Shape Ambiguities in Structure from Motion

Richard Szeliski[1] and Sing Bing Kang[2]

[1] Microsoft Corporation,
One Microsoft Way,
Redmond, WA 98052-6399
szeliski@microsoft.com

[2] Digital Equipment Corporation,
Cambridge Research Lab,
One Kendall Square, Bldg. 700,
Cambridge, MA 02139
sbk@crl.dec.com

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 will 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.

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 3-D models from imagery. Unfortunately, the quality of results available using this approach is still often very disappointing. More precisely, 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., [1]) can lead to algorithms that are numerically ill-conditioned or unstable in the presence of measurement errors. To overcome this, statistically optimal algorithms for estimating structure and motion have been developed [2, 3, 4]. 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 [5]. This algorithm assumes orthography and is implemented using an object-centered representation and singular value decomposition. It uses many points and frames, and for most sequences, a large amount of object rotation (usually 360°). However, when only a small range of viewpoints is present (e.g., the "House" sequence in [5], Fig. 7), the reconstruction no longer appears metric (the house walls are not perpendicular).

In this paper, we demonstrate that it is precisely this last factor, i.e., the overall rotation of the object, or equivalently, the variation in viewpoints, which critically determines the quality of the reconstruction. The ambiguity in object shape due to small viewpoint variation often looks like it might be a *projective* deformation of the Euclidean shape, which is interesting—several researchers have argued recently in favor of trying to recover only this projective structure [6, 7, 8]. In fact, we show that the major ambiguity in the reconstruction is a simple depth scale uncertainty, i.e., the classic *bas-relief* ambiguity which exists for two-frame structure from motion under orthographic projection [9].

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 on-line (recursive) estimation algorithms.

Our paper is structured as follows. After reviewing previous work, we present our formulation of the structure from motion problem and develop our technique for analyzing ambiguities using eigenvector analysis of the information (Hessian) matrix. We then present the results of our analysis for two different camera models: 1-D orthographic cameras and 2-D perspective cameras (more examples and results are presented in [10]). We conclude with a discussion of the main sources of errors and ambiguities, and directions for possible future work.

2 Previous work

Structure from motion has been extensively studied in computer vision. Early papers on this subject develop algorithms to compute the structure and motion from a small set of points matched in two frames using an *essential parameter* approach [1]. The performance of this approach can be significantly improved using non-linear least squares (*optimal estimation*) techniques [2, 3]. More recent research focuses on extraction of shape and motion from longer image sequences using both batch and recursive (Kalman filter) formulations [4, 5, 11, 12, 13]. Another line of research has addressed recovering affine [14] or projective [6, 7, 8] structure estimates. For a more detailed review of related work, please see [4, 10].

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 [3]. Adiv [15] and Young and Chellappa [16] analyze continuous-time (optical flow) based algorithms using the concept of the Cramer-Rao lower bound. Oliensis and Thomas [17] 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 multi-frame optimal structure from motion algorithms, which have not been studied in great detail.

3 Problem formulation and uncertainty analysis

The equation which projects the ith 3-D point $\mathbf{p}_i$ (given time-varying motion parameters $\mathbf{m}_j$) into the jth frame at location $\mathbf{u}_{ij}$ is

$$\mathbf{u}_{ij} = \mathcal{P}\left(T(\mathbf{p}_i, \mathbf{m}_j)\right). \tag{1}$$

The perspective projection $\mathcal{P}$ (defined below) is applied to a rigid transformation

$$T(\mathbf{p}_i, \mathbf{m}_j) = \mathbf{R}_j \mathbf{p}_i + \mathbf{t}_j, \tag{2}$$

where $\mathbf{R}_j$ is a rotation matrix and $\mathbf{t}_j$ is a translation applied after the rotation. A variety of alternative representations are possible for the rotation matrix [18]. In this paper, we represent the rotation matrix as a function of a quaternion, since this representation has no singularities.

The standard perspective projection equation used in computer vision is

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathcal{P}_1 \begin{pmatrix} x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} f\frac{x}{z} \\ f\frac{y}{z} \end{pmatrix}, \tag{3}$$

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 [4] is

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathcal{P}_2 \begin{pmatrix} x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} s\frac{x}{1+\eta z} \\ s\frac{y}{1+\eta z} \end{pmatrix}. \tag{4}$$

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

In our previous work, 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 [2, 3, 4, 12]. The Levenberg-Marquardt method is a standard non-linear least squares technique [19] which minimizes

$$C(\mathbf{a}) = \sum_i \sum_j c_{ij} \left|\tilde{\mathbf{u}}_{ij} - \mathbf{f}_{ij}(\mathbf{a})\right|^2, \tag{5}$$

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 $\mathbf{a}$ contains the 3-D points $\mathbf{p}_i$, the motion parameters $\mathbf{m}_j$, and any additional unknown calibration parameters. The weight c_{ij} in (5) describes the confidence in measurement $\mathbf{u}_{ij}$.

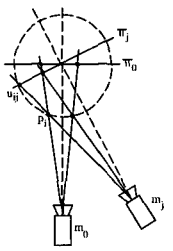


Fig. 1. Sample configuration of cameras ($\mathbf{m}_j$), 3-D points ($\mathbf{p}_i$), image planes(Π_j), and screen locations ($\mathbf{u}_{ij}$)

3.1 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 $\mathbf{A}$. This matrix is formed by computing outer products of the *Jacobians* of the measurement equations

$$\mathbf{A} = \sum_i \sum_j c_{ij} \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{a}} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{a}^T}. \tag{6}$$

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 $\frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{a}}$.

The $\mathbf{A}$ matrix has the structure

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A_p} & \mathbf{A_{pm}} \\ \hline \mathbf{A_{pm}}^T & \mathbf{A_m} \end{array}\right]. \tag{7}$$

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

$$\mathbf{A}_{\mathbf{p}_i} = \sum_j \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{p}_i} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{p}_i^T} \quad \text{and} \quad \mathbf{A}_{\mathbf{m}_j} = \sum_i \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{m}_j} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{m}_j^T}, \tag{8}$$

respectively (assuming $c_{ij} = 1$), while $\mathbf{A_{pm}}$ is dense, with entries

$$\mathbf{A}_{\mathbf{p}_i \mathbf{m}_j} = \frac{\partial \mathbf{f}_{ij}^T}{\partial \mathbf{p}_i} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{m}_j^T}. \tag{9}$$

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 [15, 16]. 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.

3.2 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. In the case of RMS reconstruction error, the positional uncertainty matrix $\mathbf{C}_{\mathbf{p}_i}$ can be computed by inverting $\mathbf{A}$ and looking at its upper left block (the block corresponding to the $\mathbf{p}_i$ variables).[1] The RMS reconstruction and motion error can also be computed directly by summing the *inverse* eigenvalues of the information $\mathbf{A}$.

What is the advantage of the second 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, 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.

3.3 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 (prior) knowledge, it is subject to certain ambiguities. The most fundamental (but most innocuous) of these is the coordinate frame (also known as pose, or Euclidean) ambiguity, i.e., we can move the origin of the coordinate system to an arbitrary place and pose and still obtain an equally valid solution.

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. 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 3.1. Absolute ambiguities will show

[1] Note that this is *not* the same as simply inverting $\mathbf{A}_{\mathbf{p}}$.

up as zero eigenvalues (unless we add additional constraints or knowledge to remove them), whereas weak ambiguities will show up as small eigenvalues.

4 Orthography: single scanline

Let us begin our analysis with an orthographic scanline camera, where the unknowns are the 2-D point positions $\mathbf{p}_i = (x_i, z_i)$ and the rotation angles θ_j.[2] The imaging equations are

$$u_{ij} = c_j x_i - s_j z_i \tag{10}$$

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

The Jacobian for the 1-D orthographic camera is

$$\mathbf{H}_{ij} = \left[\frac{\partial u_{ij}}{\partial x_i} \; \frac{\partial u_{ij}}{\partial z_i} \middle| \frac{\partial u_{ij}}{\partial \theta_j} \right]^T = \left[c_j \;\; -s_j \middle| -(c_j z_i + s_j x_i) \right]^T, \tag{11}$$

and the entries in the information matrix are

$$\mathbf{A}_{\mathbf{p}_i} = \begin{bmatrix} \sum_j c_j^2 & -\sum_j c_j s_j \\ -\sum_j c_j s_j & \sum_j s_j^2 \end{bmatrix} = \begin{bmatrix} C & -D \\ -D & S \end{bmatrix}, \tag{12}$$

$$\mathbf{A}_{\mathbf{p}_i \mathbf{m}_j} = \begin{bmatrix} -c_j^2 z_i - c_j s_j x_i \\ c_j s_j z_i + s_j^2 x_i \end{bmatrix}, \tag{13}$$

$$\mathbf{A}_{\mathbf{m}_j} = \left[\sum_i (c_j z_i + s_j x_i)^2 \right] = \left[c_j^2 Z + 2 c_j s_j W + s_j^2 X \right], \tag{14}$$

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

Before analyzing the complete information matrix, let us look at the two subblocks $\mathbf{A_p}$ and $\mathbf{A_m}$. 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 $\sigma_z^2 \propto S^{-1}$ (note that for small rotations, σ_x^2 is generally much smaller than σ_z^2). 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, the traditional scale ambiguity does not exist. However, translations along the optical axis cannot be estimated, and an overall pose (coordinate frame) ambiguity still exists. This manifests itself as the null (zero eigenvalue) eigenvector

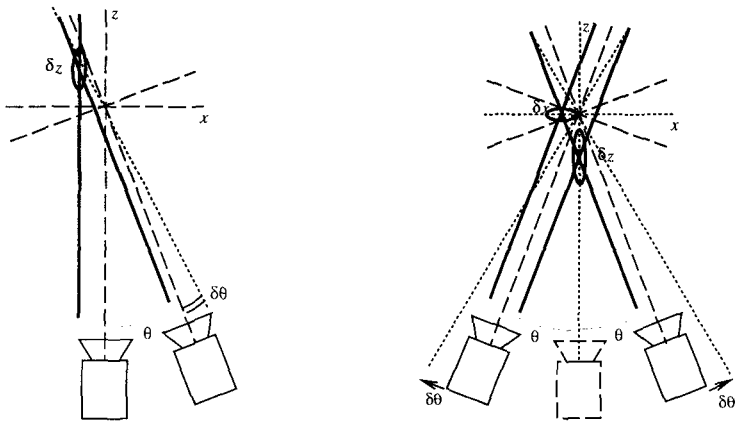
$$\mathbf{e}_0 = \left[z_0 \;\; -x_0 \;\; \cdots \;\; z_N \;\; -x_N \middle| 1 \;\; \cdots \;\; 1 \right]^T.$$

4.1 Two frames: the bas-relief ambiguity

Let us say we only have two frames, and we have fixed $\theta_0 = 0, c_0 = 1, s_0 = 0, \theta_1 = \theta, c_1 = c, s_1 = s$ (Fig. 2). Then

$$\mathbf{A}_{\mathbf{p}_i} = \begin{bmatrix} 1 + c^2 & -cs \\ -cs & s^2 \end{bmatrix}, \mathbf{A}_{\mathbf{p}_i \mathbf{m}} = \begin{bmatrix} -c^2 z_i - csx_i \\ csz_i + s^2 x_i \end{bmatrix}, \mathbf{A}_{\mathbf{m}} = \left[c^2 Z + 2csW + s^2 X \right]. \tag{15}$$

[2] We do not estimate the horizontal translation since it can be determined from the motion of the centroid of the image points [5].



(a) stationary first frame (b) antisymmetric cameras

Fig. 2. 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 $\delta\theta$).

The bas-relief ambiguity manifests itself as a null eigenvector

$$\mathbf{e}_0 = \begin{bmatrix} 0 & cz_0 + sx_0 & 0 & \cdots & cz_N + sx_N \end{bmatrix} \begin{bmatrix} -s \end{bmatrix}^T .$$

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). Note however that this uncertainty is proportional to $cz + sx$ rather than z, as can be seen by inspecting Fig. 2a.

An alternative parameterization of the two-frame problem is to set $\theta_0 = -\theta_1$ (Fig. 2b), in which case the null eigenvector is

$$\mathbf{e}_0 = \begin{bmatrix} s^2x_0 & -c^2z_i & \cdots & s^2x_N & -c^2z_N \end{bmatrix} \begin{bmatrix} cs \end{bmatrix}^T . \tag{16}$$

It shows that the primary effect of the bas-relief ambiguity is a "squashing" of the z values for a small increase in motion, with a much smaller "bulging" in the x values.[3] This squashing and bulging is an affine deformation of the true structure.

4.2 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

[3] Note that the total interframe rotation is now 2θ.

Fig. 2b with more cameras). In this case, we have

$$\mathbf{H}_{ij}^T = \left[c_j \; -s_j \middle| -j(c_j z_i + s_j x_i) \right] \tag{17}$$

$$\mathbf{A}_{\mathbf{p}_i} = \begin{bmatrix} \sum_j c_j^2 & 0 \\ 0 & \sum_j s_j^2 \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & S \end{bmatrix}, \tag{18}$$

$$\mathbf{A}_{\mathbf{p}_i\mathbf{m}} = \begin{bmatrix} -\sum_j j c_j s_j x_i \\ \sum_j j c_j s_j z_i \end{bmatrix} = \begin{bmatrix} -E x_i \\ E z_i \end{bmatrix}, \tag{19}$$

$$\mathbf{A}_{\mathbf{m}} = \left[\sum_j j^2 c_j^2 Z + \sum_j j^2 s_j^2 X \right] = \left[C' Z + S' X \right], \tag{20}$$

with $E = \sum_j j c_j s_j$, $C' = \sum_j j^2 c_j^2$, $S' = \sum_j j^2 s_j^2$, and C, D, S, Z, W, X defined as in (13–14). In this case, the smallest eigenvalue eigenvector has the form

$$\mathbf{e}_0 = \left[\alpha x_0 \;\; -\beta z_0 \;\; \cdots \;\; \alpha x_N \;\; -\beta z_N \;\middle|\; 1 \right]^T. \tag{21}$$

This will be an eigenvector if we can satisfy the matrix equation $\mathbf{A}\mathbf{e} = \lambda \mathbf{e}$, i.e.,

$$\left[\begin{array}{c|c} \mathbf{A_p} & \mathbf{A_{pm}} \\ \hline \mathbf{A_{pm}^T} & \mathbf{A_m} \end{array}\right] \left[\begin{array}{c} \alpha x_0 \\ -\beta z_0 \\ \vdots \\ -\beta z_N \\ \hline 1 \end{array}\right] = \lambda \left[\begin{array}{c} \alpha x_0 \\ -\beta z_0 \\ \vdots \\ -\beta z_N \\ \hline 1 \end{array}\right],$$

which reduces to the following three equations:

$$\begin{aligned} \alpha C - E &= \alpha\lambda \\ \beta S - E &= \beta\lambda \\ (S' - \alpha E)X + (C' - \beta E)Z &= \lambda. \end{aligned}$$

Substituting $\alpha = \frac{E}{C-\lambda}$ and $\beta = \frac{E}{S-\lambda}$ into the third equation, we obtain a cubic in λ,

$$(S-\lambda)(S'(C-\lambda)-E^2)X+(C-\lambda)(C'(S-\lambda)-E^2)Z-(S-\lambda)(C-\lambda)\lambda = 0, \tag{22}$$

which can be solved analytically using a package such as *Mathematica*® [20].

Assuming that the smallest eigenvalue is very small, we can use the approximation $\alpha \approx \frac{E}{C}$ to obtain a quadratic in λ,

$$(S - \lambda)(S'C - E^2)X + C(C'(S - \lambda) - E^2)Z - (S - \lambda)C\lambda = 0. \tag{23}$$

Furthermore, using the small angle approximations, $C \approx \sum_j 1 \equiv J_0$, $S \approx \Delta\theta^2 J_2$, $E \approx \Delta\theta J_2$, $C' \approx J_2$, and $S' \approx \Delta\theta^2 J_4$, where $J_2 = \sum_j j^2$ and $J_4 = \sum_j j^4$, we obtain after some manipulation

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

Notice that the minimum eigenvalue is related to the fourth power of $\Delta\theta$, i.e., doubling the inter-frame rotation reduces the RMS error by a factor of 4

$\lambda_{\min}$	$F=2$	$F=3$	$F=4$	$F=5$	$F=6$	$F=7$	$F=8$
$\theta_{\text{tot}}=11.5°$	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 1. Minimum eigenvalues for 1-D orthographic known equi-angular motion

(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 is somewhat surprising, and suggests that flatter objects can be reconstructed better.

We can numerically compute the values of λ for a range of J and $\Delta\theta$ values. For example, with $J=1$, $\Delta\theta=0.1$ rad $\approx 6°$, and $X=Z=1$, we have $\lambda = \{0.0000664436, 1.98064, 3.0193\}$. For the smallest eigenvalue, $\lambda = 0.0000664436$, we have a corresponding $\alpha = 0.0666676$ and $\beta = 10.0001$.

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 $\mathbf{Apm}$ is an eigenvector, e.g.,

$$\mathbf{e} = \left[x_1 \;\; 0 \;\; -x_0 \;\; 0 \;\; \cdots \;\; 0 \;\middle|\; 0 \right].$$

The eigenvalues corresponding to the pure x eigenvectors are C, while the z eigenvalues are S. 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., exactly the expected triangulation error for known camera positions.

For the above example with $J=1$ (3 frames), $\Delta\theta=0.1$ rad $\approx 6°$, and $X=Z=1$, the values for C and S are 2.98 and 0.0199, respectively. From this, we see that the correlated depth uncertainty due to the motion uncertainty is a factor of $0.0199/0.00006644 = 300$ times greater than the individual depth uncertainties. A full table of $\lambda_{\min}$ as a function of $F = 2J+1$ (the number of frames) and $\theta_{\text{tot}} = (F-1)\Delta\theta$ (the total rotation angle) is shown in Table 1.

4.3 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. The equations for $\mathbf{A}_{\mathbf{p}_i}$, $\mathbf{A}_{\mathbf{p}_i\mathbf{m}_j}$, and $\mathbf{A}_{\mathbf{m}_j}$ are given in (13–14), with $D=0$. In this case, we do not have a closed form solution. However, performing a numerical eigenvalue analysis of the $\mathbf{A}$ matrix using a set of 9 points sampled on the unit square, i.e., $\{(x,z), x, z \in \{-1,0,1\}\}$, we obtain results that are very close to those shown in Table 1 (see [10] for details).

$\lambda_{\min}$	$F=2$	$F=3$	$F=4$	$F=5$	$F=6$	$F=7$	$F=8$
$\theta_{\text{tot}}=11.5°$	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 2. Minimum eigenvalues for 3-D perspective projection, equi-angular rotation around y axis, $\eta = 0.1$.

5 Perspective in 3-D

Our full-length paper contains analyses of orthography in 3-D and the perspective scanline model [10]. Due to space limitations, we now jump directly to the full 3-D perspective model. 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.

In this section, we briefly discuss results of numerical eigenvalue analysis of pure object-centered rotation (which in camera-centered coordinates is actually both rotation and translation), and pure forward translation. Ignoring the effects of motion across the retina, these two cases capture the basic motion cues available to structure from motion. In our experiments, we used a 15-point data set consisting of the 8 corners of a unit cube, the 6 cube faces, and the origin.

5.1 Mostly rotations

The computed eigenvalues for pure rotation are shown in Table 2. Compared to the orthographic case (Table 1), we see some striking differences. First, the two-frame problem is now soluble (up to a scale ambiguity, of course). Second, for small viewing angles, there is marked improvement even for multiple frames. Third, the results for large viewing angles with small η'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.

The tables of λ_{min} with varying η are presented in [10] for the two and three frame problems. For the two-frame case, doubling the amount of perspective distortion η results in a fourfold increase in $\lambda_{\min}$ (and hence a halving of the RMS error). For the three-frame case, the results are less sensitive to η.

For a typical minimum eigenvector (e.g., a three-frame problem with $\eta = 0.1$ and $\theta_{tot} = 11.5°$), the majority of the ambiguity is depth scaling. However, the eigenvector is not a pure affine transform of the 3-D coordinates (this has been verified numerically). Our conjecture is that the minimum eigenvector may be a *projective* transformation of the 3-D points, i.e., that the main ambiguity is projective, but we have not yet found a proof for this conjecture.

5.2 Looming

The motion of a camera forward in a 3-D world creates a different kind of parallax, which can also be exploited to compute structure from motion. To

compute the ambiguities in this kind of motion, we used the same approach as before, except with no rotation and pure forward motion ($t_z \neq 0$).

Using our usual 15-point data set results in some unexpected behavior: four of the eigenvalues are zero. This is because the z coordinates of the three points on the optical axis cannot be recovered as they lie on the focus of expansion. This is a severe limitation of recovering structure from looming: points near the focus of expansion are recovered with extremely poor accuracy. In our experiments, we use a 12-point data set instead, i.e., the 15-point set with the three points $(x, y) = (0, 0)$ removed. The numerical results can be found in [10].

In one set of experiments, we calculate $\lambda_{\min}$ as a function of the number of frames F and the total extent of forward motion t_z (the object being viewed is a unit cube with coordinates $[-1, 1]^3$). The two-frame results are almost as good as the three frame results with the same extent of motion. The value of $\lambda_{\min}$ appears to depend quadratically on the total extent of motion. Overall, however, these results are much worse than those available with object-centered rotation.

In another set of experiments, we calculate $\lambda_{\min}$ as a function of η, i.e., the amount of perspective distortion. It appears that $\lambda_{\min}$ depends cubically on η, at least for small t_zs. To obtain reasonable estimates, therefore, it is necessary to both use a wide field of view and a large amount of motion relative to the scene depth.

6 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. In fact, very few results of convincing quality are available. Those cases where metrically accurate results have been demonstrated almost always use a large amount of rotation [5].

This raises the obvious question: are current structure from motion algorithms of practical significance? The situation is perhaps not that bad. For large object rotations, we can indeed recover accurate reconstructions. Furthermore, 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 (e.g., what appears to be happening under perspective for large motions). Computing

the RMS_{pos} error directly from the covariance matrix $\mathbf{A}^{-1}$ would be more useful in these cases, and we plan to carry out this analysis.

6.1 Future work

We are currently performing an error analysis on the results of an optimal structure from motion algorithm [4] with noisy data to see if they agree with the errors predicted by our analysis. In particular, we are estimating the (scaled) metric, affine, and projective reconstruction errors to determine which kinds of errors dominate.

In future work, we plan to compare results available with object-centered and camera-centered representations (Equations 3–4). Our guess is that the former will produce estimates of better quality. Similarly, we would like to analyze the effects of mis-estimating internal calibration parameters such as focal length, and to study the feasibility of estimating them as part of the reconstruction process. The results presented here have assumed for now that feature points are visible in all images. Our approach generalizes naturally to missing data points. In particular, we would like to study the effects feature tracks with relatively short lifetimes.

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 [17] but which performs significantly better than algorithms which ignore such correlations.

7 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 3-D 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. H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.
2. M. E. Spetsakis and J. Y. Aloimonos. Optimal motion estimation. In *IEEE Workshop on Visual Motion*, pp. 229–237, 1989.
3. J. Weng, N. Ahuja, and T. S. Huang. Optimal motion and structure estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(9):864–884, 1993.
4. R. Szeliski and S. B. Kang. Recovering 3D shape and motion from image streams using nonlinear least squares. *J. Vis. Commun. and Image Repr.*, 5(1):10–28, 1994.
5. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int'l J. of Computer Vision*, 9(2):137–154, 1992.
6. O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Second European Conf. Computer Vision (ECCV'92)*, pp. 563–578, 1992
7. R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR'92)*, pp. 761–764, 1992.
8. R. Mohr, L. Veillon, and L. Quan. Relative 3D reconstruction using multiple uncalibrated images. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR'93)*, pp. 543–548, 1993.
9. H. C. Longuet-Higgins. Visual motion ambiguity. *Vision Research*, 26(1):181–183, 1986.
10. R. Szeliski and S. B. Kang. Shape ambiguities in structure from motion. Technical Report 96/1, Digital Equipment Corporation, Cambridge Research Lab, Cambridge, MA, 1996.
11. N. Cui, J. Weng, and P. Cohen. Extended structure and motion analysis from monocular image sequences. In *Third Int'l Conf. Computer Vision (ICCV'90)*, pp. 222–229, 1990.
12. C. J. Taylor, D. J. Kriegman, and P. Anandan. Structure and motion in two dimensions from multiple images: A least squares approach. In *IEEE Workshop on Visual Motion*, pp. 242–248, 1991.
13. A. Azarbayejani and A. P. Pentland. Recursive estimation of motion, structure, and focal length. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(6):562–575, 1995.
14. J. J. Koenderink and A. J. van Doorn. Affine structure from motion. *J. of the Optical Society of America A*, 8:377–385538, 1991.
15. G. Adiv. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(5):477–490, 1989.
16. G.-S. Y. Young and R. Chellappa. Statistical analysis of inherent ambiguities in recovering 3-d motion from a noisy flow field. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(10):995–1013, 1992.
17. J. Oliensis and J. I. Thomas. Incorporating motion error in multi-frame structure from motion. In *IEEE Workshop on Visual Motion*, pp. 8–13, 1991.
18. N. Ayache. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception.* MIT Press, Cambridge, MA, 1991.
19. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, England, 1992.
20. Stephen Wolfram. *Mathematica: A System for Doing Mathematics by Computer.* Addison-Wesley, Redwood City, CA, 1991.