Quasi-Perspective Projection with Applications to 3D Factorization from Uncalibrated Image Sequences

Guanghui Wang† Q. M. Jonathan Wu† Guoqiang Sun‡
† Department of Electrical and Computer Engineering, University of Windsor
401 Sunset, Windsor, ON, Canada, N9B 3P4.
‡ Department of Control Engineering, Aviation University, Changchun, 130022, China.
ghwangca@gmail.com; jwu@uwindsor.ca

Abstract

The paper addresses the problem of factorization-based 3D reconstruction from uncalibrated image sequences. We propose a quasi-perspective projection model and apply the model to structure and motion recovery of rigid and nonrigid objects based on factorization of tracking matrix. The novelty and contribution of the paper lies in three aspects. First, under the assumption that the camera is far away from the object with small rotations, we propose and prove that the imaging process can be modeled by quasi-perspective projection. The model is more accurate than affine since the projective depths are implicitly embedded. Second, we apply the model to the factorization algorithm and establish the framework of rigid and nonrigid factorization under quasi-perspective assumption. Third, we propose a new and robust method to recover the transformation matrix that upgrades the factorization to the Euclidean space. The proposed method is validated and evaluated on synthetic and real image sequences and good improvements over existing solutions are observed.

1. Introduction

The problem of structure and motion recovery from image sequences is an important theme in computer vision. Great progresses have been made for different applications during the last two decades [10]. The factorization method was first proposed by Tomasi and Kanade [17] in the early 90's. The main idea of this algorithm is to factorize the tracking matrix into motion and structure matrices simultaneously by singular value decomposition (SVD) with low-rank approximation. The algorithm assumes an orthographic projection model. It was extended to weak perspective and paraperspective projection by Poelman and Kanade [13]. In case of uncalibrated cameras, Quan [15] proposed a self-calibration algorithm for affine cameras.

More generally, Christy and Horaud [5] extended the above methods to perspective camera model by incrementally performing the factorization under affine assumption. The method is an affine approximation to general perspective projection. Triggs and Sturm [16, 20] proposed a full projective reconstruction method via rank-4 factorization of a scaled tracking matrix with projective depths recovered from pairwise epipolar geometry. The method was further studied in [9, 11, 12], where subspace constraints are embedded to recover the projective depths iteratively.

The above methods work only for rigid objects and static scenes. In order to deal with the scenarios of nonrigid or dynamic, many extensions stemming from the factorization algorithm were proposed to relax the rigidity constraint [1, 6]. In the pioneer work by Bregler *et al.* [4], it is demonstrated that the 3D shape of the nonrigid object may be expressed as a weighted linear combination of a set of shape bases. Then the shape bases, weighting coefficients and camera motions were factorized simultaneously under the rank constraint of the tracking matrix. Following this idea, the method was further investigated and developed by many researchers, such as Brand [2, 3], Del Bue *et al.* [7, 8], Torresani *et al.* [18, 19], Wang *et al.* [24] and Xiao *et al.* [25, 26].

Most nonrigid factorization methods are based on affine camera model due to its simplicity. It was extended to perspective projection in [22, 26] by iteratively recovering the projective depths. The perspective factorization is more complicated and there is no guarantee that it will converge to the correct depths, especially for nonrigid scenarios [10].

In this paper, we try to solve the problem under a novel framework. We assume that the camera is far away from the object with small rotations which is similar to affine assumptions and is easily satisfied in practice. We propose a quasi-perspective projection model under this assumption. The model is more accurate than affine camera model since the projective depths are implicitly embedded in the shape matrix. However, it is computationally as cheap as

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

affine. We apply the model to the factorization algorithm and present details on recovering the structure of rigid and nonrigid objects under this framework.

2. Background on Factorization

2.1. Problem definition

Under perspective projection, a 3D point in space $\mathbf{X}_j = [x_j, y_j, z_j, 1]^T$ is projected onto image $\mathbf{x}_{ij} = [u_{ij}, v_{ij}, 1]^T$ in the i-th frame according to the equation

$$\lambda_{ij}\mathbf{x}_{ij} = \mathbf{P}_i\mathbf{X}_j = \mathbf{K}_i[\mathbf{R}_i, \mathbf{T}_i]\mathbf{X}_j \tag{1}$$

where λ_{ij} is a non-zero scale factor, commonly called projective depth; $\mathbf{P}_i$ is the projection matrix; $\mathbf{K}_i$, $\mathbf{R}_i$ and $\mathbf{T}_i$ are the corresponding calibration matrix, rotation matrix and translation vector of the camera. When the distance of the object to the camera is much greater than the depth variation of the object, we may assume affine camera model. Then the imaging process can be simplified to $\bar{\mathbf{x}}_{ij} = \mathbf{A}_i\bar{\mathbf{X}}_j + \mathbf{t}_i$ by removing the scale factor, where $\bar{\mathbf{x}}_{ij}$ and $\bar{\mathbf{X}}_j$ are the non-homogeneous form of $\mathbf{x}_{ij}$ and $\mathbf{X}_j$; $\mathbf{A}_i$ is a 2×3 matrix; $\mathbf{t}_i$ is the image of world origin. It is easy to verify that the centroid of a set of space points is projected to the centroid of their images. Thus $\mathbf{t}_i$ will vanish if we register all image points to the corresponding centroid, and the projection is further simplified to

$$\bar{\mathbf{x}}_{ij} = \mathbf{A}_i\bar{\mathbf{X}}_j \tag{2}$$

The problem of structure from motion is defined as: Given n tracked feature points of an object across a sequence of m frames $\{\mathbf{x}_{ij}|i = 1, ..., m, j = 1, ..., n\}$. We want to recover the structure $\mathbf{S} = \{\mathbf{X}_j|j = 1, ..., n\}$ and motion $\{\mathbf{R}_i, \mathbf{T}_i\}$ of the object. The factorization based algorithm is proved to be an effective method to deal with this problem. According to the camera assumption and object property, the algorithm can be formulated as: (i) rigid object under affine assumption; (ii) rigid object under perspective projection; (iii) nonrigid object under affine assumption; (iv) nonrigid object under perspective projection.

2.2. Rigid factorization

Under affine assumption (2), the projection from space to the sequence is expressed as

$$\begin{bmatrix} \bar{\mathbf{x}}_{11} & \cdots & \bar{\mathbf{x}}_{1n} \\ \vdots & \ddots & \vdots \\ \bar{\mathbf{x}}_{m1} & \cdots & \bar{\mathbf{x}}_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_m \end{bmatrix} \begin{bmatrix} \bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n \end{bmatrix} \tag{3}$$

The equation can be written concisely as $\mathbf{W}_{2m\times n} = \mathbf{M}_{2m\times 3}\bar{\mathbf{S}}_{3\times n}$, where $\mathbf{W}$ is called the tracking matrix; $\mathbf{M}$ and $\bar{\mathbf{S}}$ are called the motion matrix and shape matrix respectively. It is evident that the rank of the tracking matrix is at most 3, and the rank constraint can be easily imposed by performing SVD on $\mathbf{W}$ and truncating it to rank 3. However, the decomposition is not unique since it is only defined up to a nonsingular linear transformation matrix $\mathbf{H}_{3\times 3}$ as $\mathbf{W} = (\mathbf{MH})(\mathbf{H}^{-1}\bar{\mathbf{S}})$. Actually, the decomposition is just one of the affine reconstructions of the object. By inserting $\mathbf{H}$ into the factorization, we can upgrade the reconstruction to the Euclidean space. Many researchers utilize the metric constraints of the motion matrix to recover the transformation [13, 15], which is indeed a self-calibration process with simplified camera parameters.

Under perspective projection (1), the factorization equation can be formulated as

$$\begin{bmatrix} \lambda_{11}\mathbf{x}_{11} & \cdots & \lambda_{1n}\mathbf{x}_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{m1}\mathbf{x}_{m1} & \cdots & \lambda_{mn}\mathbf{x}_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1 \\ \vdots \\ \mathbf{P}_m \end{bmatrix} \begin{bmatrix} \bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n \\ 1, \cdots, 1 \end{bmatrix} \tag{4}$$

or concisely as $\dot{\mathbf{W}}_{3m\times n} = \mathbf{M}_{3m\times 4}\mathbf{S}_{4\times n}$, where $\dot{\mathbf{W}}$ is called the scaled tracking matrix, and its rank is at most 4 if a consistent set of projective depths are present. Obviously, any such factorization corresponds to a valid projective reconstruction which is defined up to a projective transformation matrix $\mathbf{H}_{4\times 4}$. We can still use the metric constraint to recover the matrix.

The most difficult part for perspective factorization is to recover the projective depths that are consistent with (1). One method is to estimate the depths pairwisely from the fundamental matrix and then string them together [20]. The disadvantage of the method is the computational cost and possible error accumulation. The other method is to start with initial depths $\lambda_{ij} = 1$, and iteratively refine the depths by reprojections [9, 10]. However, there is no guarantee that the procedure will converge to a global minimum.

2.3. Nonrigid factorization

When the object is nonrigid, many studies assume that the nonrigid structure is approximated by a linear combination of k rigid shape bases as $\bar{\mathbf{S}}_i = \sum_{l=1}^{k} \omega_{il}\mathbf{B}_l$ [4], where $\mathbf{B}_l \in \mathbb{R}^{3\times n}$ is the shape base that embodies the principal mode of the deformation, $\omega_{il} \in \mathbb{R}$ is the deformation weight. Under this assumption and affine camera model, the nonrigid factorization is modeled as

$$\begin{bmatrix} \bar{\mathbf{x}}_{11} & \cdots & \bar{\mathbf{x}}_{1n} \\ \vdots & \ddots & \vdots \\ \bar{\mathbf{x}}_{m1} & \cdots & \bar{\mathbf{x}}_{mn} \end{bmatrix} = \begin{bmatrix} \omega_{11}\mathbf{A}_1 & \cdots & \omega_{1k}\mathbf{A}_1 \\ \vdots & \ddots & \vdots \\ \omega_{m1}\mathbf{A}_m & \cdots & \omega_{mk}\mathbf{A}_m \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_k \end{bmatrix} \tag{5}$$

It can be expressed in matrix form as $\mathbf{W}_{2m\times n} = \mathbf{M}_{2m\times 3k}\mathbf{B}_{3k\times n}$, where $\mathbf{M}$ and $\mathbf{B}$ are called the nonrigid motion and shape matrices. It is easy to see from (5) that the rank of $\mathbf{W}$ is at most $3k$. The decomposition can be achieved by SVD with the rank constraint, which is defined up to a nonsingular transformation matrix $\mathbf{H}_{3k\times 3k}$. If the transformation is known, $\mathbf{A}_i$, ω_{il} and $\bar{\mathbf{S}}_i$ can be recovered accordingly from $\mathbf{M}$ and $\mathbf{B}$. The computation of $\mathbf{H}$ is more

complicated than that in the rigid case. Many researchers [2, 8, 19] adopt the metric constraints of the motion matrix. However, the constraints may be insufficient when the object deforms at varying speed. Xiao *et al.* [25] propose a basis constraint to solve the ambiguity.

Similarly, the factorization under perspective projection can be formulated as [26]

$$\dot{\mathbf{W}}_{3m\times n} = \begin{bmatrix} \omega_{11}\mathbf{P}_1^{(1:3)} & \cdots & \omega_{1k}\mathbf{P}_1^{(1:3)} & \mathbf{P}_1^{(4)} \\ \vdots & \ddots & \vdots & \vdots \\ \omega_{m1}\mathbf{P}_m^{(1:3)} & \cdots & \omega_{mk}\mathbf{P}_m^{(1:3)} & \mathbf{P}_m^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_k \\ \mathbf{1} \end{bmatrix} \quad (6)$$

where $\mathbf{P}_i^{(1:3)}$ and $\mathbf{P}_i^{(4)}$ denote the first three and the fourth columns of $\mathbf{P}_i$ respectively. We denote (6) as $\dot{\mathbf{W}}_{3m\times n} = \mathbf{M}_{3m\times(3k+1)}\mathbf{S}_{(3k+1)\times n}$. The rank of the correctly scaled tracking matrix is at most $3k+1$. The decomposition is defined up to a transformation $\mathbf{H}_{(3k+1)\times(3k+1)}$, which can be determined in a similar while more complicated way. Just as in rigid case, the most difficult part is to determine the projective depths. Since there is no pairwise fundamental matrix for deformable features, we can only use the iterative method to recover the depth, but it is more likely to converge to a local minimum in nonrigid situation.

3. Quasi-Perspective Projection

We will propose a quasi-perspective projection model to approximate the imaging process more accurately.

Proposition 1 *Suppose the camera undergoes small rotations with respect to the scenario, then the variation of the projective depth λ_{ij} is mainly proportional to the depth of the space point, and the projective depth of a point at any view has the same trend of variation.*

Proof: Without loss of generality, let us set the world coordinate system on the object with the camera located in the Z direction of the world frame. Suppose the camera parameters corresponding to the i-th frame are $\mathbf{K}_i = \begin{bmatrix} f_i & \varsigma_i & u_{0i} \\ 0 & \kappa_i f_i & v_{0i} \\ 0 & 0 & 1 \end{bmatrix}$, $\mathbf{R}_i = [\mathbf{r}_{1i}, \mathbf{r}_{2i}, \mathbf{r}_{3i}]^T$ and $\mathbf{T}_i = [t_{xi}, t_{yi}, t_{zi}]^T$, respectively. Then the projection matrix can be written as

$$\begin{aligned} \mathbf{P}_i &= \mathbf{K}_i[\mathbf{R}_i, \mathbf{T}_i] \\ &= \begin{bmatrix} f_i\mathbf{r}_{1i}^T+\varsigma_i\mathbf{r}_{2i}^T+u_{0i}\mathbf{r}_{3i}^T & f_it_{xi}+\varsigma_it_{yi}+u_{0i}t_{zi} \\ \kappa_if_i\mathbf{r}_{2i}^T+v_{0i}\mathbf{r}_{3i}^T & \kappa_if_it_{yi}+v_{0i}t_{zi} \\ \mathbf{r}_{3i}^T & t_{zi} \end{bmatrix} \end{aligned} \quad (7)$$

Let us decompose the rotation matrix into the rotations around three axes $\mathbf{R}(\gamma_i)\mathbf{R}(\beta_i)\mathbf{R}(\alpha_i)$, where $\alpha_i, \beta_i, \gamma_i$ denote the rotation angles around the X, Y and Z axes, respectively. Then we have

$$\begin{aligned} \mathbf{R}_i &= \mathbf{R}(\gamma_i)\mathbf{R}(\beta_i)\mathbf{R}(\alpha_i) \\ &= \begin{bmatrix} \mathcal{C}\gamma_i & -\mathcal{S}\gamma_i & 0 \\ \mathcal{S}\gamma_i & \mathcal{C}\gamma_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathcal{C}\beta_i & 0 & \mathcal{S}\beta_i \\ 0 & 1 & 0 \\ -\mathcal{S}\beta_i & 0 & \mathcal{C}\beta_i \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathcal{C}\alpha_i & -\mathcal{S}\alpha_i \\ 0 & \mathcal{S}\alpha_i & \mathcal{C}\alpha_i \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{C}\gamma_i\mathcal{C}\beta_i & \mathcal{C}\gamma_i\mathcal{S}\beta_i\mathcal{S}\alpha_i-\mathcal{S}\gamma_i\mathcal{C}\alpha_i & \mathcal{C}\gamma_i\mathcal{S}\beta_i\mathcal{C}\alpha_i+\mathcal{S}\gamma_i\mathcal{S}\alpha_i \\ \mathcal{S}\gamma_i\mathcal{C}\beta_i & \mathcal{S}\gamma_i\mathcal{S}\beta_i\mathcal{S}\alpha_i+\mathcal{C}\gamma_i\mathcal{C}\alpha_i & \mathcal{S}\gamma_i\mathcal{S}\beta_i\mathcal{C}\alpha_i-\mathcal{C}\gamma_i\mathcal{S}\alpha_i \\ -\mathcal{S}\beta_i & \mathcal{C}\beta_i\mathcal{S}\alpha_i & \mathcal{C}\beta_i\mathcal{C}\alpha_i \end{bmatrix} \end{aligned} \quad (8)$$

where $'\mathcal{S}'$ stands for sine function, and $'\mathcal{C}'$ stands for cosine function. By inserting (7) and (8) into (1), we have

$$\begin{aligned} \lambda_{ij} &= [\mathbf{r}_{3i}^T, t_{zi}]\mathbf{X}_j \\ &= -(\mathcal{S}\beta_i)x_j + (\mathcal{C}\beta_i\mathcal{S}\alpha_i)y_j + (\mathcal{C}\beta_i\mathcal{C}\alpha_i)z_j + t_{zi} \end{aligned} \quad (9)$$

When the rotation angles are small, we have $\mathcal{S}\beta_i \ll \mathcal{C}\beta_i\mathcal{C}\alpha_i$ and $\mathcal{C}\beta_i\mathcal{S}\alpha_i \ll \mathcal{C}\beta_i\mathcal{C}\alpha_i$. Thus (9) can be approximated by

$$\lambda_{ij} \approx (\mathcal{C}\beta_i\mathcal{C}\alpha_i)z_j + t_{zi} \quad (10)$$

All the features $\{x_{ij}|j = 1, ..., n\}$ in the i-th frame correspond to the same rotation $\alpha_i, \beta_i, \gamma_i$ and translation t_{zi}. It is evident from (10) that the projective depths of a point in all frames have the same trend of variation, which are in proportion to the value of z_j of the space point. ■

Corollary 2 *Under Proposition 1, if we further assume that the distance of the camera to the object is greatly larger than the depth of the object, i.e. $t_{zi} >> z_j$, then the ratio of $\{\lambda_{ij}|i = 1, ..., m\}$ corresponding to any two different frames can be approximated by a constant.*

Proof: Let us take the first frame as a reference. Since $\mathcal{C}\beta_i\mathcal{C}\alpha_i \leq 1$ and $t_{zi} >> z_j$, then from

$$\begin{aligned} \mu_i = \frac{\lambda_{1j}}{\lambda_{ij}} &\approx \frac{(\mathcal{C}\beta_1\mathcal{C}\alpha_1)z_j + t_{z1}}{(\mathcal{C}\beta_i\mathcal{C}\alpha_i)z_j + t_{zi}} \\ &= \frac{\mathcal{C}\beta_1\mathcal{C}\alpha_1(z_j/t_{zi}) + t_{z1}/t_{zi}}{\mathcal{C}\beta_i\mathcal{C}\alpha_i(z_j/t_{zi}) + 1} \approx \frac{t_{z1}}{t_{zi}} \end{aligned} \quad (11)$$

we have $\lambda_{ij} = \frac{1}{\mu_i}\lambda_{1j}$ with $\mu_1 = 1$. ■

According to Corollary 2, the projection (1) can be written as $\frac{1}{\mu_i}\lambda_{1j}\mathbf{x}_{ij} = \mathbf{P}_i\mathbf{X}_j$. Let $\ell_j = \frac{1}{\lambda_{1j}}$, and replace $\mathbf{P}_i$ with $\mu_i\mathbf{P}_i$, and $\mathbf{X}_j$ with $\ell_j\mathbf{X}_j$, we have

$$\mathbf{x}_{ij} = (\mu_i\mathbf{P}_i)(\ell_j\mathbf{X}_j) \quad (12)$$

We call (12) the quasi-perspective projection. Compared with perspective projection, the quasi-perspective assumes that the projective depths between different frames are defined up to a constant μ_i. This is more general than affine model which assumes all projective depths equal to 1.

4. Quasi-Perspective Rigid Factorization

Under quasi-perspective projection (12), the factorization equation of the tracking matrix can be expressed as

$$\begin{bmatrix} \mathbf{x}_{11} & \cdots & \mathbf{x}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{m1} & \cdots & \mathbf{x}_{mn} \end{bmatrix} = \begin{bmatrix} \mu_1\mathbf{P}_1 \\ \vdots \\ \mu_m\mathbf{P}_m \end{bmatrix} [\ell_1\mathbf{X}_1, \cdots, \ell_n\mathbf{X}_n] \quad (13)$$

or written in short as $\mathbf{W}_{3m\times n} = \mathbf{M}_{3m\times 4}\mathbf{S}_{4\times n}$, which is similar to perspective factorization (4). However, the

projective depths in (13) are embedded in the motion and shape matrices, thus we do not need to estimate them explicitly. By performing SVD on the tracking matrix and imposing the rank-4 constraint, $\mathbf{W}$ may be factorized as $\hat{\mathbf{M}}_{3m\times 4}\hat{\mathbf{S}}_{4\times n}$. However, the decomposition is not unique since it is defined up to a nonsingular linear transformation $\mathbf{H}_{4\times 4}$ as $\mathbf{M}=\hat{\mathbf{M}}\mathbf{H}$ and $\mathbf{S}=\mathbf{H}^{-1}\hat{\mathbf{S}}$.

We adopt the metric constraint [9] to compute the transformation matrix. Let us denote $\mathbf{H}_{4\times 4}=[\mathbf{H}_l|\mathbf{H}_r]$, where $\mathbf{H}_l$ and $\mathbf{H}_r$ are the first three and the last columns of $\mathbf{H}$, respectively. Suppose $\hat{\mathbf{M}}_i$ is the i-th triple rows of $\hat{\mathbf{M}}$, then from $\hat{\mathbf{M}}_i\mathbf{H}=[\hat{\mathbf{M}}_i\mathbf{H}_l|\hat{\mathbf{M}}_i\mathbf{H}_r]$, we know that

$$\hat{\mathbf{M}}_i\mathbf{H}_l=\mu_i\mathbf{P}_i^{(1:3)}=\mu_i\mathbf{K}_i\mathbf{R}_i \tag{14}$$

$$\hat{\mathbf{M}}_i\mathbf{H}_r=\mu_i\mathbf{P}_i^{(4)}=\mu_i\mathbf{K}_i\mathbf{T}_i \tag{15}$$

Let us denote $\mathbf{C}_i=\hat{\mathbf{M}}_i\mathbf{Q}\hat{\mathbf{M}}_i^T$, where $\mathbf{Q}=\mathbf{H}_l\mathbf{H}_l^T$ is a 4×4 symmetric matrix. Suppose we adopt a simplified camera model with only one parameter as $\mathbf{K}_i=diag(f_i,f_i,1)$. This is a safe assumption for most digital cameras, and it was suggested that the principal points and aspect ratios are insignificant for reconstruction [14, 26]. Then from

$$\begin{aligned}\mathbf{C}_i&=\hat{\mathbf{M}}_i\mathbf{Q}\hat{\mathbf{M}}_i^T=(\mu_i\mathbf{K}_i\mathbf{R}_i)(\mu_i\mathbf{K}_i\mathbf{R}_i)^T\\&=\mu_i^2\mathbf{K}_i\mathbf{K}_i^T=\mu_i^2\begin{bmatrix}f_i^2&0&0\\0&f_i^2&0\\0&0&1\end{bmatrix}\end{aligned} \tag{16}$$

we can obtain the following constraints.

$$\mathbf{C}_i(1,1)=\mathbf{C}_i(2,2) \tag{17}$$

$$\mathbf{C}_i(1,2)=\mathbf{C}_i(1,3)=\mathbf{C}_i(2,3)=0 \tag{18}$$

Since the factorization (13) can be defined up to a global scalar as $\mathbf{W}=\mathbf{M}\mathbf{S}=(\varepsilon\mathbf{M})(\mathbf{S}/\varepsilon)$, we may set $\mu_1=1$ to avoid the trivial solution of $\mathbf{Q}=\mathbf{0}$. Thus we have $4m+1$ linear constraints in total on the 10 unknowns of $\mathbf{Q}$, which can be solved via least squares. Ideally, $\mathbf{Q}$ is a positive semidefinite symmetric matrix, then the matrix $\mathbf{H}_l$ can be recovered from the following proposition [23].

Proposition 3 *Suppose $\mathbf{Q}$ is a 4×4 positive semidefinite symmetric matrix of rank 3. Then it can be decomposed as $\mathbf{Q}=\mathbf{H}_l\mathbf{H}_l^T$, where $\mathbf{H}_l$ is a 4×3 rank 3 matrix. Furthermore, the decomposition can be uniquely written as $\mathbf{Q}=\mathbf{H}_d\mathbf{H}_d^T$ with $\mathbf{H}_d=\begin{bmatrix}h1&h2&h3\\h4&h5&h6\\&h7&h8\\&&h9\end{bmatrix}$.*

The proposition can be taken as an extension of Cholesky Decomposition to the case of positive semidefinite symmetric matrix. From the Proposition we know that $\mathbf{Q}$ is only defined with 9 degrees of freedom. In case of noise data, the recovered matrix $\mathbf{Q}$ may be negative definite, and fail to decompose into $\mathbf{H}_l\mathbf{H}_l^T$ or $\mathbf{H}_d\mathbf{H}_d^T$. In this case, we can substitute $\mathbf{Q}$ in (16) with $\mathbf{H}_d\mathbf{H}_d^T$ and solve the problem by minimizing the following cost function

$$\begin{aligned}f(\mathbf{h})=\min\sum_{i=1}^{m}\Big(&\mathbf{C}_i^2(1,2)+\mathbf{C}_i^2(1,3)+\mathbf{C}_i^2(2,3)\\&+(\mathbf{C}_i(1,1)-\mathbf{C}_i(2,2))^2\Big)\end{aligned} \tag{19}$$

where $\mathbf{h}$ is a 9-vector composed of the 9 elements in $\mathbf{H}_d$. The minimization scheme can be solved via any nonlinear optimization techniques.

We now show how to compute $\mathbf{H}_r$. From the quasi-perspective equation (12), we have

$$\mathbf{x}_{ij}=(\mu_i\mathbf{P}_i^{(1:3)})(\ell_j\bar{\mathbf{X}}_j)+(\mu_i\mathbf{P}_i^{(4)})\ell_j \tag{20}$$

For all the features in the i-th frame, their summation is

$$\sum_{j=1}^{n}\mathbf{x}_{ij}=\mu_i\mathbf{P}_i^{(1:3)}\sum_{j=1}^{n}(\ell_j\bar{\mathbf{X}}_j)+\mu_i\mathbf{P}_i^{(4)}\sum_{j=1}^{n}\ell_j \tag{21}$$

where $\mu_i\mathbf{P}_i^{(1:3)}$ is recovered from $\hat{\mathbf{M}}_i\mathbf{H}_l$, $\mu_i\mathbf{P}_i^{(4)}=\hat{\mathbf{M}}_i\mathbf{H}_r$. Since the world system can be chosen freely, we may set $\sum_{j=1}^{n}(\ell_j\bar{\mathbf{X}}_j)=0$, which is equivalent to set the origin of the world system at the gravity center of the scaled space points. On the other hand, since the reconstruction is defined up to a global scalar, we may simply set $\sum_{j=1}^{n}\ell_j=1$. Thus (21) is simplified to

$$\hat{\mathbf{M}}_i\mathbf{H}_r=\sum_{j=1}^{n}\mathbf{x}_{ij}=\begin{bmatrix}\sum_j u_{ij}\\\sum_j v_{ij}\\n\end{bmatrix} \tag{22}$$

which provides 3 linear constraints on the four unknowns of $\mathbf{H}_r$. Therefore, we can obtain $3m$ equations from the sequence and recover $\mathbf{H}_r$ via least squares. The solution of $\mathbf{H}_r$ is not unique as it is dependant on the selections of $\sum_{j=1}^{n}(\ell_j\bar{\mathbf{X}}_j)$ and $\sum_{j=1}^{n}\ell_j$. Actually, $\mathbf{H}_r$ may be set freely and we have the following proposition.

Proposition 4 *Suppose $\mathbf{H}_l$ is already recovered. Let us set the transformation matrix as $\tilde{\mathbf{H}}=[\mathbf{H}_l|\tilde{\mathbf{H}}_r]$, where $\tilde{\mathbf{H}}_r$ is any 4-vector that is independent with the three columns of $\mathbf{H}_l$. Then $\tilde{\mathbf{M}}=\hat{\mathbf{M}}\tilde{\mathbf{H}}$ and $\tilde{\mathbf{S}}=\tilde{\mathbf{H}}^{-1}\hat{\mathbf{S}}$ must correspond to a valid motion and shape matrices.*

The proof can be found in [23]. According to Proposition 4, the value of $\mathbf{H}_r$ can be set randomly as any 4-vector that is independent to $\mathbf{H}_l$. In practice, $\mathbf{H}_r$ may be set as follows. Suppose the SVD decomposition of $\mathbf{H}_l$ is $\mathbf{U}_{4\times 4}\Sigma_{4\times 3}\mathbf{V}_{3\times 3}$, where $\mathbf{U}$ and $\mathbf{V}$ are two orthogonal matrices, Σ is a diagonal matrix of the three singular values. Let us choose σ as any value between the biggest and the smallest singular values, then we may set $\mathbf{H}_r=\sigma\mathbf{U}^{(4)}$, where $\mathbf{U}^{(4)}$ is the last column of $\mathbf{U}$. Such construction guarantees that $\mathbf{H}$ is

invertible and has the same condition number as $\mathbf{H}_l$, such that we can obtain a good precision in computing $\mathbf{H}^{-1}$. After the correct motion and shape matrix are recovered, the camera parameters and pose that correspond to each frame can be recovered as follows.

$$\mu_i = \|\mathbf{M}_{i(3)}^{(1:3)}\| \tag{23}$$

$$f_i = \frac{1}{\mu_i}\|\mathbf{M}_{i(1)}^{(1:3)}\| = \frac{1}{\mu_i}\|\mathbf{M}_{i(2)}^{(1:3)}\| \tag{24}$$

$$\mathbf{R}_i = \frac{1}{\mu_i}\mathbf{K}_i^{-1}\mathbf{M}_i^{(1:3)},\ \mathbf{T}_i = \frac{1}{\mu_i}\mathbf{K}_i^{-1}\mathbf{M}_i^{(4)} \tag{25}$$

where $\mathbf{M}_{i(t)}^{(1:3)}$ denotes the t-th row of $\mathbf{M}_i^{(1:3)}$. The result is obtained under quasi-perspective assumption, which is a close approximation to the general perspective projection. The solution may be further optimized to perspective projection by minimizing the image reprojection residuals.

$$f(\mathbf{K}_i, \mathbf{R}_i, \mathbf{T}_i, \mu_i, \mathbf{X}_j) = \min \sum_{i=1}^{m}\sum_{j=1}^{n} |\mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij}|^2 \tag{26}$$

where $\hat{\mathbf{x}}_{ij}$ denotes the reprojected image point computed from perspective projection (1). The minimization process is termed as bundle adjustment, which can be solved via Levenberg-Marquardt iterations [10].

5. Quasi-Perspective Nonrigid Factorization

For nonrigid factorization, we still follow the assumption to represent the nonrigid shape by weighted combination of k shape bases. Under quasi-perspective projection, the structure is expressed in homogeneous form with nonzero scalars. Let $\mathbf{S}_i = \begin{bmatrix} \bar{\mathbf{S}}_i \\ \ell^T \end{bmatrix} = \begin{bmatrix} \ell_1\bar{\mathbf{X}}_1, & \cdots, & \ell_n\bar{\mathbf{X}}_n \\ \ell_1, & \cdots, & \ell_n \end{bmatrix}$ be the scale weighted structure associated with the i-th frame, $\mathbf{B}_l = [\ell_1\bar{\mathbf{X}}_1^l, \cdots, \ell_n\bar{\mathbf{X}}_n^l]$ be the l-th scale weighted shape basis. Then we can easily have the following result.

Proposition 5 *The scale weighted nonrigid structure can be expressed as linear combination of k scale weighted shape bases as $\mathbf{S}_i = \begin{bmatrix} \sum_{l=1}^{k}\omega_{il}\mathbf{B}_l \\ \ell^T \end{bmatrix}$, where $\ell^T = [\ell_1 \cdots \ell_n]$.*

Under proposition 5, the quasi-perspective projection of the i-th frame can be written as

$$\begin{aligned}\mathbf{W}_i &= (\mu_i\mathbf{P}_i)\mathbf{S}_i = [\mu_i\mathbf{P}_i^{(1:3)}, \mu_i\mathbf{P}_i^{(4)}]\begin{bmatrix} \sum_{l=1}^{k}\omega_{il}\mathbf{B}_l \\ \ell^T \end{bmatrix} \\ &= [\omega_{i1}\mu_i\mathbf{P}_i^{(1:3)}, \cdots, \omega_{ik}\mu_i\mathbf{P}_i^{(1:3)}, \mu_i\mathbf{P}_i^{(4)}]\begin{bmatrix} \mathbf{B}_1 \\ \cdots \\ \mathbf{B}_k \\ \ell^T \end{bmatrix}\end{aligned} \tag{27}$$

Thus the nonrigid factorization equation under quasi-perspective projection can be expressed as

$$\mathbf{W}_{3m\times n} = \begin{bmatrix} \omega_{11}\mu_1\mathbf{P}_1^{(1:3)} & \cdots & \omega_{1k}\mu_1\mathbf{P}_1^{(1:3)} & \mu_1\mathbf{P}_1^{(4)} \\ \vdots & \ddots & \vdots & \vdots \\ \omega_{m1}\mu_m\mathbf{P}_m^{(1:3)} & \cdots & \omega_{mk}\mu_m\mathbf{P}_m^{(1:3)} & \mu_m\mathbf{P}_m^{(4)} \end{bmatrix}\begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_k \\ \ell^T \end{bmatrix}$$

We can write the equation in matrix form as $\mathbf{W}_{3m\times n} = \mathbf{M}_{3m\times(3k+1)}\mathbf{B}_{(3k+1)\times n}$, which is similar to (6). However, the difficult problem of estimating the projective depths is avoided here. The rank of the tracking matrix is at most $3k+1$, and the factorization is again defined up to a transformation matrix $\mathbf{H}_{3k+1}$. Let us denote $\mathbf{H} = [\mathbf{H}_l, \mathbf{H}_r]$, where $\mathbf{H}_l$ and $\mathbf{H}_r$ are the first $3k$ and the last columns of $\mathbf{H}$. Then $\mathbf{H}_l$ can be recovered from the metric and the basis constraints as that in [26]. The Proposition 4 is still applicable to the nonrigid case except that $\mathbf{H}_r$ is a $(3k+1)$-vector here. Thus we can compute $\mathbf{H}_r$ in a similar way as that in the rigid case. Finally, the projection matrices and deformation weights can be easily recovered from $\mathbf{M}$ by Procrustes analysis [2, 19], and the structure corresponding to each frame is thus recovered from Proposition 5.

6. Experimental Evaluations

6.1. Evaluation on quasi-perspective projection

We randomly generated 50 points within a cube of $20 \times 20 \times 20$ in space, the X, Y and Z values of the points are shown in Fig.1(a). We simulated 10 images from these points by perspective projection. The image size is set at 800×800. The camera parameters are set as follows. The focal lengths are set randomly between 800 and 1200. The three rotation angles are set randomly between $\pm 5^\circ$. The X and Y positions of the cameras are set randomly between ± 15, while the Z positions are set evenly from 200 to 220. The true projective depths λ_{ij} associated with these points across the 10 views are shown in Fig.1(b1), where the values are given after normalization so that they have unit mean value. We then estimate λ_{1j} and μ_i from (10) and (11), and construct the estimated projective depths from $\hat{\lambda}_{ij} = \frac{\lambda_{1j}}{\mu_i}$. The registered result is shown in Fig.1(b2). We can see from the results that the recovered projective depths are similar to the ground truths, and are generally proportional to the variation of the space points in Z direction. If we adopt affine camera model, it is equivalent to setting all the projective depths to 1. The error is obvious.

Influence of different imaging conditions to the quasi-perspective assumption was also investigated. First, we fix the camera position and vary the amplitude of the rotation angles from $\pm 2^\circ$ to $\pm 38^\circ$ in steps of 4°. At each step, we check the relative error of the recovered projective depths, which is defined as $e_{ij} = \frac{|\lambda_{ij} - \hat{\lambda}_{ij}|}{\lambda_{ij}} \times 100(\%)$. We carried out 100 independent tests at each step so as to obtain more statistically meaningful results. The mean and standard deviation of e_{ij} are shown in Fig.1(c1). We then fix the rotation angles at $\pm 5^\circ$ and vary the relative distance of the camera to the object (i.e. the ratio between the distance of the camera to the object center and that of the object size) from 2 to 20 in steps of 2. The mean and standard deviation of e_{ij} at each step for 100 tests are shown in Fig.1(c2). The

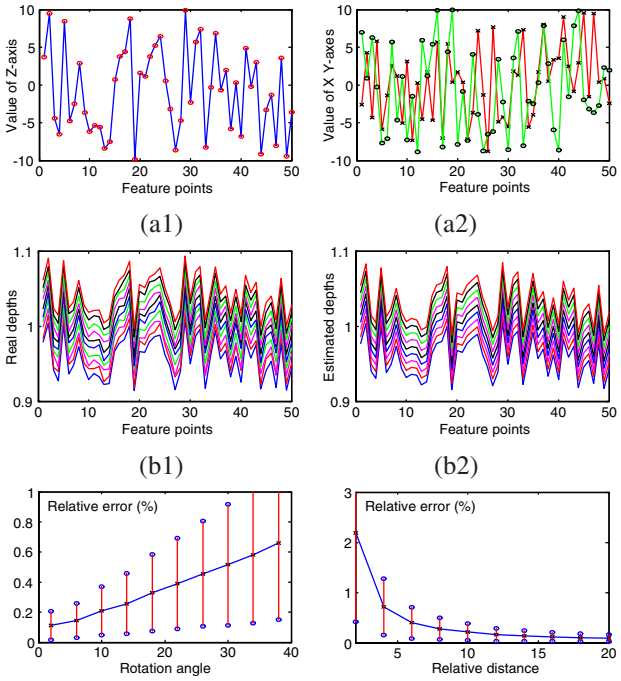


Figure 1. Evaluation on quasi-perspective projection. (a1)& (a2) Coordinates of the synthetic space points; (b1)& (b2) the real and the estimated projective depths; (c1)& (c2) the relative error of the estimated depths *vs.* rotation angle and relative distance.

results show that the quasi-perspective projection is a good approximation ($e_{ij} < 0.5\%$) when the rotation angles are less than $\pm 20^\circ$ and relative distance is larger than 8. Please note that the results assume noise free data.

6.2. Evaluation on rigid factorization

We add Gaussian white noise to the initially generated 10 frames, and vary the noise level from 0 to 3 pixels with a step of 0.5. At each noise level, we reconstructed the 3D structure of the object which is defined up to a similarity transformation with the ground truth. We register the reconstruction with the ground truth and calculate the mean pointwise distances as the reconstruction error. The mean and standard deviation of the error on 100 independent tests are shown in Fig.2. The proposed algorithm (Quasi) is compared with [13] under affine assumption (Affine) and [9] under perspective projection (Persp). We then take these solutions as initial values and perform the perspective optimization by LM iterations. It is evident that the proposed method performs much better than that of affine, the optimized solution (Quasi+LM) is very close to that of perspective projection with optimization (Persp+LM).

We compared the computation time of different algorithms. The program was implemented in Matlab 6.5 on an Intel Pentium 4 PC with 3.6GHz CPU. In this test, we select

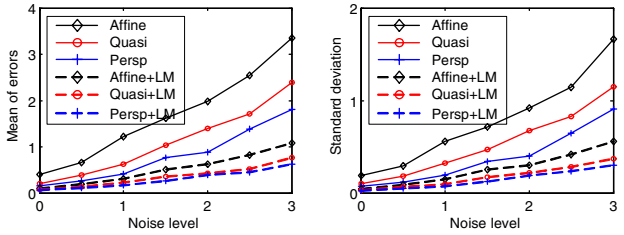


Figure 2. Evaluation on rigid factorization. The mean (a) and standard deviation (b) of the reconstruction errors by different algorithms *vs.* different noise levels.

Table 1. The average computation time of different algorithms.

Frame number		5	10	50	100	150	200
Time (s)	Quasi	0.015	0.016	0.047	0.156	0.297	0.531
	Affine	0.015	0.015	0.031	0.097	0.156	0.219
	Persp	0.281	0.547	3.250	6.828	10.58	15.25

200 space points and vary the frame number from 5 to 200. The real computation time (seconds) for different data set are listed in Table 1, where the time for perspective projection is taken at the 10th iteration. Clearly, the computation time of the proposed model is comparative to that of affine, while the perspective factorization is computationally more intensive than the other two models.

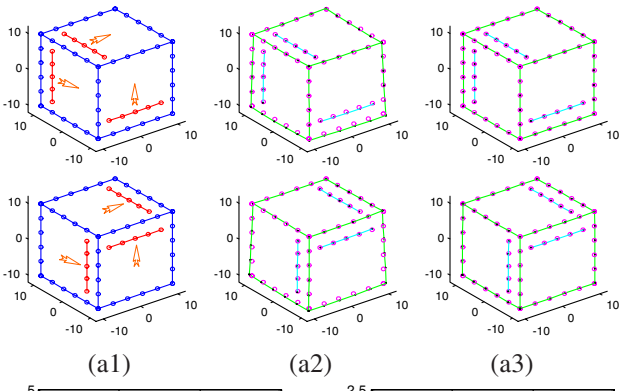
6.3. Evaluation on nonrigid factorization

We generated a synthetic cube with 7 evenly distributed points on each visible edge. There are three sets of moving points on the adjacent surfaces of the cube that move on the surfaces at constant speed as shown in Fig.3(a1). The object can be taken as nonrigid with 2 shape bases. We generated 10 frames with the same camera parameters as in the rigid case. We reconstructed the structure associated with each frame by the proposed method as shown in Fig.3(a2) and (a3). We can see that the structure after optimization is visually the same as the ground truth, while the result before optimization is a little bit deformed due to perspective effect. We compared the method with that under affine assumption [25] and that under perspective projection [26]. The reconstruction errors at different noise levels are shown in Fig.3(b), we have the same conclusion as in the rigid case that the proposed method performs better than that of affine.

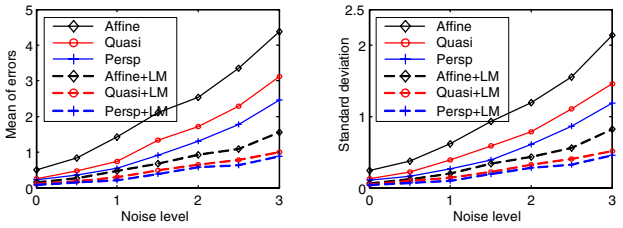
6.4. Evaluation on real image sequences

The method was tested on many real sequences. We will report three results here. Please refer to the supplemental videos for details of the test results.

The first test is on a post sequence with 8 images captured by Canon Powershot G3 camera. The image resolution is 1024×768. The post is a rigid object and we estab-



(a1) (a2) (a3)



(b1) (b2)

Figure 3. Evaluation on nonrigid factorization. (a1) Two synthetic cubes in space; (a2) the quasi-perspective factorization result of the two frames superimposed with the ground truth; (a3) the structure after optimization. (b1)&(b2) the mean and standard deviation of the reconstruction errors *vs.* noise level.

lished the correspondences by the system [21]. Totally 3693 reliable features were tracked across the sequence. Fig.4 shows the tracked features and the reconstructed 3D model from different viewpoints. The recovered structure is visually plausible and realistic.

The second test is on a grid sequence also captured by Canon G3 camera. There are 12 images with a resolution of 1024×768. The scenario is three objects moving linearly in three directions on an orthogonal background. We established 206 tracked features interactively across the sequence [24], where 140 features belong to the static background and 66 features belong to the three moving objects. Fig.5 shows the reconstructed VRML models and the corresponding triangulated wireframes of two frames by the proposed method. The dynamic structure of the scene is correctly recovered by the algorithm.

The background of this sequence is two orthogonal sheets with square grids. We take this as a ground truth and compute the angle (unit: degree) between the two reconstructed surfaces of the background, the length ratio of the two diagonals of each grid and the angle formed by the two diagonals. The mean errors of these three values are denoted by $\mathbf{E}_{\alpha 1}$, $\mathbf{E}_{rat}$ and $\mathbf{E}_{\alpha 2}$ respectively. We also calculated the relative reprojection error $\mathbf{E}_{rep}$. The comparative results obtained by the three algorithms are listed in Table 2. The proposed method performs better than that of affine, and is very close to that of perspective projection.

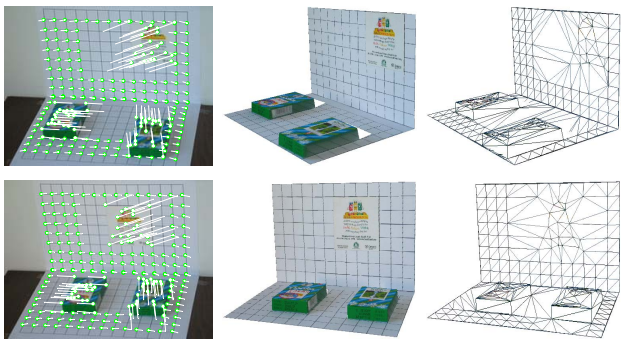
Table 2. Performance comparison on grid sequence.

Method	$\mathbf{E}_{\alpha 1}$	$\mathbf{E}_{\alpha 2}$	$\mathbf{E}_{rat}$	$\mathbf{E}_{rep}$
Affine/Affine+LM	2.35/0.96	0.92/0.37	0.15/0.07	5.66/2.25
Quasi/Quasi+LM	1.62/0.58	0.75/0.26	0.12/0.04	4.37/1.53
Persp/Persp+LM	1.28/0.52	0.63/0.24	0.10/0.04	3.64/1.46

The third test was on Franck sequence, which was downloaded from the European working group on face and gesture recognition (www-prima.inrialpes.fr/FGnet/). We select 60 frames with various facial expressions for the test. The image resolution is 720×576, and there are 68 tracked feature across the sequence. Fig.6 shows the reconstructed models of 2 frames by the propose method. Different facial expressions are correctly recovered, though some points are not very accurate due to tracking errors. The result could be used for visualization and recognition.

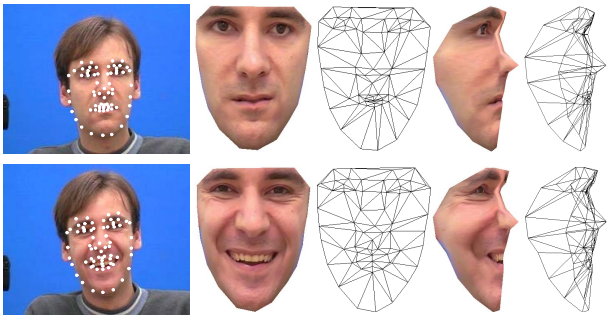
(a) (b)

Figure 4. Reconstruction result of the post sequence. (a) Two frames from the sequence overlaid with tracked features and the relative disparities shown in white lines; (b) the reconstructed VRML model and wireframe shown from different viewpoints.



(a) (b) (c)

Figure 5. Reconstruction result of the grid sequence. (a) Two frames from the sequence with tracked features, please note the three moving objects; (b)&(c) the reconstructed VRML models and wireframes of the two frames.



(a) (b) (c)

Figure 6. Reconstruction of different facial expressions in Franck sequence. (a) Two frames of the sequence overlaid with the tracked features; (b)&(c) the front and the side views of the reconstructed VRML models and the corresponding wireframes.

7. Conclusions

In this paper, under the assumption that the camera is far away from the object with small rotations, we proposed and proved a quasi-perspective projection model. We applied the model to rigid and nonrigid factorization and presented a new method to recover the transformation matrix. The proposed method is more accurate than that of affine, and it avoids the difficult problem of computing the projective depths in perspective factorization. Experiments demonstrated the advantages and improvements over previous methods. It should be noted that the assumption can usually be satisfied in many real applications. For long sequence taken around the object, we can simply divide the sequence into several subsequences with small movements, then register and merge the result of each subsequence to obtain the structure of the whole object.

Acknowledgment

Thanks to the reviewers for the valuable comments. The work is supported in part by the Canada Research Chair program, the NSERC Discovery Grant, and the National Natural Science Foundation of China under grant no. 60575015.

References

[1] B. Bascle and A. Blake. Separability of pose and expression in facial tracing and animation. In *ICCV*, 1998. 1

[2] M. Brand. Morphable 3D models from video. In *CVPR(2)*, pp. 456–463, 2001. 1, 3, 5

[3] M. Brand. A direct method for 3D factorization of nonrigid motion observed in 2D. In *CVPR(2)*, pp. 122–128, 2005. 1

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR(2)*, pp. 690–696, 2000. 1, 2

[5] S. Christy and R. Horaud. Euclidean shape and motion from multiple perspective views by affine iterations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(11):1098–1104, 1996. 1

[6] J. Costeira and T. Kanade. A multibody factorization method for independent moving objects. *IJCV*, 29(3):159–179, 98. 1

[7] A. Del Bue, X. Lladó, and L. de Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *CVPR(1)*, pp. 1191–1198, 2006. 1

[8] A. Del Bue, F. Smeraldi, and L. Agapito. Non-rigid structure from motion using nonparametric tracking and non-linear optimization. In *ANM*, pp. 8–15, 2004. 1, 3

[9] M. Han and T. Kanade. Creating 3D models with uncalibrated cameras. In *WACV*, pp. 178–185, 2000. 1, 2, 4, 6

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Second edition, 2004. 1, 2, 5

[11] A. Heyden, *et al.* An iterative factorization method for projective structure and motion from image sequences. *Image Vision Comput.*, 17(13):981–991, 1999. 1

[12] S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *CVPR(2)*, 430–437, 2000. 1

[13] C. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(3):206 – 218, 1997. 1, 2, 6

[14] M. Pollefeys, R. Koch, and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters. *IJCV*, 32(1):7–25, 1999. 4

[15] L. Quan. Self-calibration of an affine camera from multiple views. *IJCV*, 19(1):93–105, 1996. 1, 2

[16] P. F. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *ECCV(2)*, pp. 709–720, 1996. 1

[17] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992. 1

[18] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3D shape from 2D motion. In *NIPS*, 2004. 1

[19] L. Torresani, *et al.* Tracking and modeling non-rigid objects with rank constraints. In *CVPR*, pp. 493–500, 2001. 1, 3, 5

[20] B. Triggs. Factorization methods for projective structure and motion. In *CVPR*, pp. 845–851, 1996. 1, 2

[21] G. Wang. A hybrid system for feature matching based on SIFT and epipolar constraints. *Tech. Rep. Department of ECE, University of Windsor*, 2006. 7

[22] G. Wang, Y. Tian, and G. Sun. Modelling Nonrigid Object from Video Sequence Under Perspective Projection. In *ACII'05, LNCS*, 3784: 64–71, 2005. 1

[23] G. Wang and J. Wu. Quasi-perspective projection model with applications to structure and motion factorization of rigid and nonrigid objects. *Tech. Rep. University of Windsor*, 2007. 4

[24] G. Wang and J. Wu. Stratification Approach for 3-D Euclidean Reconstruction of Nonrigid Objects From Uncalibrated Image Sequences. *IEEE T-SMCB*, 38(1): 90–101, 2008. 1, 7

[25] J. Xiao, J.-X. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *ECCV(4)*, pp. 573–587, 2004. 1, 3, 6

[26] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *ICCV(2)*, pp. 1075–1082, 2005. 1, 3, 4, 5, 6