achieved reasonable accuracy, *e.g.* the maximum reconstruction error was less than 15% when the noise level was 20%. As we expected, under the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

6.3 Qualitative Evaluation on Real Video Sequences

Finally we examined 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 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 linear motions. 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 99% of the energy of $\tilde{W}$ could remain after the factorization using the rank constraint. The number of bases was thus determined by $K = \frac{rank(\tilde{W})}{3}$. Then the camera rotations and dynamic scene structures were reconstructed using the proposed method. With the recovered shapes, we could view the scene appearance from any novel directions. An example is shown in Fig.4.(b) and (e). The wireframes show the scene shapes and the yellow lines show the trajectories of the moving objects from the beginning of the sequence until the present frames. The reconstruction was 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 recovered shapes of the boxes were distorted and the plane was 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 appeared.

Human faces are highly non-rigid objects and 3D face shapes can be regarded as weighted combinations of certain shape bases that refer to various facial expressions. Thus our approach is capable of reconstructing the deformable 3D face shapes from the 2D image sequence. One example is shown in Fig.5. The sequence consisted of 236 images that contained facial expressions like eye blinking and mouth opening. 60 feature points were tracked using an



(*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 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, was mistakenly located underneath the slope.

efficient 2D Active Appearance Model (AAM) method [1]. Fig.5.(a) and (d) display two of the input images with marked feature points. After reconstructing the shapes and poses, we could view the 3D face appearances in any novel poses. Two examples are shown respectively in Fig.5.(b) and (e). Their corresponding 3D shape wireframes, as shown in Fig.5.(c) and (f), exhibit the recovered facial deformations such as mouth opening and eye closure. Note that the feature correspondences in these experiments were noisy, especially for those features on the sides of the face. The reconstruction performance of our approach demonstrates its robustness to the image noise.

7 Conclusion and Discussion

This paper proposes a linear closed-form solution to the problem of non-rigid shape and motion recovery from a single-camera video. In particular, we have proven that enforcing only the rotation constraints results in ambiguous and invalid solutions. We thus introduce the basis constraints to resolve this ambiguity. We have also proven that imposing both the linear constraints leads to a unique reconstruction of the non-rigid shape 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 deformations. A shape basis is degenerate, if its rank is either 1 or 2, *i.e.* it limits the shape deformation within a 2D plane. Degenerate deformations occur in some applications. For example, when a scene consists of several buildings and one car moving on a straight street, the shape basis referring to the rank-1 linear motion is degenerate. It is conceivable that, in such degenerate cases, the

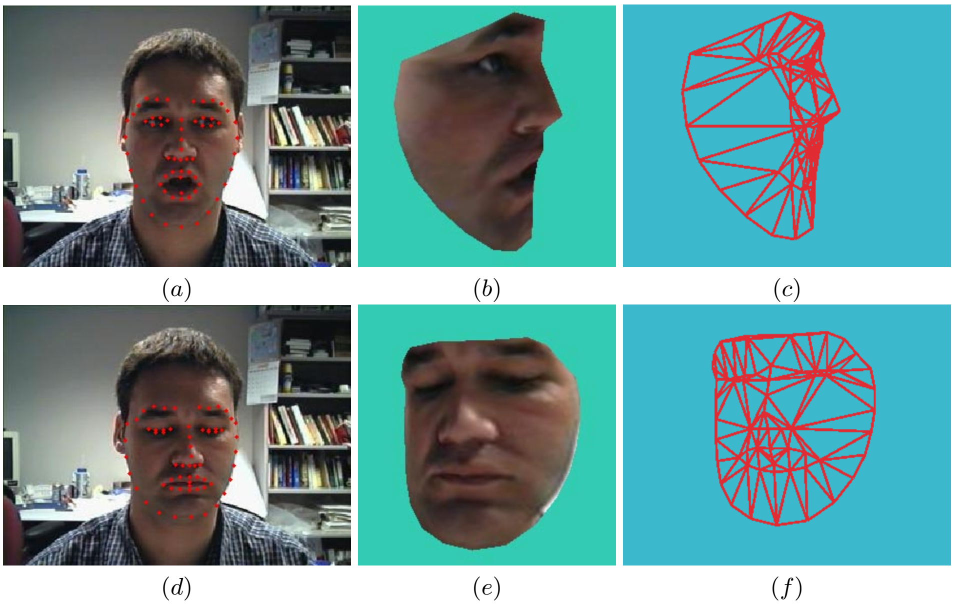


Fig. 5. Reconstruction of shapes of human faces carrying expressions. (*a*)&(*d*) Input images. (*b*)&(*e*) Reconstructed 3D face appearances in novel poses. (*c*)&(*f*) Shape wireframes demonstrating the recovered facial deformations such as mouth opening and eye closure.

basis constraints cannot completely resolve the ambiguity of the rotation constraints. We are now exploring how to extend the current method to reconstructing the shapes involving degenerate deformations. Another limitation of our approach is that we assume the weak perspective projection model. It would be interesting to see how the proposed solution could be extended to the full perspective projection model.

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. P. H. Schönemann, “ A generalized solution of the orthogonal procrustes problem,” *Psychometrika*, 31(1):1-10, 1966.
13. 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.
14. 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.
15. B. Triggs, “ Factorization methods for projective structure and motion,” *Proc. Int. Conf. Computer Vision and Pattern Recognition*,1996.
16. 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.
17. L. Wolf, A. Shashua, “ Two-body segmentation from two perspective views,” *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
18. 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.