# Shape Ambiguities in Structure From Motion

Richard Szeliski and Sing Bing Kang

**Abstract**—This paper examines the fundamental ambiguities and uncertainties inherent in recovering structure from motion. By examining the eigenvectors associated with null or small eigenvalues of the Hessian matrix, we can quantify the exact nature of these ambiguities and predict how they affect the accuracy of the reconstructed shape. Our results for orthographic cameras show that the bas-relief ambiguity is significant even with many images, unless a large amount of rotation is present. Similar results for perspective cameras suggest that three or more frames and a large amount of rotation are required for metrically accurate reconstruction.

**Index Terms**—Structure from motion, ambiguities, bas-relief ambiguity, uncertainty analysis, eigenvalue analysis.

## 1 Introduction

STRUCTURE from motion is one of the classic problems in computer vision, and has received a great deal of attention over the last decade. It has wide-ranging applications including robot vehicle guidance and obstacle avoidance, and the reconstruction of 3D models from imagery. Unfortunately, while the qualitative estimates of structure and motion look reasonable, the actual quantitative (*metric*) estimates can be significantly distorted.

Much progress has been made recently in identifying the sources of errors and instabilities in the structure from motion process. It is now widely understood that the arbitrary algebraic manipulation of the imaging equations to derive closed-form solutions (e.g., [4]) can be unstable. To overcome this, statistically optimal algorithms for estimating structure and motion have been developed [11], [17], [16], [12]. It is also understood that using more feature points and images results in better estimates, and that certain configurations of points (at least in the two frame case) are pathological and cannot be reconstructed.

An example of an algorithm which generates very good results is the factorization approach of Tomasi and Kanade [16]. This algorithm, which assumes orthography, uses many points and frames, and for most sequences, a large amount of object rotation. However, when only a small range of viewpoints is present, the reconstruction appears distorted.

A number of authors has also analyzed the role of degenerate point configurations called *critical surfaces* in structure from motion ambiguities [6]. Here, we concentrate more on the ambiguities (uncertainties) which remain even when the points are well distributed in space.

In this paper, we demonstrate that it is the overall variation in viewpoint or object rotation which critically determines the quality of the reconstruction. The ambiguity in object shape due to small viewpoint variation often resemble a *projective* deformation of the Euclidean shape. In fact, we show that the major ambiguity in the reconstruction is a simple depth scale uncertainty.

- *R. Szeliski is with Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399. E-mail: szeliski@microsoft.com.*
- *S.B. Kang is with Digital Equipment Corporation, Cambridge Research Lab, One Kendall Square, Bldg. 700, Cambridge, MA 02139. E-mail: sbk@crl.dec.com.*

*Manuscript received Feb. 23, 1996; revised Jan. 2, 1997. Recommended for acceptance by S. Peleg.*
*For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P97012.*

To derive our results, we use eigenvalue analysis of the covariance matrix for the structure and motion estimates. Our results are significant for two reasons. First, we show how to theoretically derive the expected ambiguity in a reconstruction, and also derive some intuitive guidelines for selecting imaging situations which can be expected to produce reasonable results. Second, since the primary ambiguities are very well characterized by a small number of modes, this information can be used to construct better online (recursive) estimation algorithms.

## 2 Previous Work

Structure from motion has been extensively studied in computer vision. Early papers on this subject [4] develop algorithms to compute the structure and motion from a small set of points matched in two frames using an *essential parameter* approach. The performance of this approach can be significantly improved using nonlinear least squares (*optimal estimation*) techniques [17], [11]. Recent research focuses on extraction of shape and motion from longer image sequences [16].

The nature of structure and motion errors, which is the main focus of this paper, has also previously been studied. Weng et al. perform some of the earliest and most detailed error analyses of the two-frame essential parameter approach [17]. Adiv [1], Daniilidis and Nagel [3], and Young and Chellappa [19] analyze continuous-time (optical flow) based algorithms using the concept of the Cramer-Rao lower bound. Oliensis and Thomas [7] show how modeling the motion error can significantly improve the performance of recursive algorithms.

In this paper, we extend these previous results using an eigenvalue analysis of the covariance matrix. This analysis can pinpoint the exact nature of structure from motion ambiguities and the largest sources of reconstruction error. We also focus on multiframe optimal structure from motion algorithms, which have not been studied in great detail.

## 3 Problem Formulation and Uncertainty Analysis

Structure from motion can be formulated as the recovery of a set of 3D structure parameters $\mathbf{p}_i$ and time-varying motion parameters $\mathbf{m}_j$ from a set of observed image features $\mathbf{u}_{ij}$. In this section, we present the forward equations which map 3D points into 2D image points. We also describe our uncertainty analysis using the Jacobians of the forward equations and how our results relate to classical structure from motion ambiguities.

### 3.1 Problem Formulation

The equation which projects the $i$th 3D point $\mathbf{p}_i$ into the $j$th frame at location $\mathbf{u}_{ij}$ is

$$\mathbf{u}_{ij} = \mathcal{P}(T(\mathbf{p}_i, \mathbf{m}_j)) = \mathcal{P}(\mathbf{R}_j\mathbf{p}_i + \mathbf{t}_j) \tag{1}$$

where $\mathbf{m}_j$ are the motion parameters for camera position $j$, $\mathcal{P}$ the perspective projection (defined below), $\mathbf{R}_j$ a rotation matrix, and $\mathbf{t}_j$ a translation. In this paper, $\mathbf{R}_j$ is represented using the quaternion $\mathbf{q} = [w, (q_0, q_1, q_2)]$, since this representation has no singularities. For our 1D examples, we use the rotation angle around the vertical axis.

The standard perspective projection equation used in computer vision is

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathcal{P}_1\begin{pmatrix} x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} f\frac{x}{z} \\ f\frac{y}{z} \end{pmatrix} \tag{2}$$

where $f$ is a product of the focal length of the camera and the pixel scale factor (assuming that pixels are square). An alternative object-centered formulation, which we introduced in [12] is

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathcal{P}_2\begin{pmatrix} x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} s\frac{x}{1+\eta z} \\ s\frac{y}{1+\eta z} \end{pmatrix} \tag{3}$$

Here, we assume that the $(x, y, z)$ coordinates before projection are with respect to a reference frame $\Pi_j$ that has been displaced away from the camera by a distance $t_z$ along the optical axis, with *scale factor* $s = f/t_z$ and *perspective distortion factor* $\eta = 1/t_z$ (Fig. 1). This formulation allows us to model both orthographic and perspective cameras.

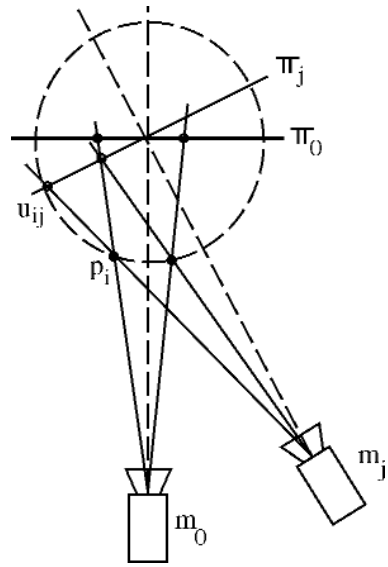



Fig. 1. Sample configuration cameras ($\mathbf{m}_j$) 3D points ($\mathbf{p}_i$), image planes ($\Pi_j$), and screen locations ($\mathbf{u}_{ij}$).

In our previous work [12], we used the iterative Levenberg-Marquardt algorithm to estimate $\{\mathbf{p}_i, \mathbf{m}_j\}$ from $\{\mathbf{u}_{ij}\}$, since it provides a statistically optimal solution [17], [11], [14], [12]. The Levenberg-Marquardt method is a standard non-linear least squares technique [8] which directly minimizes an objective function

$$C(\mathbf{a}) = \sum_i \sum_j c_{ij} \left| \tilde{\mathbf{u}}_{ij} - \mathbf{f}_{ij}(\mathbf{a}) \right|^2 \tag{4}$$

where $\tilde{\mathbf{u}}_{ij}$ is the observed image measurement, $\mathbf{f}_{ij}(\mathbf{a}) = \mathbf{u}(\mathbf{p}_i, \mathbf{m}_j)$ is given in (1), and the vector $\mathbf{a}$ contains the 3D points $\mathbf{p}_i$, the motion parameters $\mathbf{m}_j$, and any additional unknown calibration parameters. The weight $c_{ij}$ in (4) describes the confidence in measurement $\mathbf{u}_{ij}$, and is normally set to the inverse variance $\sigma_{ij}^{-2}$ (or zero for missing measurements). This same approach has long been used in the photogrammetric community under the name of *bundle adjustment* [9].

### 3.2 Uncertainty Analysis

Regardless of the solution technique, the uncertainty in the recovered parameters—assuming that image measurements are corrupted by small Gaussian noise errors—can be determined by computing the inverse covariance or *information* matrix $\mathbf{A}$ [10]. Assuming that the measurement errors are independent identically distributed Gaussian errors with variance $\sigma_{ij} = 1$, the information matrix can be formed by computing outer products of the *Jacobians* of the measurement equations

$$\mathbf{A} = \sum_i \sum_j c_{ij} \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{a}} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{a}^T} \tag{5}$$

For notational succinctness, we use the symbol

$$\mathbf{H}_{ij} = \begin{bmatrix} \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{p}_i} \\ \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{m}_j} \end{bmatrix}$$

to denote the non-zero portion of the full Jacobian $\frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{a}}$.

If we list the structure parameters $\{\mathbf{p}_i\}$ first, followed by the motion parameters $\{\mathbf{m}_j\}$, the $\mathbf{A}$ matrix has the structure

$$\mathbf{A} = \left[ \begin{array}{c|c} \mathbf{A_p} & \mathbf{A_{pm}} \\ \hline \mathbf{A_{pm}}^T & \mathbf{A_m} \end{array} \right] \tag{6}$$

The matrices $\mathbf{A_p}$ and $\mathbf{A_m}$ are block diagonal, with diagonal entries

$$\mathbf{A}_{\mathbf{p}_i} = \sum_j \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{p}_i} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{p}_i^T} \text{ and } \mathbf{A}_{\mathbf{m}_j} = \sum_i \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{m}_j} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{m}_j^T} \tag{7}$$

respectively, while $\mathbf{A_{pm}}$ is dense, with entries

$$\mathbf{A}_{\mathbf{p}_i \mathbf{m}_j} = \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{p}_i} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{m}_j^T} \tag{8}$$

The $\mathbf{A}$ matrix is the same Hessian matrix as is used in the inner loop of the Levenberg-Marquardt iterative non-linear least squares algorithm [8], [12].

The information matrix has previously been used in the context of structure from motion to determine *Cramer-Rao lower bounds* on the parameter uncertainties by taking the inverse of the diagonal entries [1], [19]. The Cramer-Rao bounds, however, can be arbitrarily weak, especially when $\mathbf{A}$ is singular or near-singular. In this paper, we use eigenvector analysis of $\mathbf{A}$ to find the dominant directions in the uncertainty (covariance) matrix and their magnitudes, which gives us more insight into the exact nature of structure from motion ambiguities (see Section 3.4).

### 3.3 Estimating Reconstruction Errors

An important benefit of uncertainty analysis is that we can easily quantify the expected amount of reconstruction (and motion) error for an optimal structure from motion algorithm. For example, the expected sum of squared error in reconstructed 3D point positions is

$$S_{pos}^2 \equiv \left\langle \sum_i \left\| \tilde{\mathbf{p}}_i - \mathbf{p}_i^* \right\|^2 \right\rangle \tag{9}$$

where $\tilde{\mathbf{p}}_i$ are the estimated (recovered) positions and $\mathbf{p}_i^*$ the true positions (the angle brackets $\langle\ \rangle$ indicate the expected value). The positional uncertainty matrix $\mathbf{C_p}$ can be computed by inverting $\mathbf{A}$ and looking at its upper left block (the block corresponding to the $\mathbf{p}_i$ variables).[1] If we perform an eigenvalue analysis of $\mathbf{C_p}$, we obtain

$$\mathbf{C_p} = \mathbf{E_p}^T \mathbf{\Lambda_p} \mathbf{E_p} \tag{10}$$

where $\mathbf{E_p}$ is the matrix of eigenvectors, and $\mathbf{\Lambda_p}$ is the diagonal matrix containing the eigenvalues of $\mathbf{C_p}$. Since $S_{pos}^2$ is a Euclidean norm, its value is unaffected by orthogonal coordinate transformations such as $\mathbf{E_p}$. The value of $S_{pos}^2$ can thus be computed as either the trace of $\mathbf{C_p}$ or the trace of $\mathbf{\Lambda_p}$, i.e., the sum of the eigenvalues of $\mathbf{C_p}$.

In practice, we do not need to compute $\mathbf{C_p}$. Instead, the sum of squared reconstruction and motion error,

$$S_{all}^2 \equiv \left\langle \sum_i \left\| \tilde{\mathbf{p}}_i - \mathbf{p}_i^* \right\|^2 + \sum_j \left\| \tilde{\mathbf{m}}_j - \mathbf{m}_j^* \right\|^2 \right\rangle \tag{11}$$

can be computed directly summing the *inverse* eigenvalues of the information matrix $\mathbf{A}$. By choosing an appropriate scaling for the parameters being estimated, we can make $S_{all}$ be close to $S_{pos}$.[2]

What is the advantage of this approach, if computing eigenvalues is just as expensive as inverting matrices? First, we can compute the first few eigenvalues more cheaply (and in less space) than the matrix inverse using an inverse iteration algorithm [2], [8], and these tend to dominate the overall reconstruction error. Second, it justifies the approach in the paper, which is to look at the minimum eigenvalue as the prime indicator of reconstruction error. We can therefore study how much certain ambiguities (such as the bas-relief ambiguity) contribute to the overall reconstruction error. We can also obtain much tighter lower bounds on the reconstruction error than would be possible by using the Cramer-Rao bounds.

### 3.4 Ambiguities in Structure From Motion

Because structure from motion attempts to recover both the structure of the world and the camera motion without any external or prior knowledge, it is subject to certain ambiguities. The most fundamental of these is the coordinate frame (also known as pose, or Euclidean) ambiguity.

The next most common ambiguity is the scale ambiguity (for a perspective camera) or the depth ambiguity (for an orthographic camera). This ambiguity can be removed with a small amount of additional knowledge, e.g., the absolute distance between camera positions. A third ambiguity, and the one we focus on in this paper, is the *bas-relief ambiguity*. In its pure form, this ambiguity occurs for a two frame problem with an orthographic camera, and is a confusion between the *relative depth* of the object and the amount of object rotation (i.e., $\{(x_i, y_i, z_i)\}$ and $\{(x_i, y_i, az_i)\}$ are valid solutions).

In this paper, we focus on the *weak* form of this ambiguity, i.e., the very large *bas-relief uncertainty* which occurs with imperfect measurements even when we use more than two frames and/or perspective cameras. A central result of this paper is that the bas-relief ambiguity captures the largest uncertainties arising in structure from motion. However, when examined in detail, it appears that a larger class of deformations (i.e., projective) more fully characterizes the errors which occur in structure from motion.

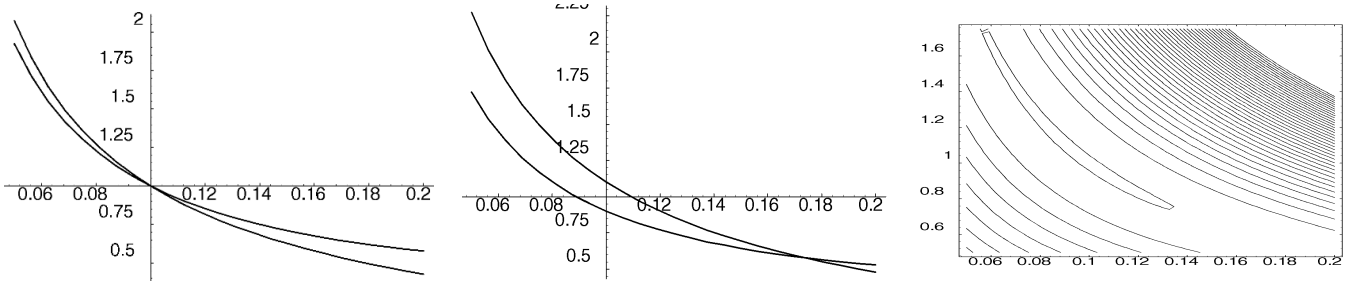
To characterize these ambiguities, we will use eigenvector analysis of the information matrix, as explained in Section 3.2. Absolute ambiguities will show up as zero eigenvalues, whereas weak ambiguities will show up as small eigenvalues.

Consider, for example, an eigenvector of the $\mathbf{A}$ matrix which has ones in locations corresponding to the $z$ coordinates of the 3D point positions $\mathbf{p}_i$ and zeros in the $x$ and $y$ coordinates (ignoring motion parameter entries). If the eigenvalue associated with this eigenvector is zero, we can add any constant vector to the $z$ coordinates in our structure estimate (and some corresponding vector to the motion estimates), and still obtain an equally valid solution. We thus can interpret an eigenvector of $\mathbf{A}$ with zero eigenvalue as a structure/motion ambiguity.

Similarly, if a null eigenvector of $\mathbf{A}$ has entries in the $z$ structure coordinates proportional to the values of the $z$ estimates themselves (and zero values in the $x$ and $y$ coordinates), we can add any multiple of the $z$ values to the solution and still have a valid solution. We thus identify this case with a *bas-relief ambiguity*. If the eigenvalue associated with such an eigenvector is small but not zero, we call this situation a *bas-relief uncertainty*. We will use this kind of analysis to establish all of the main results in this paper.

1. Note that this is *not* the same as simply inverting $\mathbf{A_p}$.

2. Adding translation and rotation errors is not well founded, which is one of the reasons why this paper concentrates on uncertainties and errors in the *structure* instead of the *motion*.




(a) Noise-Free (b) Noise = 0.01 (c) Contour Plot

Fig. 2. Constraint curves and objective function for simple two-parameter example. The –axis is the angle $\Delta\theta$ and the *y*-axis is the scale factor *a*.

## 4 A Two Parameter Example

To develop an intuitive understanding of the basic bas-relief ambiguity, we start with a simple two-parameter example. Assume that we have an orthographic scanline camera which measures the $x$ component of 2D points $(x, z)$. Furthermore, assume that we already know the shape up to a scale factor in depth, $\mathbf{p}_i = (x_i, az_i)$ and that the rotation angles are uniform, $\theta_j = j\Delta\theta$. The projection equation is then

$$u_{ij} = c_j x_i - s_j a z_i \tag{12}$$

with $c_j = \cos\theta_j$ and $s_j = \sin\theta_j$.

What happens when we try to estimate the scale factor $a$ and the angle $\Delta\theta$ from a set of noisy measurements $\{u_{ij}\}$? First, let us examine the very simplest case, which is a single point, say at $(x, z) = (1, 1)$. Each new image gives us a constraint of the form

$$c_j - as_j = c_j^* - a^* s_j^* + n_j \tag{13}$$

where $c_j^*$, $s_j^*$, and $a^*$ are the true values and $n_j$ is random noise. Fig. 2a shows the two constraint curves for $j = \pm 1$ assuming the noise-free case, i.e., the locus of the solutions to (13) with $n_j = 0$, with $a = 1$ and $\Delta\theta = 0.1$ rad. Fig. 2b shows the constraint lines for $n_{-1} = n_1 = 0.01$. As can be seen, the estimate for $(\Delta\theta, a)$ is very sensitive to noise. This can also be seen in the contour plot of the objective function $C(\Delta\theta, a)$ (see (4)) shown in Fig. 2c, which can be computed by summing the constraints in (13).

To characterize the shape of the error surface near its minimum, we compute the information matrix $\mathbf{A}$. The Jacobian for $(a, \Delta\theta)$ is straightforward,

$$\mathbf{H}_{ij} = \begin{bmatrix} \dfrac{\partial u_{ij}}{\partial a} \\ \dfrac{\partial u_{ij}}{\partial \Delta\theta} \end{bmatrix} = \begin{bmatrix} -s_j z_i \\ -j\left(ac_j z_i + s_j x_i\right) \end{bmatrix} \approx -j\begin{bmatrix} \Delta\theta z_i \\ az_i + j\Delta\theta x_i \end{bmatrix} \tag{14}$$

if we assume small rotation angles, $|\theta_j| \ll 1$, so that $s_j \approx j\Delta\theta$ and $c_j \approx 1$. The inverse covariance (information) matrix is then

$$\mathbf{A} \approx J_2 Z \begin{bmatrix} \Delta\theta^2 & a\Delta\theta \\ a\Delta\theta & a^2 + \Delta\theta^2 \dfrac{J_4 X}{J_2 Z} \end{bmatrix} \tag{15}$$

where $J_2 = \sum_j j^2$, $J_4 = \sum_j j^4$, $X = \sum_i x_i^2$, and $Z = \sum_i z_i^2$ (assuming that $\sum_j j = 0$). Assuming that $\Delta\theta^2 \ll a^2$, we can compute [13] the approximate eigenvalues of $\mathbf{A}$ as

$$\lambda_{\min} \approx \Delta\theta^4 J_4 X / a^2 \text{ and } \lambda_{\max} \approx J_2 Z a^2 \tag{16}$$

## 5 Orthography: Single Scanline

Let us now turn to a true structure from motion problem where both the structure and motion are unknown. For simplicity, we analyze the orthographic scanline camera first, where the unknowns are the 2D point positions $\mathbf{p}_i = (x_i, z_i)$ and the rotation angles $\theta_j$.[3] The imaging equations are

$$u_{ij} = c_j x_i - s_j z_i \tag{17}$$

with $c_j = \cos\theta_j$ and $s_j = \sin\theta_j$.

The Jacobian for the 1D orthographic camera is

$$\mathbf{H}_{ij} = \left[\frac{\partial u_{ij}}{\partial x_i} \quad \frac{\partial u_{ij}}{\partial z_i} \quad \frac{\partial u_{ij}}{\partial \theta_j}\right]^T = \left[c_j \quad -s_j \quad -\left(c_j z_i + s_j x_i\right)\right]^T \tag{18}$$

and the entries in the information matrix are

$$\mathbf{A}_{\mathbf{p}_i} = \begin{bmatrix} \sum_j c_j^2 & -\sum_j c_j s_j \\ -\sum_j c_j s_j & \sum_j c_j^2 \end{bmatrix} = \begin{bmatrix} C & -D \\ -D & S \end{bmatrix} \tag{19}$$

$$\mathbf{A}_{\mathbf{p}_i \mathbf{m}_j} = \begin{bmatrix} -c_j^2 z_i - c_j s_j x_i \\ c_j s_j z_i + s_j^2 x_i \end{bmatrix} \tag{20}$$

$$\mathbf{A}_{\mathbf{m}_j} = \left[\sum_i \left(c_j z_i + s_j x_i\right)^2\right] = \left[c_j^2 Z + 2c_j s_j W + s_j^2 X\right] \tag{21}$$

with $C = \sum_j c_j^2$, $D = \sum_j c_j s_j$, $S = \sum_j s_j^2$, $Z = \sum_i z_i^2$, $W = \sum_j z_i x_i$, and $X = \sum_i z_i^2$.

If we know the motion, the structure uncertainty is determined by $\mathbf{A}_{\mathbf{p}_i}$ and is simply the triangulation error, i.e., $\sigma_x^2 \propto C^{-1}$ and $\sigma_z^2 \propto S^{-1}$ (for small rotations, $\sigma_x^2 \ll \sigma_z^2$). The result that depth uncertainty is primarily in $z$ is well known [5]. If we know the structure, the motion accuracy is determined by $\mathbf{A}_{\mathbf{m}_j}$ and is inversely proportional to the variance in depth along the viewing direction $(s_j, c_j)$.

What about ambiguities in the solution? Under orthography, scale ambiguity does not exist. However, translations along the optical axis cannot be estimated, and an overall pose ambiguity

3. We do not estimate the horizontal translation since it can be estimated from the motion of the centroid of the image points [16]. While the error in this estimate does contribute to the overall reconstruction error, we have neglected it in our analysis since its magnitude is much smaller than the error caused by the bas-relief ambiguity.

still exists. Unless we add additional constraints, we can always rotate the coordinate system by $\Delta\theta$ and add the same amount to $\{\theta_j\}$. This manifests itself as the null (zero eigenvalue) eigenvector

$$\mathbf{e}_0 = [\, z_0 \quad -x_0 \quad \cdots \quad z_N \quad -x_N \,|\, 1 \quad \cdots \quad 1]^T$$

### 5.1 Two Frames: The Bas-Relief Ambiguity

Let us say we only have two frames, and we have fixed $\theta_0 = -\theta_1$, $c_0 = c_1 = c$, $-s_0 = s_1 = s$ (Fig. 3). Then

$$\mathbf{A}_{\mathbf{p}_i} = \begin{bmatrix} 2c^2 & 0 \\ 0 & 2s^2 \end{bmatrix}, \mathbf{A}_{\mathbf{p}_i\mathbf{m}} = \begin{bmatrix} -2csx_i \\ 2csz_i \end{bmatrix}, \mathbf{A}_{\mathbf{m}} = \begin{bmatrix} 2c^2 Z + 2s^2 X \end{bmatrix} \qquad (22)$$

The bas-relief ambiguity manifests itself as a null eigenvector

$$\mathbf{e}_0 = [s^2 x_0 \quad -c^2 z_0 \quad \cdots \quad s^2 x_N \quad -c^2 z_N \;\mid\; cs]^T \qquad (23)$$

as can be verified by inspection. This is as we expected, i.e., the primary uncertainty in the structure is entirely in the depth ($z$) direction, and is a scale uncertainty (proportional to $z$). There is also a much smaller "bulging" in the $x$ values (at least for small inter-frame rotations). This squashing and bulging is an affine deformation of the true structure.

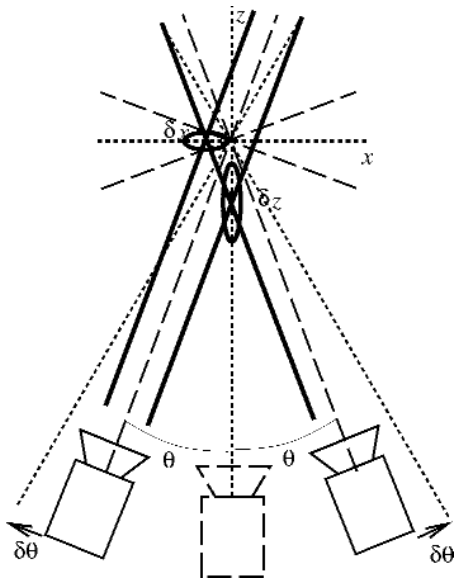



Fig. 3. Orthographic projection, two frames. The solid lines indicate the viewing rays, while the thin lines indicate the optical axes and image planes. The diagonal dashed lines are the displaced viewing rays, while the ellipses indicate the positional uncertainty in the reconstruction due to uncertainty in motion (indicated as $\delta\theta$).

### 5.2 More Than Two Frames, Equi-Angular Motion Constraint

To simplify the analysis, we assume for the moment that we know we have an equi-angular image sequence, i.e., that the rotation angles are given by $\theta_j = j\Delta\theta$, $j \in \{-J, \ldots, J\}$, $J = \frac{F+1}{2}$, where $F$ is the total number of frames (imagine Fig. 3b with more cameras). Skipping the derivation, which can be found in [13], we obtain

$$\lambda_{\min} \approx \frac{\Delta\theta^4 X J_2 \left(J_0 J_4 - J_2^2\right)}{J_0 J_2 Z + \Delta\theta^2 \left[X\left(J_0 J_4 - J_2^2\right) + J_0 J_2\right]} \qquad (24)$$

where $J_0 = \sum_j 1$, $J_2 = \sum_j j^2$, $J_4 = \sum_j j^4$, $X = \sum_i x_i^2$, and $Z = \sum_i z_i^2$ (see also (15)). Notice that doubling the inter-frame rotation reduces the RMS (root mean square) error by a factor of four (assuming that $Z \gg \Delta\theta^2$). Increasing the extent of the $x_i$ compared to the $z_i$ directly increases the minimum eigenvalue, i.e., it decreases the structure uncertainty. This result suggests that flatter objects can be reconstructed better.

Once the smallest eigenvalue and eigenvector have been computed, we can easily determine some additional eigenvectors. Any vector which consists purely of $x_i$ or $z_i$ values which is also orthogonal to $\mathbf{A}_{\mathbf{pm}}$ is an eigenvector, e.g.,

$$\mathbf{e} = [x_1 \quad 0 \quad -x_0 \quad 0 \quad \cdots \quad 0 \,|\, 0]$$

The eigenvalues corresponding to the pure $x$ eigenvectors are $C$, while the $z$ eigenvalues are $S$ (see the definition following (21)). In other words, once the global bas-relief uncertainty has been accounted for (squashing in $z$ and smaller bulging in $x$), the variance in $x$ position estimates is proportional to $C^{-1}$ and in $z$ positions is proportional to $S^{-1}$, i.e., the expected triangulation error for known camera positions.

For the above example with $J = 1$ (three frames), $\Delta\theta = 0.1$ rad $\approx 6^\circ$, and $X = Z = 1$, the values for $C$, $S$, and $\lambda_{min}$ are 2.98, 0.0199, and 0.00006644, respectively. Thus, the correlated depth uncertainty due to the motion uncertainty is a factor of $0.0199 / 0.00006644 = 300$ times greater than the individual (triangulation) depth uncertainties. The table of $\lambda_{\min}$ as a function of $F = 2J + 1$ (number of frames) and $\theta_{tot} = (F-1)\Delta\theta$ (total rotation angle) is shown in Table 1.

### 5.3 More Than Two Frames, Without Motion Constraint

If we take the same data set as above, but remove the additional knowledge of equi-angular steps, we end up solving for each motion (angle) estimate separately. In this case, we do not have a closed form solution, since we have $2J + 3$ equations in three unknowns. However, if we assume a small angle approximation and $W = 0$ (i.e., that the 3D point cloud is rotationally symmetric with respect to the middle frame), then we get the same eigenvectors as with the known equiangular motion constraint.

This behavior can be verified numerically [13], with results that are quite similar to those shown in Table 1. To obtain these results, we computed the **A** matrix explicitly using a set of nine points sampled on the unit square, i.e., $\{(x, z), x, z \in \{-1, 0, 1\}\}$, and then numerically computed the eigenvalues.

## 6 Perspective in 3D

The full length version of our paper [13] performs the analysis of the orthographic camera in 3D, and the perspective scanline camera. The results on the former are similar to the orthographic scanline results presented in the previous section. For the perspective scanline camera, we make the observation that the two-frame problem is still ambiguous, since the optical rays will still intersect for any camera configuration. In this section, we analyze the most interesting case, that of a perspective camera operating in a 3D environment. Here, we know that the two-frame problem has a solution, although our results on the simpler camera models suggest that the reconstructions may be particularly sensitive to noise.

The forward imaging equations are given in (1) and (3). In this paper, we present some numerical results on $\lambda_{\min}$ and discuss their significance. These results were obtained using the *Mathematica*® package [18]. For this example, we used a 15-point data set consisting of the eight corners of a unit cube, the six cube faces, and the origin.

We present the results here for the special case of pure object-centered rotation (which in camera-centered coordinates is actually both rotation and translation). The results for pure forward translation are given in [13].

Table 2 shows the eigenvalues computed for various rotation

TABLE 1
MINIMUM EIGENVALUES FOR 1D ORGHOGRAPHIC KNOWN EQUI-ANGULAR MOTION

| $\lambda_{\min}$ | $F = 2$ | $F = 3$ | $F = 4$ | $F = 5$ | $F = 6$ | $F = 7$ | $F = 8$ |
|---|---|---|---|---|---|---|---|
| $\theta_{\text{tot}} = 11.5°$ | 0.000000 | 0.000067 | 0.000079 | 0.000088 | 0.000096 | 0.000104 | 0.000112 |
| $\theta_{\text{tot}} = 22.9°$ | 0.000000 | 0.001087 | 0.001283 | 0.001418 | 0.001547 | 0.001677 | 0.001810 |
| $\theta_{\text{tot}} = 34.4°$ | 0.000000 | 0.005618 | 0.006597 | 0.007277 | 0.007931 | 0.008594 | 0.009269 |
| $\theta_{\text{tot}} = 45°$ | 0.000000 | 0.016854 | 0.019688 | 0.021673 | 0.023596 | 0.025552 | 0.027547 |
| $\theta_{\text{tot}} = 60°$ | 0.000000 | 0.054679 | 0.063442 | 0.069678 | 0.075782 | 0.082017 | 0.088389 |
| $\theta_{\text{tot}} = 90°$ | 0.000000 | 0.272977 | 0.316453 | 0.348500 | 0.380039 | 0.412200 | 0.444997 |

TABLE 2
MINIMUM EIGENVALUES FOR 3D PERSPECTIVE PROJECTION, EQUI-ANGULAR ROTATION AROUND Y AXIS, $\eta = 0.1$

| $\lambda_{\min}$ | $F = 2$ | $F = 3$ | $F = 4$ | $F = 5$ | $F = 6$ | $F = 7$ | $F = 8$ |
|---|---|---|---|---|---|---|---|
| $\theta_{\text{tot}} = 11.5°$ | 0.000175 | 0.000214 | 0.000239 | 0.000269 | 0.000299 | 0.000331 | 0.000364 |
| $\theta_{\text{tot}} = 22.9°$ | 0.000690 | 0.001289 | 0.001462 | 0.001633 | 0.001803 | 0.001981 | 0.002158 |
| $\theta_{\text{tot}} = 34.4°$ | 0.001512 | 0.004372 | 0.004972 | 0.005491 | 0.006009 | 0.006510 | 0.007024 |
| $\theta_{\text{tot}} = 45°$ | 0.002512 | 0.009905 | 0.011282 | 0.012020 | 0.012959 | 0.013460 | 0.014070 |
| $\theta_{\text{tot}} = 60°$ | 0.004234 | 0.020246 | 0.022853 | 0.021650 | 0.021870 | 0.020495 | 0.019727 |
| $\theta_{\text{tot}} = 90°$ | 0.008381 | 0.032074 | 0.032623 | 0.027976 | 0.026149 | 0.023367 | 0.021596 |

TABLE 3
MINIMUM EIGENVALUES FOR 3D PERSPECTIVE PROJECTION, EQUI-ANGULAR ROTATION AROUND *Y* AXIS, TWO FRAMES ($F = 2$), VARYING $\eta$. $\phi$ IS THE CAMERA'S FIELD OF VIEW

| $\lambda_{\min}$ | $\eta = 0.025$ $\phi = 3^\circ$ | $\eta = 0.05$ $\phi = 6^\circ$ | $\eta = 0.1$ $\phi = 12^\circ$ | $\eta = 0.2$ $\phi = 28^\circ$ | $\eta = 0.3$ $\phi = 46^\circ$ | $\eta = 0.4$ $\phi = 67^\circ$ | $\eta = 0.5$ $\phi = 90^\circ$ |
|---|---|---|---|---|---|---|---|
| $\theta_{\text{tot}} = 11.5°$ | 0.000010 | 0.000041 | 0.000175 | 0.000899 | 0.002648 | 0.003899 | 0.002947 |
| $\theta_{\text{tot}} = 22.9°$ | 0.000040 | 0.000161 | 0.000690 | 0.003505 | 0.010216 | 0.015504 | 0.011702 |
| $\theta_{\text{tot}} = 34.4°$ | 0.000087 | 0.000354 | 0.001512 | 0.007578 | 0.021758 | 0.034461 | 0.025941 |
| $\theta_{\text{tot}} = 45°$ | 0.000145 | 0.000591 | 0.002512 | 0.012402 | 0.035035 | 0.057861 | 0.043377 |
| $\theta_{\text{tot}} = 60°$ | 0.000247 | 0.001002 | 0.004234 | 0.020494 | 0.056570 | 0.097234 | 0.072229 |
| $\theta_{\text{tot}} = 90°$ | 0.000492 | 0.001993 | 0.008381 | 0.039718 | 0.105540 | 0.144799 | 0.111384 |

angles (the rotation axis is perpendicular to the optical axis). Compared to the orthographic case (Table 1), we see some striking differences. First, the two-frame problem is now soluble (up to a scale). Second, for small viewing angles, there is marked improvement even for multiple frames. Third, the results for large viewing angles with small $\eta$s are significantly inferior to the orthographic results. This appears to be caused by ambiguities in camera motion along the optical axis ($t_z$), which are neglected in the orthographic case.

This table only shows us the results for a particular value of $\eta$. The dependence of $\lambda_{\min}$ on $\eta$ is presented in Table 3 for the two frame problems (results for three frames are given in [13]). In this table, the fields of view equivalent to each $\eta$ were computed from the horizontal spread of the data points on the unit cube and the distance of the cube from the camera $\eta^{-1}$ using the formula $\phi = 2\tan^{-1}\frac{\eta}{1-\eta}$. As can be seen for the two-frame case, doubling the amount of perspective distortion $\eta$ results in a fourfold increase in $\lambda_{\min}$ (and hence a halving of the RMS error). For the three-frame case [13], the results are less sensitive to $\eta$.

The full-length version of this paper [13] presents the above re-

sults in graphical form, shows what the typical eigenvectors look like (the main ambiguity appears to be z-scaling, but is not exactly affine), and computes the RMS structure errors by computing and inverting the full information matrix **A**, as described in Section 3.2. It also analyses the case of pure forward motion (looming). For looming, the results are much worse than those available with object-centered rotation (e.g., $\lambda_{\min}$ appears to depend quadratically on the total extent of motion [13]).

## 7 Discussion

The results presented in this paper suggest that in many situations where structure from motion might be applied, the solutions are extremely sensitive to noise. Those cases where metrically accurate results have been demonstrated almost always involve a large amount of rotation [16]. For scene reconstruction, using cameras with large fields of view, several camera mounted in different directions, or even panoramic images, should remove most of the ambiguities.

The general approach developed in this paper, i.e., eigenvalue analysis of the Hessian (information) matrix appears to explain most of the known ambiguities in structure from motion. However, there are certain ambiguities (e.g., depth reversals under orthography, or multiplicities of solutions with few points and frames) which will not be detected by this analysis because they correspond to multiple local minima of the cost function in the parameter space. Furthermore, analysis of the information matrix can only predict the sensitivity of the results to *small* amounts of image noise. Further study using empirical methods is required to determine the limitations of our approach.

Using the minimum eigenvalue to predict the overall reconstruction error may fail when the dominant ambiguities are in the motion parameters. Computing the $RMS_{pos}$ error directly from the covariance matrix $\mathbf{A}^{-1}$ is more useful in these cases.

In future work, we would like to compare results available with camera-centered and object-centered representations ((2) and (3), respectively). Our intuition is that the object-centered representation will produce estimates of better quality. Similarly, we would like to analyze the effects of mis-estimating internal calibration parameters such as focal length.

Finally, it appears that the portion of the uncertainty matrix which is correlated can be accounted for by a small number of modes. This suggest that an efficient recursive structure from motion algorithm could be developed which avoids the need for using full covariance matrices [15] but which performs significantly better than algorithms which ignore such correlations.

## 8 Conclusions

This paper has developed new techniques for analyzing the fundamental ambiguities and uncertainties inherent in structure from motion. Our approach is based on examining the eigenvalues and eigenvectors of the Hessian matrix in order to quantify the nature of these ambiguities. The eigenvalues can also be used to predict the overall accuracy of the reconstruction.

Under orthography, the bas-relief ambiguity dominates the reconstruction error, even with large numbers of frames. This ambiguity disappears, however, for large object-centered rotations. For perspective cameras, two-frame solutions are possible, but there must still be a large amount of object rotation for best performance. Using three of more frames avoids some of the sensitivities associated with two-frame reconstructions. Translations towards the object are an alternative source of shape information, but these appear to be quite weak unless large fields of views and large motions are involved.

When available, prior information about the structure or motion (e.g., absolute distances, perpendicularities) can be used to improve the accuracy of the reconstructions. Whether 3D reconstruction errors (for modeling) or motion estimation errors (for navigation) are most significant for a given application determines the conditions which produce acceptable results. In any case, careful error analysis is essential in ensuring that the results of structure from motion algorithms are sufficiently reliable to be used in practice.

## References


[1] G. Adiv, "Inherent Ambiguities in Recovering 3D Motion and Structure From a Noisy Flow Field," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, pp. 477–490, May 1989.

[2] K.-J. Bathe and E. L. Wilson, *Numerical Methods in Finite Element Analysis*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1976.

[3] K. Daniilidis and H.-H. Nagel, "Analystical Results on Error Sensitivity of Motion Estimation From Two Views," *Image and Vision Computing*, vol. 8, no. 4, pp. 297–303, Nov. 1990.

[4] H. C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene From Two Projections," *Nature*, no. 293, pp. 133–135, 1981.

[5] L. Matthies and S.A. Shafer, "Error Modeling in Stereo Navigation," *IEEE J. Robotics and Automation*, vol. 3, no. 3, pp. 239–248, June 1987.

[6] S. Maybank, *Theory of Reconstruction From Image Motion*. Berlin: Springer-Verlag, 1993.

[7] J. Oliensis and J.I. Thomas, "Incorporating Motion Error in Multi-Frame Structure From Motion, *IEEE Workshop on Visual Motion*, pp. 8-13, Princeton, N.J., Los Alamitos, Calif.: CS Press, Oct. 1991.

[8] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, England: Cambridge Univ. Press, 1992.

[9] C.C. Slama, ed. *Manual of Photogrammetry*, 4th ed. Falls Church, Va.: American Society of Photogrammetry, 1980.

[10] H.W. Sorenson. *Parameter Estimation, Principles and Problems*. New York: Marcel Dekker, 1980.

[11] M.E. Spetsakis and J.Y. Aloimonos, "Optimal Motion Estimation," *IEEE Workshop on Visual Motion*, Irvine, Calif., March 1989. Los Alamitos, Calif.: CS Press, pp. 229–237, 1989.

[12] R. Szeliski and S.B. Kang, "Recovering 3D Shape and Motion From Image Streams Using Nonlinear Least Squares," *J. Visual Communication and Image Representation*, vol. 5, no. 1, pp. 10–28, Mar. 1994.

[13] R. Szeliski and S.B. Kang, "Shape Ambiguities in Structure From Motion," Technical Report 96/1, Digital Equipment Corporation, Cambridge Research Lab, Cambridge, Mass., Jan. 1996.

[14] C.J. Taylor and D.J. Kriegman, "Structure and Motion From Line Segments in Multiple Images," *IEEE Int'l Conf. Robotics and Automation*, Nice, France, May 1992. Los Alamitos, Calif.: CS Press, pp. 1,615–1,621, 1992.

[15] J.I. Thomas, A. Hanson, and J. Oliensis, "Understanding Noise: The Critical Role of Motion Error in Scene Reconstruction," *Fourth Int'l Conf. Computer Vision (ICCV '93)*, Berlin, Germany, May 1993. Los Alamitos, Calif.: CS Press, pp. 325–329, 1993.

[16] C. Tomasi and T. Kanade, "Shape and Motion From Image Streams Under Orthography: A Factorization Method," *Int'l J. Computer Vision*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[17] J. Weng, N. Ahuja, and T.S. Huang, "Optimal Motion and Structure Estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15, no. 9, pp. 864–884, Sept. 1993.

[18] S. Wolfram, *Mathematica®: A System for Doing Mathematics by Computer*, 2nd ed. Redwood City, Calif.: Addison-Wesley, 1991.

[19] G.-S.Y. Young and R. Chellappa, "Statistical Analysis of Inherent Ambiguities in Recovering 3D Motion From a Noisy Flow Field," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 10, pp. 995–1,013, Oct. 1992.