Uncalibrated Perspective Reconstruction of Deformable Structures

Jing Xiao
Epson Palo Alto Laboratory
Palo Alto, CA 94304
xiaoj@erd.epson.com

Takeo Kanade
Robotics Institute, CMU
Pittsburgh, PA 15213
tk@cs.cmu.edu

Abstract

Reconstruction of 3D structures from uncalibrated image sequences has a wealthy history. Most work has been focused on rigid objects or static scenes. This paper studies the problem of perspective reconstruction of deformable structures such as dynamic scenes from an uncalibrated image sequence. The task requires decomposing the image measurements into a composition of three factors: 3D deformable structures, rigid rotations and translations, and intrinsic camera parameters. We develop a factorization algorithm that consists of two steps. In the first step we recover the projective depths iteratively using the sub-space constraints embedded in the image measurements of the deformable structures. In the second step, we scale the image measurements by the reconstructed projective depths. We then extend the linear closed-form solution for weak-perspective reconstruction [23] to factorize the scaled measurements and simultaneously reconstruct the deformable shapes and underlying shape model, the rigid motions, and the varying camera parameters such as focal lengths. The accuracy and robustness of the proposed method is demonstrated quantitatively on synthetic data and qualitatively on real image sequences.

1. Introduction

Perspective reconstruction of 3D structure and motion from a 2D image sequence has been studied for decades. Great successes have been achieved for rigid objects or static scenes [15, 20, 18, 9, 25, 10]. A critique of various approaches is presented in [11]. One class of methods assumes calibrated or partially calibrated cameras and directly imposes metric constraints for perspective reconstruction [15, 20, 11]. In many applications, precise camera calibration is difficult to achieve. Such applications require simultaneous calibration of camera parameters and reconstruction of 3D structures. Hartley [8] and Faugeras [6] have shown that directly recovering all the structure and camera parameters leads to an extremely complicated non-linear optimization process, of which the performance greatly relies on the quality of the initial estimate.

The stratified approaches were introduced to avoid this difficulty [8, 6, 12, 2]. This class of approaches consists of two steps. The first step recovers the projective structure and motion from the image measurements, and the second step enforces the metric constraints to reconstruct the Euclidean structure and calibrate the camera parameters. A cluster of stratified methods uses the factorization technique to uniformly utilize all the available image measurements for reliable reconstruction [18, 14, 7, 9]. The factorization technique was originally introduced by Tomasi and Kanade for orthographic or weak-perspective reconstruction of rigid structures [16]. Triggs and Sturm [18, 14] presented that by scaling the image measurements with the projective depths, full perspective reconstruction can be achieved by the factorization technique. As the first step of their method, the projective depths were recovered using pairwise constraints among images. To recover the projective depths more reliably, the methods in [7, 9] utilized the subspace constraints embedded in the entire set of measurements.

In reality, many biological objects and natural scenes vary their shapes: expressive human faces, cars running beside buildings, etc. Perspective reconstruction of such non-rigid structures from images has much interest recently. Successful methods have been proposed for applications where individual objects contain sufficient sample points and independently rotate and translate [21, 19], or all objects move at constant velocities [22]. More generally, it has been shown that many non-rigid objects or dynamic scenes deform their shapes as linear combinations of certain shape bases [1, 23]. A number of methods have been proposed to recover such structures, assuming the weak-perspective camera model [4, 17, 3, 23]. When the camera is close or the scene is large in space, the assumption of weak-perspective projection no longer holds and the existing methods yield distorted reconstruction due to the perspective effect.

This paper presents a two-step factorization approach for perspective reconstruction of deformable structures composed by linear combinations of shape bases. As in the rigid cases, the first step recovers the projective depths. Our analysis shows that scaling the image measurements by the associated projective depths leads to a scaled measurement matrix of which the rank is determined by the number of the

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 IEEE

underlying bases. This subspace constraint is then used iteratively to recover the projective depths. In the second step, we factorize the scaled measurement matrix to reconstruct the 3D deformable shapes and rigid motions and simultaneously calibrate the camera focal lengths, by extending the linear closed-form solution for weak-perspective reconstruction [23]. The main extension is that this step recovers the varying camera focal lengths across images as well as the deformable structures and rigid motions, while the original work in [23] reconstructs only the latter. Since our method works for deformable structures and allows cameras to be unsynchronized and to freely vary the focal lengths, it provides a powerful tool for applications such as dynamic camera networks over large-scale dynamic scenes.

2 Recovery of Projective Depths

Suppose the structure consists of n points and their homogeneous coordinates across m perspective cameras are given. The j_{th} point is projected in the i_{th} image as follows,

$$U_{ij} = \frac{1}{d_{ij}} P_i X_{ij} \tag{1}$$

where $U_{ij} = (u_{ij}, v_{ij}, 1)^T$ and $X_{ij} = (x_{ij}, y_{ij}, z_{ij}, 1)^T$ are respectively the homogeneous image coordinate and 3D world coordinate of the j_{th} point in the i_{th} image. P_i is the 3×4 perspective projection matrix associated with the i_{th} image. $d_{ij} = P_{i(3)} X_{ij}$ is a non-zero scalar, where $P_{i(3)}$ denotes the third row of the projection matrix P_i. This projection representation is only determined up to an arbitrary 4×4 projective transformation Ω, *i.e.,* $U_{ij} = \frac{1}{d_{ij}}(P_i\Omega)(\Omega^{-1}X_{ij})$. But d_{ij} is independent of the choice of Ω, thus this parameter is commonly called projective depth.

2.1 Rank of the Scaled Measurement Matrix

Scaling the image coordinates by the corresponding projective depths and stacking them together as follows, we obtain the $3m \times n$ scaled measurement matrix,

$$W_s = \begin{pmatrix} d_{11}U_{11} & ... & d_{1n}U_{1n} \\ \vdots & \vdots & \vdots \\ d_{m1}U_{m1} & ... & d_{mn}U_{mn} \end{pmatrix} = \begin{pmatrix} P_1 S_1 \\ \vdots \\ P_m S_m \end{pmatrix} \tag{2}$$

where $S_i, i = 1, \ldots, m$ is a $4 \times n$ matrix that denotes the homogeneous world coordinates of the n 3D points in the i_{th} image. The first three rows of S_i refer to the 3D structure consisting of all points and the last row is a vector of all ones. Each column refers to a respective point. All the points in one image share a single projection matrix P_i.

For rigid structures, all the images share a single 3D structure, *i.e.,* $S_1 = \ldots = S_m$. The rank of W_s thus equals 4, the rank of the single structure [18, 14, 7, 9]. Accordingly the projective depths can be recovered using this rank constraint [7, 9]. When the structures are deformable and vary at different images, the rank-4 constraint in rigid cases no longer holds. Intuitively, to derive similar constraints for recovering the projective depths, we need to analyze what is shared by the deformable shapes across all the images.

It has been shown in [1, 4, 23] that the non-rigid objects, *e.g.,* expressive human faces, or dynamic scenes, *e.g.,* cars running on a straight road, often deform their structures as a linear combination of a set of shape bases, *i.e.,* $S_{i(1\sim3)} = \Sigma_{j=1}^{K} c_{ij} B_j, i = 1, \ldots, m$, where $S_{i(1\sim3)}$ denotes the first three rows of S_i, K is the number of bases, $B_j, j = 1, \ldots, K$ are the K $3 \times n$ shape bases, and c_{ij} are the corresponding combination weights. For any image number i, $P_i S_i = \Sigma_{j=1}^{K} c_{ij} P_i^{(1\sim3)} B_j + P_i^{(4)} \cdot \mathbf{1}$, where $P_i^{(1\sim3)}$ and $P_i^{(4)}$ denote the first three and the fourth columns of P_i respectively and $\mathbf{1}$ is a n-dimensional vector of all ones. Therefore what is shared by all the deformable shapes is the set of shape bases. We then rewrite Eq. (2) as follows,

$$W_s = \begin{pmatrix} c_{11}P_1^{(1\sim3)} & ... & c_{1K}P_1^{(1\sim3)} & P_1^{(4)} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m1}P_m^{(1\sim3)} & ... & c_{mK}P_m^{(1\sim3)} & P_m^{(4)} \end{pmatrix} \begin{pmatrix} B_1 \\ \vdots \\ B_K \\ \mathbf{1} \end{pmatrix} \tag{3}$$

We call the first matrix on the right side of Eq. (3) the scaled projection matrix and the second matrix the basis matrix, denoted as M and B, respectively a $3m \times (3K+1)$ matrix and a $(3K+1) \times n$ matrix. Under non-degenerate situations, both M and B are of full rank, respectively $min\{3m, 3K+1\}$ and $min\{3K+1, n\}$. Thus their product, W_s, is of rank $min\{3K+1, 3m, n\}$. In practice the image number m and point number n are usually much larger than the basis number K such that $3m > 3K+1$ and $n > 3K+1$. Thus the rank of W_s is $3K+1$ and the basis number K is determined by $\frac{rank(W_s)-1}{3}$. It is consistent with the previous conclusion for rigid cases ($K = 1$) that the rank is 4.

2.2 Iterative Projective Reconstruction

The rank of W_s has been used as the only constraint to recover the projective depths for rigid structures successfully [7, 9]. In deformable situations, assuming the basis number K is known, the constraint is nothing different except that the rank is now $3K+1$ instead of 4. Thus, similar to the rigid cases, we develop an iterative projective factorization algorithm. Its goal is to determine a set of projective depths d_{ij} that minimize $E = \|W_s - \hat{M}\hat{B}\|^2$, where $\hat{M}$ and $\hat{B}$ are respectively a $3m \times (3K+1)$ matrix and a $(3K+1) \times n$ matrix. As in [7, 9], the minimization is achieved by iteratively alternating two steps: estimating $\hat{M}$ and $\hat{B}$ given d_{ij} and updating d_{ij} given $\hat{M}$ and $\hat{B}$. The main difference

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

from the previous methods is that, we minimize E under the constraint that in alternative steps, the projective depths of **all points in any single image** or of **any single point in all images**, have unit norms such that minimization of E is simply an eigenvalue problem. Compared to the previous methods, this constraint works better in avoiding trivial solutions, *e.g.,* all the projective depths are set as zeros.

At initialization we set all the projective depths $d_{ij} = 1$, *i.e.,* we start with the weak-perspective approximation for the camera projection model. The scaled measurement matrix W_s is then computed according to Eq. (2). By singular value decomposition (SVD), we obtain the rank-$(3K+1)$ approximation $\hat{M}\hat{B}$ of W_s that minimizes E.

Let Φ_u be a $3m \times (3K+1)$ matrix whose columns are orthonormal and span the columns of $\hat{M}$ and Φ_v be a $(3K+1) \times n$ matrix whose rows are orthonormal and span the rows of $\hat{B}$. We then update the projective depths such that all columns of the updated W_s are spanned by the columns of Φ_u and all rows are spanned by the rows of Φ_v. Thus for any column $W_s^{(i)}$ and any row $W_{s(j)}$ of W_s, we need to minimize $\frac{\|\Phi_u \Phi_u^T W_s^{(i)} - W_s^{(i)}\|^2}{\|D^{(i)}\|^2}$ and $\frac{\|W_{s(j)} \Phi_v^T \Phi_v - W_{s(j)}\|^2}{\|D_{(j)}\|^2}$ respectively, where $D^{(i)}$ is an $m \times 1$ vector denoting the projective depths of the i_{th} point in all the m images. $D_{(j)}$ is a $1 \times n$ vector referring to the projective depths of all the n points in the j_{th} image. We normalize the minimization by the norm of $D^{(i)}$ and $D_{(j)}$ respectively so that the trivial solutions such as $D^{(i)} = 0$ or $D_{(j)} = 0$ are avoided. Substituting $W_s^{(i)}$ and $W_{s(j)}$ by Eq. (2), we rewrite the two minimization objectives as follows,

$$\min_{D^{(i)}} \frac{D^{(i)T} \Omega_{ui} D^{(i)}}{D^{(i)T} D^{(i)}}, \qquad \min_{D_{(j)}} \frac{D_{(j)} \Omega_{vj} D_{(j)}^T}{D_{(j)} D_{(j)}^T} \tag{4}$$

where Ω_{ui} is an $m \times m$ matrix and the entries $\Omega_{ui(kl)} = U_{ki}^T A_u^{(kl)} U_{li}$, $k,\ l = 1, \ldots, m$. $A_u = I - \Phi_u \Phi_u^T$ and is partitioned into 3×3 blocks $A_u^{(kl)}$, where I is an identity matrix. Ω_{vj} is an $n \times n$ matrix and the entries $\Omega_{vj(kl)} = A_{v(kl)} U_{jk}^T U_{jl}$, $k, l = 1, \ldots, n$. $A_v = I - \Phi_v^T \Phi_v$, consisting of $n \times n$ entries, $A_{v(kl)}$.

Because $D^{(i)}$ and $D_{(j)}$ have common elements, it is difficult to minimize the two objectives simultaneously, however each objective alone can be minimized by simply solving an eigenvalue problem, where the solution is the basis for null space of Ω_{ui} or Ω_{vj}, *i.e.,* the eigenvector associated with the smallest eigenvalue of Ω_{ui} or Ω_{vj}. We thus iteratively alternate these two minimizations to update the projective depths. Our algorithm is summarized as follows:

1) Set $d_{ij} = 1, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$;
2) Compute W_s by Eq. (2) and perform rank-$(3K+1)$ factorization on it by SVD to determine Φ_u;
3) Compute $A_u = I - \Phi_u \Phi_u^T$. For each of $D^{(i)}$, $i = 1, \ldots, n$, compute Ω_{ui} by $\Omega_{ui(kl)} = U_{ki}^T A_u^{(kl)} U_{li}$, $k,\ l = 1, \ldots, m$. Update $D^{(i)}$ with the eigenvector of Ω_{ui} associated with the smallest eigenvalue;
4) Compute W_s using the updated projective depths. Determine Φ_v by rank-$(3K+1)$ factorization on W_s;
5) Compute $A_v = I - \Phi_v^T \Phi_v$. For each of $D_{(j)}$, $j = 1, \ldots, m$, compute Ω_{vj} by $\Omega_{vj(kl)} = A_{v(kl)} U_{jk}^T U_{jl}$, $k,\ l = 1, \ldots, n$. Update $D_{(j)}^T$ with the eigenvector of Ω_{vj} associated with the smallest eigenvalue;
6) Stop if the difference between the estimated projective depths and those in the previous iteration is less than a preset small number. Otherwise go to Step 2).

It indeed avoided the trivial solutions in our extensive experiments. Note that this algorithm assumes that the basis number K is known. In cases where K is not given, we estimate it by $K = \frac{rank(W_s)-1}{3}$ whenever W_s is updated. The columns $D^{(i)}$ and rows $D_{(j)}$ are recovered up to scales, since projective reconstruction does not enforce any constraint directly on the structures and motions and scaling on $D^{(i)}$ or $D_{(j)}$ does not alter the rank of W_s. Thus the corresponding scaled projection matrix and basis matrix are scaled as follows,

$$M = \begin{pmatrix} \lambda_1 c_{11} P_1^{(1\sim3)} & \ldots & \lambda_1 c_{1K} P_1^{(1\sim3)} & \lambda_1 P_1^{(4)} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_m c_{m1} P_m^{(1\sim3)} & \ldots & \lambda_m c_{mK} P_m^{(1\sim3)} & \lambda_m P_m^{(4)} \end{pmatrix}$$
$$B = \begin{pmatrix} \gamma_1 B_1^1 & \ldots & \gamma_n B_1^n \\ \vdots & \vdots & \vdots \\ \gamma_1 B_K^1 & \ldots & \gamma_n B_K^n \\ \gamma_1 & \ldots & \gamma_n \end{pmatrix} \tag{5}$$

where $\lambda_1, \ldots, \lambda_m$ are the scalars for $D_{(1)}, \ldots, D_{(m)}$ and $\gamma_1, \ldots, \gamma_n$ are the scalars for $D^{(1)}, \ldots, D^{(n)}$ respectively. For simplicity, we keep the notations W_s, M, and B.

3 Perspective Reconstruction

Given W_s, we compute its rank-$(3K+1)$ approximation $\hat{M}\hat{B}$ by SVD. This decomposition is not unique. Any non-singular $(3K+1) \times (3K+1)$ matrix could be inserted between $\hat{M}$ and $\hat{B}$ to obtain a new eligible factorization. Thus $W_s = \hat{M}\hat{B} = \hat{M}GG^{-1}\hat{B} = MB$, where the non-singular $(3K+1) \times (3K+1)$ matrix G is called the corrective transformation. Once G is determined, we obtain the true scaled projection matrix $M = \hat{M}G$ and the true basis matrix $B = G^{-1}\hat{B}$. We then impose metric constraints on M and B to recover the deformable shapes, rigid rotations and translations, and intrinsic camera parameters.

The perspective projection matrix $P_i \sim \Lambda_i(R_i|T_i)$, where R_i is the 3×3 orthonormal rotation matrix, T_i is the 3×1 translation vector, and Λ_i is the 3×3 camera matrix

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

as follows,

$$\Lambda_i = \begin{pmatrix} f_i & \mu_i & u_{0i} \\ 0 & \alpha_i f_i & v_{0i} \\ 0 & 0 & 1 \end{pmatrix} \quad (6)$$

where f_i is the focal length, α_i is the aspect ratio, μ_i is the skew parameter, and (u_{0i}, v_{0i}) is the principle point. In practice the skews are usually assumed as zeros. Theoretically we can calibrate all the other parameters from W_s. However, as discussed in [12, 2, 11] and also observed in our experiments, the principle points and aspect ratios are insignificant for perspective reconstruction and their estimates are highly unreliable, and generally the principle points are close to the image centers and the aspect ratios are close to 1. As pointed out in [11], such information should be used and even an approximation of these parameters helps achieve better reconstruction than treating them as free variables. Thus we move the image origins at the image centers and set $\mu_i = 0$, $\alpha_i = 1$, and $(u_{0i}, v_{0i}) = (0, 0)$.

Denote G as $(g_1, \ldots, g_K, g_L)$, where $g_k,\ k = 1, \ldots, K$ are triple columns of G and g_L is the last column. They are independent on each other because G is non-singular. Denoting $\hat{M}_i,\ i = 1, \ldots, m$, the m triple rows of $\hat{M}$, due to Eq. (5), g_k and g_L satisfy,

$$\hat{M}_i g_k = \lambda_i c_{ik} \begin{pmatrix} f_i R_{i(1)} \\ f_i R_{i(2)} \\ R_{i(3)} \end{pmatrix}, \quad \hat{M}_i g_L = \lambda_i \begin{pmatrix} f_i T_{i(1)} \\ f_i T_{i(2)} \\ T_{i(3)} \end{pmatrix} \quad (7)$$

Let us first recover $g_k,\ k = 1, \ldots, K$, respectively. Denoting $Q_k = g_k g_k^T$, we have,

$$\hat{M}_i Q_k \hat{M}_j^T = \alpha_{ijk} \begin{pmatrix} f_i f_j R_{i(1,2)} R_{j(1,2)}^T & f_i R_{i(1,2)} R_{j(3)}^T \\ f_j R_{i(3)} R_{j(1,2)}^T & R_{i(3)} R_{j(3)}^T \end{pmatrix} \quad (8)$$

where $i, j = 1, \ldots, m$, and $\alpha_{ijk} = \lambda_i \lambda_j c_{ik} c_{jk}$. As in [16, 4, 3, 23], we enforce the orthonormality constraints on the rotation matrices and obtain,

$$\hat{M}_i Q_k \hat{M}_i^T = \lambda_i^2 c_{ik}^2 \begin{pmatrix} f_i^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \quad (9)$$

where I denotes a 2×2 identity matrix. Because λ_i, c_{ik}, and f_i are all unknowns, the 3 diagonal entries of Eq. (9) yield 1 linear constraint on Q_k, *i.e.,* the equivalence of the first and second diagonal entries. Due to symmetry of Q_k, the 6 off-diagonal elements provide 3 linear constraints and the other half are redundant. Thus, for m images, we have $4m$ linear constraints on Q_k, due to orthonormality of rotations.

Following the idea in [23], we prove that enforcing the orthonormality constraints alone is insufficient to determine Q_k. To eliminate the ambiguity, we impose the basis constraints that uniquely specify the shape bases, as in [23]. Firstly we select K images, with which the associated scaled measurement matrix, *i.e.,* the corresponding $3K \times n$ sub-matrix in W_s, has a small condition number. A small (close to 1) condition number means these images contain independent structures and none of them is dominant. We then regard the K included shapes as the bases. Note that we have not recovered the explicit values of the bases, but determined in which images they are observed. Denoting these images as the first K observations, the corresponding combination weights are,

$$\begin{aligned} c_{ii} &= 1,\ i = 1, ..., K \\ c_{ij} &= 0,\ i, j = 1, ..., K,\ i \neq j \end{aligned} \quad (10)$$

Combining Eq. (8) and (10), we have,

$$\hat{M}_i Q_k \hat{M}_j^T = \mathbf{0}, \quad i = 1, ..., K, i \neq k, j = 1, ..., m. \quad (11)$$

where $\mathbf{0}$ means a 3×3 zero matrix. It leads to $9m(K-1)$ linear constraints on Q_k, including a small number of redundant ones due to symmetry of Q_k. Combining with the orthonormality constraints, we have totally $(9K-5)m$ linear constraints.

For weak-perspective reconstruction, the translations are estimated and eliminated from measurements before factorization, and then enforcing the uniqueness together with the orthonormality constraints leads to a linear closed-form solution [23]. In the perspective cases, Q_k contains more unknowns because the unknown translations are involved in the factorization step. However, the last row of B corresponding to the translations can be regarded as a rank-1 degenerate shape basis [24]. Then our problem resembles the factorization problem for weak-perspective reconstruction of shapes composed of K rank-3 and one rank-1 bases. As presented in [24], the problem has a unique solution by enforcing both the uniqueness and the orthonormality constraints. Because Eq. (9) and (11) contain not only constraints similar to those in weak-perspective cases but also many extra ones, totally $(9K-5)m$ vs $(4K-2)m$, Q_k is also determined uniquely, by solving the linear equations in Eq. (9) and (11). For more detailed analysis on the constraints and solutions, please refer to [23, 24].

Because Q_k equals $g_k g_k^T$, by SVD we can compute its rank-3 decomposition, $U_k S_k U_k^T$, where U_k and S_k are respectively $(3K+1) \times 3$ and 3×3 matrices. Then g_k is recovered as $U_k S_k^{\frac{1}{2}}$. Note that g_k is determined only up to an arbitrary 3×3 orthonormal transformation Ψ_k, *i.e.,* $Q_k = (g_k \Psi_k)(g_k \Psi_k)^T$.

Let us denote $M_i^k = \hat{M}_i g_k$. According to Eq. (7), we recover the focal length, scaled weights, and rotations as follows,

$$f_i = \frac{|M_{i(1)}^k|}{|M_{i(3)}^k|} = \frac{|M_{i(2)}^k|}{|M_{i(3)}^k|} \quad (12)$$

$$\lambda_i c_{ik} = \pm |M_{i(3)}^k| \quad (13)$$

$$R_i = \begin{pmatrix} M_{i(1,2)}^k / (\lambda_i c_{ik} f_i) \\ M_{i(3)}^k / (\lambda_i c_{ik}) \end{pmatrix} \quad (14)$$

where $M_{i(j)}^k,\ j = 1, 2, 3$, refer to the rows of M_i^k. For each of $g_k,\ k = 1, \ldots, K$, a full set of rotations for all images

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

are computed and they are also determined up to the 3×3 orthonormal transformation Ψ_k. Thus, between each two of the rotation sets, there is a 3×3 orthonormal transformation. According to Eq. (13), the sign of $\lambda_i c_{ik}$ is also undetermined. To resolve these ambiguities, we specify one of the rotation sets as the reference. The orthonormal transformations between the other sets and the reference one are computed by Orthogonal Procrustes Analysis (OPA) [13]. They transform $g_1, \ldots, g_K$ to be under a common coordinate system and eliminate the ambiguity. The signs of $\lambda_i c_{ik}$ is determined in such a way that they are consistent across all the rotation sets.

We now determine the last column g_L of G. Let us set the origin of the world coordinate system at the center of the scaled 3D structure $S_{i(1\sim3)}$. Then,

$$\bar{S}_{i(1\sim3)} = \Sigma_{j=1}^{K} \lambda_i c_{ij} \bar{B}_j = 0 \tag{15}$$

where $\bar{S}_{i(1\sim3)}$ and $\bar{B}_j$ denote the center (mean) of the 3D coordinates in $S_{i(1\sim3)}$ and B_j. B_j is the j_{th} scaled shape basis, *i.e.,* j_{th} triple rows of B in Eq. (5). We then have,

$$\bar{W}_{si} = P_i^{(1\sim3)} \Sigma_{j=1}^{K} \lambda_i c_{ij} \bar{B}_j + \lambda_i P_i^{(4)} \bar{\Gamma} = \lambda_i P_i^{(4)} \bar{\Gamma} \tag{16}$$

where W_{si} means the scaled measurements in the i_{th} image, *i.e.,* the i_{th} triple rows of W_s. Γ denotes the last row of B and $\bar{\Gamma}$ is a constant. Since uniformly scaling B does not violate the factorization of W_s, *i.e.,* $W_s = MB = (\bar{\Gamma} M)(B/\bar{\Gamma})$, we set $\bar{\Gamma} = 1$. Due to Eq. (7,16), we have,

$$\hat{M} g_L = \bar{W}_s \tag{17}$$

where $\bar{W}_s$ is the mean of the columns of W_s. Thus g_L is determined by $g_L = \hat{M}^+ \bar{W}_s$, where $\hat{M}^+$ means the pseudoinverse of $\hat{M}$. The scaled translations $\lambda_i T_i$ are computed by Eq. (7). The scaled projection matrix M in Eq. (5) is now complete. The scaled basis matrix B is determined by $M^+ W_s$. The last row of B contains the scalars $\gamma_1, \ldots, \gamma_n$. We thus normalize the columns of B respectively by the last elements to obtain the true shape bases in Eq. (3).

Combining the shape bases by the scaled weights, we obtain the deformable structures up to scalars, $S_{i(1\sim3)} = \lambda_i \Sigma_{j=1}^{K} c_{ij} B_j$. As shown above, the translations are recovered up to the same scalars, $\lambda_i T_i$. This scaling ambiguity is inherent since according to Eq. (1) and (3), scaling the shape and translation simultaneously does not vary the image measurements. We need one reference and align all the shapes with it to eliminate the scaling ambiguity. Such reference can be any segment connecting two static points. When all points are allowed to deform, we set the first of the recovered shapes as the reference. The other shapes are then aligned with it by Extended Orthogonal Procrustes Analysis (EOPA) [13], where only the scalars are unknowns.

Given W_s, our perspective reconstruction algorithm is summarized as follows,

1) Determine the K basis images and compute rank-$(3K+1)$ approximation of $W_s = \hat{M}\hat{B}$ by SVD;
2) Determine Q_k, $k = 1, \ldots, K$, respectively by solving Eq. (9, 11) via the linear least square method and then compute g_k by SVD.
3) For each of g_k, compute f_i, $\lambda_i c_{ik}$, and R_i, $i = 1, \ldots, m$, by Eq. (12, 13, 14) and transform them to a common coordinate system by OPA;
4) Compute $g_L = \hat{M}^+ \bar{W}_s$ and $\lambda_i T_i = \hat{M}_i g_L$ that complete M;
5) Compute $B = M^+ W_s$ and normalize its columns respectively by their last elements;
6) Align the shapes by EOPA to eliminate the scaling ambiguity on reconstruction of shapes and translations.

This algorithm leads to a linear closed-form solution. Note that it is not necessary to check all possible K images to specify the most independent shapes as the bases. We only need to find K images with which the condition number of the associated scaled measurements is small enough below certain threshold, since the point is to specify any set of independent shapes.

4 Performance Evaluation

In this section we demonstrate the performance of the proposed approach quantitatively on synthetic data and qualitatively on real image sequences.

4.1 Quantitative Evaluation on Synthetic Data

The accuracy and robustness of our method is first evaluated with respect to different measurement noise levels and basis numbers, using synthetic data respectively in rigid settings involving 1 shape basis and non-rigid settings involving 2, 3, ... , and 6 bases.

In each setting the 3D bases are randomly generated and normalized ($\|B_i\| = 1$). The combination weights are randomly generated in such a way that the weights for different bases have the same order of magnitude, *i.e.,* none of the bases is dominant. The deformable structures are constructed and projected by the cameras placed randomly around the scene from left (-45°) and upper (30°) to right (45°) and bottom (-30°), and at distances ranging from 1 to 3 times of the maximum inner shape depths. The focal lengths are randomly selected between 1000 and 2000. Assuming a Gaussian white noise, we represent the strength of the measurement noise by the ratio between the Frobenius norm of the noise and the image measurements, *i.e.,*

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

$\frac{\|noise\|}{\|\{U_{ij}\}\|}$, where $\{U_{ij}\}$ means the collection of all measurements. For each basis setting we examine the approach under 4 noise levels, 0%, 5%, 10%, and 15%.

Given the measurements, we first recover the projective depths using the iterative projective reconstruction algorithm in Section 2.2, and construct the scaled measurement matrix W_s. The rank-$(3K+1)$ approximation of W_s is projected back to the images for evaluation. The average re-projection errors relative to the ground truth at each setting are shown in Fig.1. No matter how many bases are involved, for noiseless settings (0%), the projective depths are precisely recovered. The error increases when the noise gets stronger. Because the bases equally contributed to the shape composition, the bigger the basis number, the more "non-rigid" the shapes and the stronger the noise relative to each individual basis. Thus the error also increases when the basis number is bigger. Yet our method achieves reasonable accuracy in all noise settings, *e.g.,* in the worst case of 6 bases and 15% noise level, the re-projection error is about 13%. Note that the initial step of the method assumes the weak-perspective model and yields significant errors, since the cameras are not far from the scene.

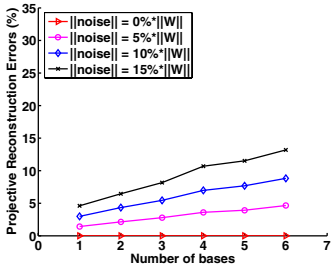

Figure 1: Projective reconstruction errors versus noise and basis numbers. Bigger basis numbers and stronger noise lead to greater errors.

We then factorize W_s and reconstruct the 3D deformable shapes, rigid rotations and translations, and focal lengths using the perspective reconstruction algorithm in Section 3. The average reconstruction errors on deformable shapes and rigid rotations are shown in Fig.2. The errors on rigid rotations are measured as the Riemannian distance in degrees, *i.e.,* $e(R_{est}, R_{truth}) = arccos(\frac{trace(R_{est}R_{truth}^T)}{3})$, because the space of rotations is a manifold. The errors on shapes are computed as the relative percentage with respect to the ground truth, *i.e.,* $e(S_{est}, S_{truth}) = \frac{\|S_{est}-S_{truth}\|}{\|S_{truth}\|}$.

As shown in Fig.2, our approach yields precise perspective reconstruction for noiseless settings and the reconstruction error increases when the noise gets stronger and the basis number is bigger. In all noise settings, reasonable accuracy is achieved, *e.g.,* in the worst case of 6 bases and 15% noise level, the average error on shapes is about 14% and on rotations about 7.5 degrees. Our method achieves similar accuracy on translations and focal lengths as on shapes, when the noise is weak (0%, 5%). When the noise is strong, their estimates in some images are unreliable, especially when not much perspective distortion is present in those images. One possible explanation is that the focal length and translation in depth are not shared by all images and tend to cancel each other in perspective projection. On the other hand, their estimates only have a small effect on the reconstruction of shapes and rotations [2, 12].

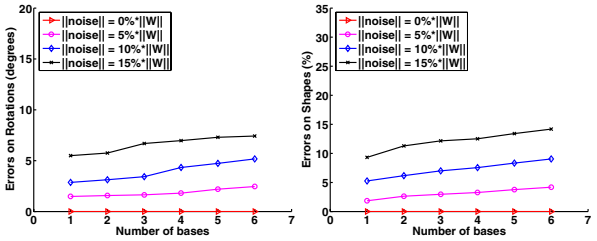

Figure 2: Perspective reconstruction errors on rotations and shapes versus noise and basis numbers. Bigger basis numbers and stronger noise lead to greater errors.

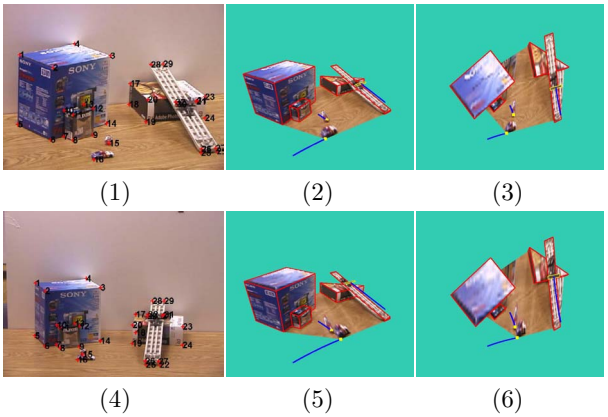

Figure 3: Perspective reconstruction of dynamic scene structures. (1)&(4) Two input images. (2)&(5) Side view of the scene appearance recovered by our method. (3)&(6) Top view of the reconstructed scene appearance. The red and yellow wireframes mean the static and moving objects respectively, and the blue lines show the moving trajectories. The straight motions and the rectangle shape of the box top are recovered correctly.

4.2 Qualitative Evaluation on Real Image Sequences

Reconstruction of the 3D deformable structures of dynamic scenes from sequences of 2D images is important for tasks like robot navigation and visual surveillance. Such scenes

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

often consist of both static objects, *e.g.,* buildings, and objects moving straight, *e.g.,* vehicles running or pedestrians walking on the road. The scene structures are linear combinations of two classes of shape bases: static points and linear trajectories of moving points. Our approach is thus capable of perspective reconstruction of such dynamic scene structures from the associated image sequences.

One example is shown in Fig.3. The scene contains three objects moving along respective directions simultaneously, two on top of the table and one along the slope. The rest of the scene are static. The scene structure is thus composed of two bases, one representing the static objects and initial locations of the moving objects and another representing the three linear motion vectors. Eighteen images of the scene were taken by a handhold camera. Two of them are shown in Fig.3.(1, 4). Thirty-two feature points represent the scene structure and are tracked across the image sequence.

According to the deformable structures recovered by our method, we transform the scene appearance to side and top views. The side views of the scene in Fig.3.(1, 4) are shown in Fig.3.(2, 5) and the top views in Fig.3.(3, 6). The red wireframes show the static objects, the yellow ones refer to the moving objects, and the blue lines mean the moving trajectories from the beginning of the sequence till the present frames. The reconstruction is consistent with the observation, *e.g.,* the three objects move straight on the table and slope respectively and the top of the biggest box is close to a rectangle. As shown in [23], the weak-perspective reconstruction achieves similar results, just slightly worse than the perspective reconstruction in Fig.3, because the distance of the scene from the camera is large relative to its inner depth variance and so the perspective effect is weak.

In practice, the deformation of a dynamic scene is often degenerate, *i.e.,* involving shape bases of rank 1 or 2 [24]. For example, when several people walk straight independently along different directions on the 2D ground, each of the linear translations refers to a respective rank-1 basis. Suppose K_1 out of K bases are such rank-1 bases, the scaled measurement matrix W_s satisfies a rank of $3K - 2K_1 + 1$. Our projective reconstruction algorithm using only the rank constraint is thus capable of recovering the projective depths. Then, based on the analysis in Section 3 and [24], we extend our perspective reconstruction algorithm for such degenerate deformations. One example is shown in Fig.4. The scene consists of a static table and two boxes being translated independently along respective table borders. The scene structure is composed of 3 bases: one rank-3 basis for the static table and initial positions of the two boxes and two rank-1 bases respectively for the linear translations. The rank of the scaled measurements is thus 6. Thirty images were taken by an automatic focusing camera and 18 feature points were tracked across the sequence.

Fig.4.(1, 4) show two of the images. Their corresponding scene appearances reconstructed by our method are viewed from the top in Fig.4.(2, 5). The yellow wireframes mean the static table, and the red and pink ones refer to the two moving boxes respectively. The reconstruction is consistent with the truth, *e.g.,* the boxes are translated along respective table borders and the shapes of the table top and box tops are close to rectangles. Fig.4.(3, 6) demonstrate the weak-perspective reconstruction of the scene by the method in [24]. Because the images were taken from a relatively small distance, the perspective effect is strong and the recovered structures are apparently distorted, *e.g.,* the shape of table top is far from rectangle. Fig.5 shows that our method is able to recover the varying focal lengths of the automatic focusing camera. The variance is not big because the close distance between the scene and the camera did not change much across the sequence.

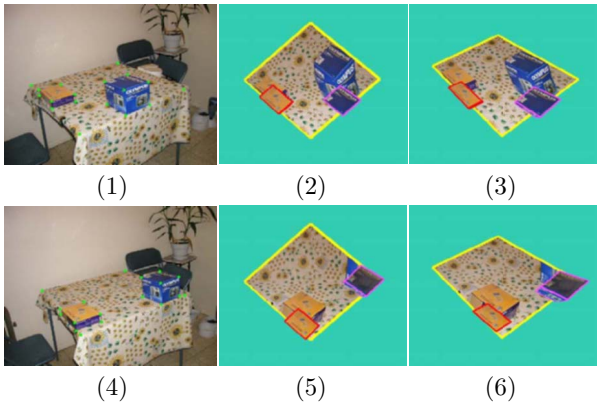

Figure 4: Perspective reconstruction of degenerate scene deformations. (1)&(4): Two input images. (2)&(5): Top view of the scene appearance recovered by our method. The yellow wireframes mean the static table. The red and pink ones refer to the moving boxes. (3)&(6): Weak-perspective reconstruction by the method in [24], where the perspective distortion is notable.

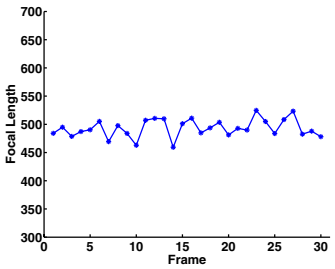

Figure 5: The recovered focal lengths of the moving box sequence vary across the sequence as expected.

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

5. Conclusions and Discussions

In this paper we present a 2-step factorization algorithm for perspective reconstruction of deformable structures from uncalibrated images. The first step recovers the projective depths using the sub-space constraints embedded in the image measurements of the deformable structures. The second step factorizes the image measurements scaled by the recovered projective depths to reconstruct the deformable 3D structures, rigid rotations and translations, and varying focal lengths simultaneously. Since this method allows varying structures and unsynchronized and automatic focusing cameras, it provides a powerful tool for applications such as dynamic camera networks over large-scale dynamic scenes.

The current algorithm updates the projective depths of one point or in one image individually. It is inefficient when the number of images or points is big. We are working on how to segment the points or images into groups and update the projective depths group by group. We believe such segmentation also helps detect and eliminate outliers. Another problem is how to properly use the constraints on orthonormality of rotations and those on uniqueness of bases in the perspective reconstruction step. Presently we equivalently enforces these two constraints. However one of them comes from all images and another is from only the first K images. When noise exists, they might have different stabilities and thus different importance to the solution. We are exploring a way of weighting the constraints to improve the performance. Another benefit of using the weights is that we can accordingly sample the constraints to improve the efficiency of the algorithm.

References

[1] V. Blanz, T. Vetter, " A morphable model for the synthesis of 3D faces," *SIGGRAPH*, 1999.

[2] S. Bougnoux, " From Projective to Euclidean Space under Any Practical Situation, A Criticism of Self-Calibration," *ICCV*, 1998.

[3] M. Brand, " Morphable 3D models from video," *CVPR*, 2001.

[4] C. Bregler, A. Hertzmann, H. Biermann, " Recovering non-rigid 3D shape from image streams," *CVPR*, 2000.

[5] J. Costeira, T. Kanade, " A Multibody Factorization Method for Independently Moving-Objects," *IJCV*, 29(3):159-179, 1998.

[6] O.D. Faugeras, " What can Be Seen in Three Dimensions with An Uncalibrated Stereo Rig," *ECCV*, 1992.

[7] M. Han, T. Kanade, " Creating 3D Models with Uncalibrated Cameras," *WACV*, 2000.

[8] R. Hartley, R. Gupta, T. Chang, " Stereo from Uncalibrated Cameras," *CVPR*, 1992.

[9] S. Mahamud, M. Hebert, " Iterative Projective Reconstruction from Multiple Views," *CVPR*, 2000.

[10] D.D. Morris, K. Kanatani, T. Kanade, " Uncertainty Modeling for Optimal Structure from Motion," *Vision Algorithms: Theory and Practice*, 1999.

[11] J. Oliensis, " A Multi-frame Structure from Motion Algorithm under Perspective Projection," *IJCV*, 34(2/3):163-192, 1999.

[12] M. Pollefeys, R. Koch, L. Van Gool, " Self-Calibration and Metric Reconstruction in Spite of Varying and Unknown Internal Camera Parameters," *ICCV*, 1998.

[13] P. H. Schönemann, R. M. Carroll, " Fitting One Matrix to Another under Choice of A Central Dilation and A Rigid Motion," *Psychometrika*, 35(2):245-255, 1970.

[14] P.F. Sturm, B. Triggs, " A Factorization-Based Algorithm for Multi-Image Projective Structure and Motion," *ECCV*, 1996.

[15] R. Szeliski, S. B. Kang, " Recovering 3D shape and motion from image streams using non-linear least squares," *VCIR*, 5(1):10-28, 1994.

[16] C. Tomasi, T. Kanade, " Shape and motion from image streams under orthography: A factorization method," *IJCV*, 9(2):137-154, 1992.

[17] L. Torresani, D. Yang, G. Alexander, C. Bregler, " Tracking and modeling non-rigid objects with rank constraints," *CVPR*, 2001.

[18] B. Triggs, " Factorization methods for projective structure and motion," *CVPR*,1996.

[19] 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.

[20] D. Weinshall, P. Anandan, M. Irani, " From Ordinal to Euclidean Reconstruction with Partial Scene Calibration," *ECCV Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, 1998.

[21] L. Wolf, A. Shashua, " Two-body segmentation from two perspective views," *CVPR*, 2001.

[22] L. Wolf, A. Shashua, " On projection matrices $P^k \rightarrow P^2, k = 3, \ldots, 6$, and their applications in computer vision," *IJCV*, 48(1):53-67, 2002.

[23] J. Xiao, J. Chai, T. Kanade, " A Closed-Form Solution to Non-Rigid Shape and Motion Recovery," *ECCV*, 2004.

[24] J. Xiao, T. Kanade, " Non-Rigid Shape and Motion Recovery: Degenerate Deformations," *CVPR*, 2004.

[25] Z. Zhang, R. Deriche, O. Faugeras, Q. Luong, " A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *AI*, 78(1/2):87-119, 1995.

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**