

# Statistical Error Propagation in 3D Modeling From Monocular Video

Amit Roy Chowdhury, Rama Chellappa  
Center for Automation Research  
University of Maryland, College Park, MD 20742

## Abstract

A significant portion of recent research in computer vision has focused on issues related to sensitivity and robustness of existing techniques. In this paper, we study the classical structure from motion problem and analyze how the statistics representing the quality of the input video propagates through the reconstruction algorithm and affects the quality of the output reconstruction. Specifically, we show that it is possible to derive analytical expressions of the first and second order statistics (bias and error covariance) of the solution as a function of the statistics of the input. We concentrate on the case of reconstruction from a monocular video, where the small baseline makes any algorithm very susceptible to noise in the motion estimates from the video sequence. We derive an expression relating the error covariance of the reconstruction to the error covariance of the feature tracks in the input video. This is done using the implicit function theorem of real analysis and does not require strong statistical assumptions. Next, we prove that the 3D reconstruction is statistically biased, derive an expression for it and show that it is numerically significant. Combining these two results, we also establish a new bound on the minimum error in the depth reconstruction. We present the numerical significance of these analytical results on real video data.

## 1. Introduction

Despite the existence of numerous algorithms for structure from motion (SfM) [5, 10], constructing accurate 3D models reliably from images using SfM is still a challenging problem. Many researchers have analyzed the sensitivity and robustness of many of the existing algorithms. The work of Weng et al [26] is one of the earliest instances of estimating the standard deviation of the error in reconstruction using first order perturbations in the input. The Cramer-Rao lower bounds on the estimation error variance of the structure and motion parameters from a sequence of monocular images was derived in [2]. Young and Chellappa [27] derived bounds for the estimation error for structure and motion parameters for two images under perspective projection as well as from a sequence of stereo images. The

coupling of the translation and rotation for small field of view was studied in [3]. Zhang’s work [28] on determining the uncertainty in estimation of the fundamental matrix is another important contribution in this area. Haralick showed how well-known estimation techniques could be used to propagate additive random perturbations through different vision algorithms [9]. Soatto and Brockett [20] have analyzed SfM in order to obtain provably convergent and optimal algorithms. Oliensis emphasized the need to understand algorithm behavior and the characteristics of the natural phenomenon that is being modeled [16]. Ma, Kosecka and Sastry [13] also addressed the issues of sensitivity and robustness in their motion recovery algorithm. Sun, Ramesh and Tekalp [22] have proposed an error characterization of the factorization method for 3-D shape and motion recovery from image sequences using matrix perturbation theory. Morris and Kanatani have extended the covariance-based uncertainty calculations to account for the geometric indeterminacies like scale change [14]. A different source of error which has not received much attention in the computer vision community is the bias in the depth estimation. This has been observed by psycho-physicists who note “that it is hard to explain ... the existence of systematic biases in observers’ magnitude estimation of perceived depth” [23]. Some authors, notably Daniilidis and Spetsakis [3] and Kanatani [12], have proved that there exists a bias in the translation and rotation estimates from stereo.

In this paper, we study the problem of 3D modeling from a monocular video stream with a small baseline. There are two main sources of error for this problem: one arising from the geometrical indeterminacies and the second from statistical inaccuracies. Since the first source of errors has been well studied before [5, 10], we concentrate on the second one in this paper. For the reconstruction using a short baseline, the 2D motion estimates are often very small and even small perturbations in their estimation can lead to large errors in the camera motion and 3D reconstruction solutions. In the first part of this paper (Section 3), we show that it is possible to derive analytical expressions for the error covariance of the depth and motion estimates as a function of the error covariance of the optical flow or feature correspondence estimates, for the case of reconstruction from two frames. The result can be extended to obtain a multi-frame

error estimate, which is akin to the rate-distortion characteristic of information theory. This result does not require the assumptions of Gaussianity and is thus an extension of previous ones [27], and can be applied in practical scenarios where the noise is often non-Gaussian. In the second part of the paper, we establish a new result which shows that under many solution strategies often adopted for this problem, the 3D estimate is statistically biased and the bias is numerically significant. Recently, it has been proposed that the bias in the optical flow field can be a possible explanation for many geometrical optical illusions [6]. Our result takes this a step further. We show that as a consequence of the bias in the 2D motion estimates, the 3D reconstruction, estimated from this motion, is also statistically biased. This also leads us to modify some existing results on the minimum error variance of the structure estimate [27], since they assumed that the estimate was unbiased. We derive a *generalized* Cramer-Rao lower bound (CRLB) after incorporating the effect of this bias. We show the effect of the statistical analysis on real-life face reconstruction problems.

## 2. Problem Formulation

Consider a coordinate frame O-XYZ attached rigidly to a camera with the origin at the center of perspective projection and the Z-axis perpendicular to the image plane o-xy. Assume that the camera is in motion with respect to the rigid body imaged scene with translational velocity  $\mathbf{V} = [v_X, v_Y, v_Z]$  and rotational velocity  $\mathbf{\Omega} = [\omega_X, \omega_Y, \omega_Z]$  We assume that the camera motion between two consecutive frames in a video sequence is small, and use optical flow for motion field analysis. If  $p(x, y)$  and  $q(x, y)$  are the horizontal and vertical velocity fields of a point  $(x, y)$  in the image plane, they are related to the 3D object motion and scene depth by [15]

$$\begin{aligned} p(x, y) &= (x - fx_f)h(x, y) + \frac{1}{f}xy\omega_X - (f + \frac{1}{f}x^2)\omega_Y \\ &\quad + y\omega_Z \\ q(x, y) &= (y - fy_f)h(x, y) + (f + \frac{1}{f}y^2)\omega_X - \frac{1}{f}xy\omega_Y \\ &\quad - x\omega_Z, \end{aligned} \quad (1)$$

where  $h(x, y) = v_Z/z(x, y)$  is the scaled inverse scene depth,  $f$  is the focal length of the camera, and  $(x_f, y_f) = (\frac{v_X}{v_Z}, \frac{v_Y}{v_Z})$  is known as the *focus of expansion* (FOE). For  $N$  such points, normalizing linear distances with respect to the

focal length and defining <sup>1</sup>,

$$\begin{aligned} \mathbf{h} &= (h_1, h_2, \dots, h_N)_{N \times 1}^T \\ \mathbf{u} &= (p_1, q_1, p_2, q_2, \dots, p_N, q_N)_{2N \times 1}^T \\ \mathbf{r}_i &= (x_i y_i, -(1 + x_i^2), y_i)_{3 \times 1}^T \\ \mathbf{s}_i &= (1 + y_i^2, -x_i y_i, -x_i)_{3 \times 1}^T \\ \mathbf{\Omega} &= (\omega_X, \omega_Y, \omega_Z)_{3 \times 1}^T \\ \mathbf{Q} &= [r_1 \ s_1 \ r_2 \ s_2 \ \dots \ r_N \ s_N]_{2N \times 3}^T \\ \mathbf{P} &= \text{diag} \left[ \begin{array}{c} x_i - x_f \\ y_i - y_f \end{array} \right]_{2N \times N, i=1, \dots, N} \\ \mathbf{B} &= [\mathbf{P} \ \mathbf{Q}]_{2N \times (N+3)} \\ \mathbf{z} &= \left[ \begin{array}{c} \mathbf{h} \\ \mathbf{\Omega} \end{array} \right]_{(N+3) \times 1}, \end{aligned} \quad (2)$$

it can be shown that

$$\mathbf{u} = \mathbf{P}\mathbf{h} + \mathbf{Q}\mathbf{\Omega} = [\mathbf{P} \ \mathbf{Q}] \left[ \begin{array}{c} \mathbf{h} \\ \mathbf{\Omega} \end{array} \right] \triangleq \mathbf{B}\mathbf{z}. \quad (3)$$

We want to compute  $\mathbf{z}$  from  $\mathbf{u}$ .

Consider the cost function which minimizes the reprojection error (i.e. bundle adjustment)

$$C = \sum_{i=1}^N [(p_i - \hat{p}_i)^2 + (q_i - \hat{q}_i)^2] = \frac{1}{2} \sum_{i=1}^n C_i^2 \quad (4)$$

where  $b_{ij}$  is the  $(i, j)$ <sup>th</sup> element of  $\mathbf{B}$ .  $(\hat{p}_i, \hat{q}_i)$  are the projections of the depth and motion estimates,  $\mathbf{z}$ , onto the image plane and are obtained from the right hand side of the equations (1). The above mentioned cost function requires a non-linear optimization, which rarely gives a good solution unless a very good initial condition is available. Different methods have been proposed to deal with this. These involve estimating the camera motion first followed by the depth, recursively updating the camera motion and the depth one at a time using the previously available estimate of the other, etc. [10]. For the case of reconstruction from monocular video that we deal with, it is often possible to estimate the FOE from the first two/three frames and assume it to be constant over the next few which are used to reconstruct the structure. Knowledge of the FOE makes the system of equations in (3) linear, because of the bilinear parameterization in (1). Thus, this is an important special case of the 3D reconstruction problem and will receive significant attention in this paper.

<sup>1</sup>The  $i$ <sup>th</sup> point is represented by the subscript  $i$ .

Diagonal matrices will be very frequently used in the calculations. A diagonal matrix of size  $N \times N$  consisting of the diagonal terms  $a_1, \dots, a_N$  will be represented as  $\text{diag} [a_1, \dots, a_N]$  or  $\text{diag} [a_i]_{i=1, \dots, N}$ .

### 3 Error Covariance Calculation

We now state a result which gives a precise relationship between the error in image correspondences  $\mathbf{R}_u$  and the error in depth and motion estimate  $\mathbf{R}_z$ . We will first assume that the FOE is known, thus leading to a linear system of equations in (3). We will then show how the results can be extended for the case where the FOE is unknown.

**Theorem 1** *Define*

$$\begin{aligned} A_{\bar{i}p} &= [-(x_{\bar{i}} - x_f)\mathbf{I}_{\bar{i}}(N) - \mathbf{r}_{\bar{i}}] = [A_{\bar{i}ph} | A_{\bar{i}pm}] \\ A_{\bar{i}q} &= [-(y_{\bar{i}} - y_f)\mathbf{I}_{\bar{i}}(N) - \mathbf{s}_{\bar{i}}] = [A_{\bar{i}qh} | A_{\bar{i}qm}] \end{aligned} \quad (5)$$

where  $\bar{i} = \lceil i/2 \rceil$  is the upper ceiling of  $i$  ( $\bar{i}$  will then represent the number of feature points  $N$  and  $i = 1, \dots, n = 2N$ ) and  $\mathbf{I}_n(N)$  denotes a 1 in the  $n^{\text{th}}$  position of the array of length  $N$  and zeros elsewhere. The subscript  $p$  in  $A_{\bar{i}p}$  and  $q$  in  $A_{\bar{i}q}$  denotes that the elements of the respective vectors are derived from the  $p^{\text{th}}$  and  $q^{\text{th}}$  components of the motion in (1). Then

$$\begin{aligned} \mathbf{R}_z &= \mathbf{H}^{-1} \left( \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}} \mathbf{R}_u \frac{\partial C_i^T}{\partial \mathbf{u}} \frac{\partial C_i}{\partial \mathbf{z}} \right) \mathbf{H}^{-T} \\ &= \mathbf{H}^{-1} \left( \sum_{\bar{i}=1}^N (A_{\bar{i}p}^T A_{\bar{i}p} R_{u\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q} R_{u\bar{i}q}) \right) \mathbf{H}^{-T}, \end{aligned}$$

and

$$\mathbf{H} = \sum_{\bar{i}=1}^N (A_{\bar{i}p}^T A_{\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q}), \quad (6)$$

where  $\mathbf{R}_u = \text{diag}[R_{u\bar{i}p}, R_{u\bar{i}q}]_{\bar{i}=1, \dots, N}$ .

#### 3.1 Proof of Error Covariance Result

We use the implicit function theorem [25] to prove the above result. It has been used previously in vision for the derivation of the uncertainty in the fundamental matrix [5] and for establishing partial results on the uniqueness of the structure and motion parameters when a long sequence is used [1]. It was used in [7] for error calculations in medical imaging applications. We use it here to derive explicit expressions for error covariance in terms of the parameters of (1).

**Implicit Function Theorem:** The implicit function theorem states that if  $f$  is a continuously differentiable mapping,  $f(x, y) = 0$  can be solved uniquely for  $y$  in terms of  $x$  under certain conditions. We state the theorem precisely as described by Rudin in [25].

Let  $\mathbf{f}$  be a  $\mathcal{C}'$  mapping of an open set  $E \subset \mathbb{R}^{n+m}$  into  $\mathbb{R}^n$ , such that  $\mathbf{f}(\mathbf{a}, \mathbf{b}) = \mathbf{0}$  for some point  $(\mathbf{a}, \mathbf{b}) \in E$ . Put  $A = \mathbf{f}'(\mathbf{a}, \mathbf{b})$  and assume that  $A_x$  (the derivative matrix of

$\mathbf{f}$  with respect to its first argument  $\mathbf{x} \in \mathbb{R}^n$ ) is invertible. Then there exist open sets  $U \in \mathbb{R}^{n+m}$  and  $W \in \mathbb{R}^m$ , with  $(\mathbf{a}, \mathbf{b}) \in U$  and  $\mathbf{b} \in W$ , having the following property: To every  $\mathbf{y} \in W$  there corresponds a unique  $\mathbf{x}$  such that  $\mathbf{f}(\mathbf{g}(\mathbf{y}), \mathbf{y}) = \mathbf{0}$  and

$$\mathbf{g}'(\mathbf{b}) = -(A_x)^{-1} A_y. \quad (7)$$

Let  $\mathbf{z} = \psi(\mathbf{u})$ . Expanding  $\psi$  in a Taylor series around  $E[\mathbf{u}]$ ,

$$\psi(\mathbf{u}) = \psi(E[\mathbf{u}]) + D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}]) + \mathcal{O}(\mathbf{u} - E[\mathbf{u}])^2, \quad (8)$$

where  $\mathcal{O}(x^2)$  denotes terms of order 2 or higher in  $\mathbf{x}$  and  $D_\psi(\mathbf{x}) = \frac{\partial \psi}{\partial \mathbf{x}}$ . Up to a first-order approximation,

$$\psi(\mathbf{u}) - \psi(E[\mathbf{u}]) = D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}]). \quad (9)$$

The covariance of  $\mathbf{z}$  can then be written as

$$\begin{aligned} \mathbf{R}_z &= E[(\psi(\mathbf{u}) - E[\psi(\mathbf{u})])(\psi(\mathbf{u}) - E[\psi(\mathbf{u})])^T] \\ &= E[D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}])^T (D_\psi(E[\mathbf{u}]))^T] \\ &= D_\psi(E[\mathbf{u}]) \mathbf{R}_u D_\psi(E[\mathbf{u}])^T, \end{aligned} \quad (10)$$

where we have used the first order approximation that  $E[\mathbf{z}] = \psi(E[\mathbf{u}])$ .

For our problem, we desire to obtain our parameter of interest  $\mathbf{z}$  by minimizing  $C$  in (4). Choosing  $\mathbf{a} = E[\mathbf{z}]$  and  $\mathbf{b} = E[\mathbf{u}]$  (this is the point at which all the derivatives are computed), let

$$\phi = \frac{\partial C^T}{\partial \mathbf{z}}, \quad \text{and} \quad \mathbf{H} = \frac{\partial \phi}{\partial \mathbf{z}}. \quad (11)$$

$\phi$  is a  $m \times 1$  vector and  $\mathbf{H}$  is a symmetric  $m \times m$  matrix. Then from the implicit function theorem

$$D_\psi(\mathbf{u}) = -\mathbf{H}^{-1} \frac{\partial \phi}{\partial \mathbf{u}}. \quad (12)$$

Thus (10) becomes

$$\mathbf{R}_z = \mathbf{H}^{-1} \frac{\partial \phi}{\partial \mathbf{u}} \mathbf{R}_u \frac{\partial \phi^T}{\partial \mathbf{u}} \mathbf{H}^{-T}. \quad (13)$$

Then from (4) and (11),

$$\begin{aligned} \phi &= \frac{\partial C^T}{\partial \mathbf{z}} = \sum_i C_i \frac{\partial C_i^T}{\partial \mathbf{z}} \\ \mathbf{H} &= \frac{\partial \phi}{\partial \mathbf{z}} = \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{z}} + \sum_i \frac{\partial^2 C_i^T}{\partial \mathbf{z}^2} \\ &\approx \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{z}} \\ \frac{\partial \phi}{\partial \mathbf{u}} &\approx \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}}. \end{aligned} \quad (14)$$

Thus equation (13) becomes

$$\mathbf{R}_z = \mathbf{H}^{-1} \left( \sum_{ij} \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}} \mathbf{R}_u \frac{\partial C_j^T}{\partial \mathbf{u}} \frac{\partial C_j}{\partial \mathbf{z}} \right) \mathbf{H}^{-T}, \quad (15)$$

which gives a precise relationship between the uncertainty of the image correspondences  $\mathbf{R}_u$  and the uncertainty of the depth and motion estimates  $\mathbf{R}_z$ . Substituting our cost function from (4), we get

$$\frac{\partial C_i}{\partial \mathbf{z}} = \begin{cases} A_{\bar{i}p}, & i \text{ odd} \\ A_{\bar{i}q}, & i \text{ even} \end{cases}, \quad (16)$$

as a  $1 \times (N+3)$  dimensional vector and

$$\begin{aligned} \frac{\partial C_i}{\partial \mathbf{u}} &= \left[ \frac{\partial C_i}{\partial p_1} \quad \frac{\partial C_i}{\partial q_1} \quad \dots \quad \frac{\partial C_i}{\partial p_N} \quad \frac{\partial C_i}{\partial q_N} \right], \\ &= \mathbf{I}_i(2N), \end{aligned} \quad (17)$$

as a  $1 \times 2N$  dimensional array. Hence the Hessian in (14) becomes

$$\mathbf{H} = \sum_{\bar{i}=1}^N (A_{\bar{i}p}^T A_{\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q}). \quad (18)$$

The above expression can be represented as  $\mathbf{H} = \mathbf{B}^T \mathbf{B}$ , which can be derived by vector calculus techniques. However, as is clear from (14), the expression for the Hessian in (18) is an approximation from the implicit function theorem. This method of derivation allows easy extension to the unknown FOE (and thus more general) case, where the advantages of a linear system are lost.

Assume that the feature points as well as the components of the motion vector at each feature point are uncorrelated with each other, i.e.  $\mathbf{R}_u = \text{diag}[R_{u\bar{i}p}, R_{u\bar{i}q}]_{\bar{i}=1, \dots, N}$ . Then we can obtain a simpler relationship for the error covariances in (15):

$$\begin{aligned} \mathbf{R}_z &= \mathbf{H}^{-1} \left( \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}} \mathbf{R}_u \frac{\partial C_i^T}{\partial \mathbf{u}} \frac{\partial C_i}{\partial \mathbf{z}} \right) \mathbf{H}^{-T} \\ &= \mathbf{H}^{-1} \left( \sum_{\bar{i}=1}^N (A_{\bar{i}p}^T A_{\bar{i}p} R_{u\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q} R_{u\bar{i}q}) \right) \mathbf{H}^{-T}. \end{aligned} \quad (19)$$

Equations (18) and (19) prove the statement of Theorem 1.

**Corollary 1** *If we make the even stronger assumption that the components of  $\mathbf{R}_u$  are all identical (with variance  $r^2$ ), i.e.  $\mathbf{R}_u = r^2 \mathbf{I}_{2N \times 2N}$ , then (19) simplifies to*

$$\begin{aligned} \mathbf{R}_z &= \mathbf{H}^{-1} (r^2 \mathbf{H}) \mathbf{H}^{-T} \\ &= r^2 \mathbf{H}^{-1}. \end{aligned} \quad (20)$$

This is precisely the expression for the covariance and Cramer-Rao lower bound (CRLB) derived in [27] under an IID Gaussian noise assumption, which is a special case of (19). It should be noted that the assumption of uncorrelatedness of the noise in the features is invoked only at the end of the calculations. Thus, depending upon the validity of the assumptions, different expressions for the covariance in (15), (19) or (20) can be used. The  $\mathbf{R}_z$  thus obtained has an interesting structure as a result of our partitioning the vectors  $A_{\bar{i}p}$  and  $A_{\bar{i}q}$  into structure and motion components. It is easy to show that we can obtain a partition for  $\mathbf{R}_z$  as

$$\mathbf{R}_z = \begin{bmatrix} \mathbf{R}_h & \mathbf{R}_{hm} \\ \mathbf{R}_{hm}^T & \mathbf{R}_m \end{bmatrix} \quad (21)$$

**Corollary 2** *When the focus of expansion in (1) is unknown, the vector  $\mathbf{z} = [\mathbf{h}, x_f, y_f, \Omega]^T = [\mathbf{h}, \mathbf{m}]^T$ . The form of the error covariance result remains exactly the same as in (19) except that we need to redefine the two vectors  $A_{\bar{i}p}$  and  $A_{\bar{i}q}$  as follows:*

$$\begin{aligned} A_{\bar{i}p} &= [-(x_{\bar{i}} - x_f) \mathbf{I}_{\bar{i}}(N) \quad | \quad h_{\bar{i}} \quad 0 \quad -\mathbf{r}_{\bar{i}}], \\ A_{\bar{i}q} &= [-(y_{\bar{i}} - y_f) \mathbf{I}_{\bar{i}}(N) \quad | \quad 0 \quad h_{\bar{i}} \quad -\mathbf{s}_{\bar{i}}], \end{aligned} \quad (22)$$

A very important distinction for the unknown FOE case compared to the known FOE one is that  $A_{\bar{i}p}$  and  $A_{\bar{i}q}$  are now functions of the inverse depth estimates  $h_{\bar{i}}$ .

The covariance of the feature points,  $\mathbf{R}_u$ , is obtained using the standard method for estimating the error covariance using the inverse of the Hessian matrix of the second partial derivatives of the intensity along  $x$  and  $y$  axes [22].

Large systematic errors in feature correspondences are not considered in the above analysis. Such errors need to be addressed using robust estimation techniques. A detailed analysis of robust estimation of structure and motion using least median of squares can be found in [18].

## 3.2 Rate-Distortion Characteristic for Multiple Frames

We will extend the above result, obtained for the reconstruction for a pair of frames, to the situation where  $L$  such two-frame reconstructions are fused together. Let the two-frame inverse depth values for a particular feature point be denoted by  $X^1, X^2, \dots, X^L$ . We will derive an expression for the error in the multi-frame estimate as a function of the number of frames and the error in the image correspondences. Now  $E[\bar{X}] = \frac{1}{N} \sum_{i=1}^L E[X^i]$  and  $\text{Cov}[\bar{X}] = E[(\bar{X} - X^*)^2] = E[\bar{X}^2] - E[\bar{X}]^2$ . Then, under the assumption of independence of the two-frame observations, the covariance of the estimate of the  $j$ -th feature point is

$$\text{Cov}[\bar{X}_j] = \frac{1}{L^2} \left[ \sum_{i=1}^L \mathbf{R}_h^i(j, j) \right], \quad (23)$$

where  $\mathbf{R}_h^i(j, j)$  is the  $j$ -th diagonal term obtained from (21) for the  $i$ -th and  $(i + 1)$ -st frames. The average distortion in the reconstruction over  $N$  feature points and  $L$  frames (i.e. the rate-distortion characteristic) is

$$\begin{aligned} E_{L,N}[(\bar{X} - E[\bar{X}])^2] &= E_N[E_L[(\bar{X} - E[\bar{X}])^2 | \bar{X} = \bar{X}_j]] \\ &= \frac{1}{NL^2} \sum_{j=1}^N \sum_{i=1}^L \mathbf{R}_h^i(j, j) = \frac{1}{NL^2} \sum_{i=1}^L \text{trace}(\mathbf{R}_h^i) \end{aligned} \quad (24)$$

### 3.3 Experimental Results

The variance in the inverse depth computed using our theoretical analysis of Section 3 is shown in Figure 1(a). The diameters of the circles indicate the variance in the inverse depth estimate for the points which were tracked across the video sequence. The depth was computed using our algorithm (not described in this paper) described in [18]. It essentially consists of intelligently fusing the two-frame depth estimates.<sup>2</sup> A plot of the covariance matrix is also shown in the same figure so that it is possible to compare the relative magnitudes of the errors. The error covariance of the optical flow was estimated a priori over the first few frames of the video sequence, which were not used in the reconstruction. It was done over a sampled grid of points (rather than the dense flow) so as to simplify calculations. The technique used is similar to the gradient-based method of [22], except that, for more accurate results, it was repeated for each of these initial frames and the final estimate was obtained using bootstrapping techniques [4]. The rate-distortion characteristic for this particular sequence is shown in Figure 1(b) and an example reconstruction in (c). More detailed results can be found in [18] and [17].

## 4 Bias in 3D Reconstruction from Monocular Video

We now proceed to prove the second major result of this paper. Because of the fact that feature positions are never tracked perfectly, the 3D reconstruction, in most situations, is statistically biased and the bias is significant. We will state a precise expression for the bias and outline a proof of it in the Appendix.

As mentioned earlier, the solution of the cost function (4) involves a non-linear optimization which is extremely difficult, unless very good initial conditions are available. One of the common methods used is to estimate the camera motion and then the depth. Another strategy is to update the camera motion and the depth one at a time using

<sup>2</sup>For some points with relatively smooth texture, the variance is small, which is counter-intuitive. However, on close observation, it becomes clear that these regions have brighter illumination, and hence the points are tracked better. Also, extremely small variances are not reliable and are not used.

the previous estimate of the other, till a convergence criterion is reached [10]. For monocular video sequences, it is often possible to first estimate the direction of motion (i.e. FOE) and then estimate the depth and rotational motion [21]. These deterministic optimization methods will usually guarantee convergence to a local minimum. The point to note is that if we assume an estimate of the camera motion or FOE and solve for the depth or vice versa, we essentially are solving a linear system of equations. This can be easily seen from the bilinear parameterization of (1). It is a well known fact that the least squares solution to a linear system of the form  $Ax = b$  with errors in the system matrix  $A$  is biased [8]. When the SfM problem is posed in a least squares framework, the matrix  $A$  involves the image coordinates, which almost always have measurement errors. Thus it should be expected that the solution of the SfM problem would also have a bias. To the best of our knowledge, this is the first attempt to compute the bias in the depth reconstruction from monocular video, explicitly.

The actual value of the bias would be different for the different situations explained above. We consider the particular situation where the camera motion is known and derive the expression. We do this because one of the most common approaches to solving the 3D reconstruction problem is to first estimate the camera motion, and then the depth. Expressions for other algorithms can be similarly derived. For algorithms which estimate the parameters alternatively and update based on the previous estimates, the bias will propagate through the reconstruction strategy. For the case when the camera motion is known, (3) can be written as  $\mathbf{b} = \mathbf{A}\mathbf{h}$ , with  $\mathbf{A} \triangleq \mathbf{P}$  and  $\mathbf{b} \triangleq [p_1 - \mathbf{r}_1^T \boldsymbol{\Omega}, q_1 - \mathbf{s}_1^T \boldsymbol{\Omega}, \dots, p_N - \mathbf{r}_N^T \boldsymbol{\Omega}, q_N - \mathbf{s}_N^T \boldsymbol{\Omega}]^T$ . We now state the main result in the form of a theorem.

**Theorem 2** Consider the LS solution  $\hat{\mathbf{h}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$ . For convenience, let us define

$$\begin{aligned} \mathbf{M} &\triangleq \mathbf{A}^T \mathbf{A} = \text{diag} [(x_i - x_f)^2 + (y_i - y_f)^2] \\ &= \text{diag} [m_{ii}]_{i=1, \dots, N}, \end{aligned} \quad (25)$$

$$\begin{aligned} \mathbf{v} &\triangleq \mathbf{A}^T \mathbf{b} \triangleq [(x_i - x_f)v_{pi} + (y_i - y_f)v_{qi}]^T \\ &\triangleq [v_1, \dots, v_N]^T, \end{aligned} \quad (26)$$

where  $v_{pi} = p_i - \mathbf{r}_i^T \boldsymbol{\Omega}$  and  $v_{qi} = q_i - \mathbf{s}_i^T \boldsymbol{\Omega}$ . The bias in the inverse depth estimate,  $\hat{\mathbf{h}}$ , is  $b(\hat{\mathbf{h}}) = E[\hat{\mathbf{h}}] - \bar{\mathbf{h}}$ , where  $\bar{\mathbf{h}}$  is the true value. If  $\sigma_i^2 = E[\delta x_i^2] = E[\delta y_i^2]$  is the variance in the image coordinate measurements, then under the assumptions of the above formulation, the bias in the inverse depth estimate,  $\hat{\mathbf{h}}_i$ , of the  $i^{\text{th}}$  feature point is given

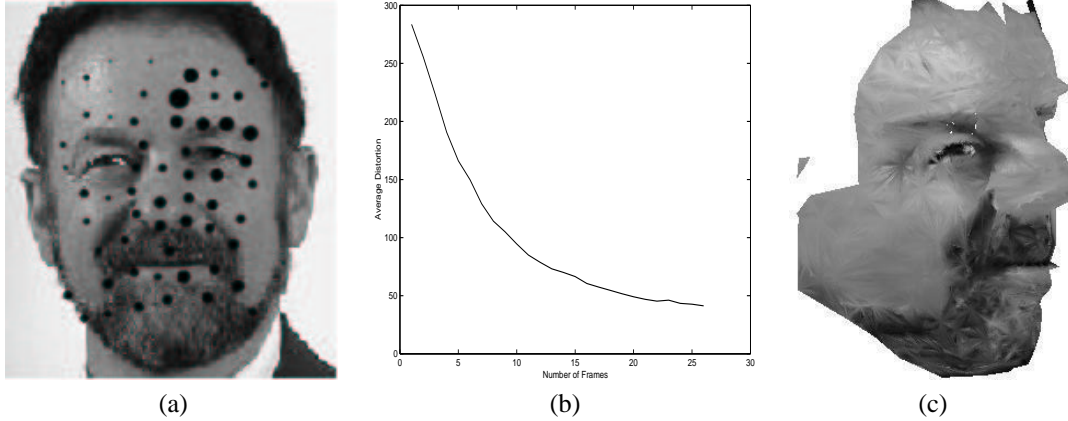


Figure 1: (a) Plot of the variance of the inverse depth for different features in a face sequence. The diameter of the circle at each feature point is proportional to the variance at that feature point. (b) represents the rate-distortion characteristic for this particular sequence. (c) depicts one view from the reconstructed model.

by

$$\begin{aligned}
 [\text{Bias}]_i &= b(\hat{\mathbf{h}}_i) = \frac{2\sigma_i^2 v_i}{m_{ii}^2} \\
 &+ \frac{2\sigma_i^2}{m_{ii}^2} [(x_i - x_f)^2 \mathbf{r}_{ix}^T + (y_i - y_f)^2 \mathbf{s}_{iy}^T] \Omega \\
 &+ \frac{2\sigma_i^2}{m_{ii}^2} [(x_i - x_f)(y_i - y_f)(\mathbf{s}_{ix}^T + \mathbf{r}_{iy}^T)] \Omega \\
 &+ \frac{\sigma_i^2}{m_{ii}} [(x_i - x_f)\omega_Y - (y_i - y_f)\omega_X - (\mathbf{r}_{ix}^T + \mathbf{s}_{iy}^T)\Omega], \tag{27}
 \end{aligned}$$

where  $f_{ix}$  represents the derivative of a function  $f_i$  with respect to  $x$ .

#### 4.1 Analysis of the Bias

**Effect of Camera Motion** The bias is a function of the camera motion parameters. It is most affected by the rotational motion of the camera. As can be seen from the expression in equation (27), the bias is negligibly small when the angular motion is zero (or very small).

**Bias Compensation** Once the structure and motion estimates are obtained, the bias can be computed and subtracted out of the estimate. For the biased estimate,  $E[\hat{\mathbf{h}}] = \bar{\mathbf{h}} + b(\hat{\mathbf{h}})$ , where  $\bar{\mathbf{h}}$  is the true value. If  $\hat{\mathbf{h}}_c = \hat{\mathbf{h}} - b(\hat{\mathbf{h}})$  is the bias compensated estimate, then  $E[\hat{\mathbf{h}}_c] = E[\hat{\mathbf{h}}] - b(\hat{\mathbf{h}}) = \bar{\mathbf{h}}$ , thus leading to an unbiased estimate.

**Bias and Total Least Squares (TLS)** The TLS has emerged as an alternative to least squares since it is capable of handling errors in both the observations,  $b$ , and the system variables,  $A$ , in a linear system

$Ax = b$ . However, the TLS estimate is unbiased only if the error in estimating  $A$  is equal in variance to the error in estimating  $b$  [24]. Such a condition would be very difficult to maintain in (1). Also, estimating the bias of a TLS estimate is extremely cumbersome. Also, as argued in [24], the covariance of an unbiased TLS estimate is larger than that of the LS estimate, in the first order approximation as well as in simulations. Hence, there is no fundamental gain in choosing the TLS over the LS solution. Thus using the TLS criterion cannot be a solution to the problem of bias in the 3D estimate. A very thorough analysis of bias in vision problems and the applicability of different techniques to remove the bias can be found in [6].

**Bias in Multi-frame Reconstruction** Suppose now that

we have  $L$  two-frame reconstructions. Let  $(\hat{\mathbf{h}}^1, \dots, \hat{\mathbf{h}}^L)$  be the two-frame estimates aligned with respect to a particular frame of reference. Let the true value be  $\bar{\mathbf{h}}$  and the bias in (27) be represented by  $(b(\hat{\mathbf{h}}^1), \dots, b(\hat{\mathbf{h}}^L))$ , i.e.  $E[\hat{\mathbf{h}}^i] = \bar{\mathbf{h}} + b(\hat{\mathbf{h}}^i)$ ,  $i = 1, \dots, L$ . Assume that the estimates and the true value have the same scale (so that the problem of scale ambiguity does not arise). Then the least squares estimate for the structure over all  $L$  observations is  $\hat{\mathbf{h}} = \frac{1}{L} \sum_{i=1}^L \hat{\mathbf{h}}^i$ . Taking expectations on both sides, we see that the bias in the multi-frame estimate is  $b^L(\hat{\mathbf{h}}) = \frac{1}{L} \sum_{i=1}^L b(\hat{\mathbf{h}}^i)$ , where  $b(\hat{\mathbf{h}}^i)$  is obtained from (27) for the  $i$  and  $(i+1)^{\text{st}}$  frame.

#### 4.2 Numerical Significance of the Bias

We first present some results on simulation data, as it will help in understanding the effect the bias can have on the

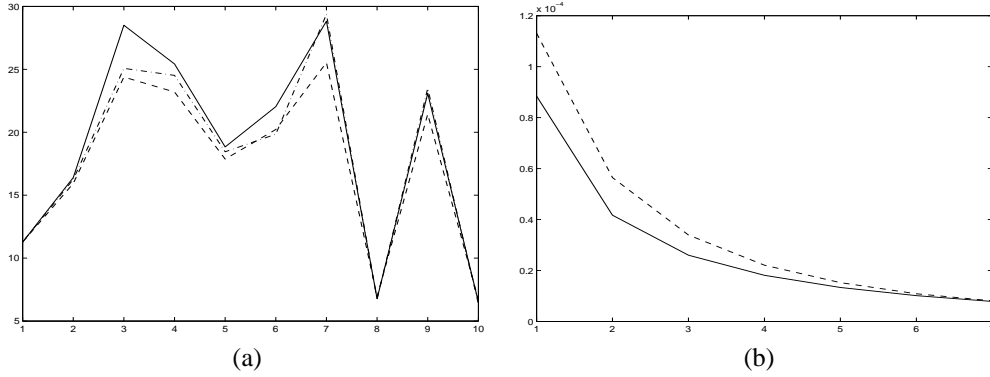


Figure 2: (a) Plot of the 3D reconstruction for a set of ten points tracked over 15 frames. The camera motion parameters are  $x_f = 10$ ,  $y_f = 10$ ,  $\omega_X = \omega_Y = \omega_Z = 1$ degree/frame. The solid lines indicate the true depth values, the dashed lines indicate the reconstruction without bias compensation, and the dashed and dotted lines indicate the reconstruction with bias compensation. (b) Plots of the CRLB of the inverse depth as a function of the number of frames for the same camera motion parameters. The solid line shows the CRLB without the bias and the dotted line shows the generalized CRLB with the bias accounted for. The number of frames shown on the x-axis is scaled down by a factor of 10.

depth values. In this set of experiments, we plotted the actual reconstruction estimate, with and without bias compensation, against the true depth. A set of 10 random 3D points were generated and their 2D projections were computed at different camera positions. The set of feature points were tracked across a few frames. Depth estimates from each pair of frames were obtained (by solving the least squares problem) and then combined together to get the maximum likelihood (ML) estimate over the entire sequence. To fix the scale of the reconstruction, the depth at the first point was used. In these experiments we considered the case of non-zero but constant linear and angular camera motion. Figure 2(a) shows the 3D reconstruction with and without bias compensation. It can be seen that bias compensation makes the estimate closer to the true value and gives significant advantages for some of the points. The error in the reconstruction by neglecting the bias can be as high as 20% of the true value.

We have also studied the impact of the bias on a practical 3D face reconstruction problem. We considered the database available on the World Wide Web at <http://sampl.eng.ohio-state.edu/sampl/data/3DDB/RID/minolta/faces-hands.1299/index.html>. The advantage of using this database is that it includes the true 3D depth model obtained from a range scanner. This allows us to compare the effect of the bias against the true value. For want of space, we show the results on one particular sequence, which is termed as “frame001” on the website. The depth was reconstructed using the algorithm (not described in this paper) described in [17]. The depth map of the subject is shown in the first column of Figure 3. The bias is shown in the form of a depth map in the second column of the same

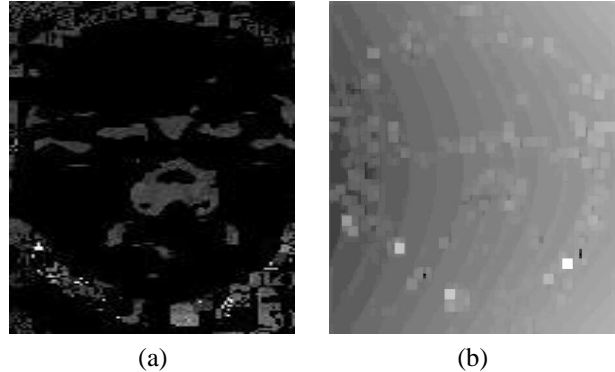


Figure 3: The reconstructed 3D depth maps are shown in the first column. The second column shows the estimated bias.

figure. The bias is a function of the camera motion; hence the particular values seen here are specific to this case and may be very different (better or worse) for a different kind of motion. We found that the average error (measured as the standard deviation of the estimated depth values from the true one), measured as a percentage of the true depth, did not change much depending on whether the bias was compensated for or not. The average error decreased from 3.8% to 3.6% for this particular case. However, the peak percentage bias was about 30% of the true depth, which was almost completely removed. Thus, even though the bias is concentrated at only a few points, it is important to compensate for them because even a single point can affect the reconstruction significantly.

## 5 Minimum Error Bounds for SfM

The minimum error variance (namely the CRLB) of SfM estimates was derived in [27]. This was done assuming that the estimate was unbiased. This is the usual case when deriving the CRLB of any estimate because it is difficult to know the bias of an estimator. The general expression for the CRLB after incorporating the bias in the estimate and under the proper regularity assumptions is [19]

$$\begin{aligned} \Sigma_\theta(g) &\geq b_\theta(g)b_\theta(g)^T \\ &+ (I + \nabla_\theta b_\theta(g))M^{-1}(\theta)(I + \nabla_\theta b_\theta(g))^T \end{aligned} \quad (28)$$

where  $g$  is the estimate of the parameter  $\theta$ ,  $b$  is the bias of the estimate,  $M$  the Fisher information (FI) matrix,  $I$  is an identity matrix of suitable size and  $\nabla_\theta$  is the gradient with respect to  $\theta$ .

We suggest a modification of the result in [27] after taking into account the effect of the bias. Let  $\hat{\mathbf{h}}$  denote the multi-frame estimate of  $\bar{\mathbf{h}}$  (the true value) with respect to a particular scale. Let the bias in the multi-frame estimate be denoted by  $b(\hat{\mathbf{h}})$ . Since the bias does not depend on  $\bar{\mathbf{h}}$ ,  $\nabla_{\bar{\mathbf{h}}}b(\hat{\mathbf{h}}) = 0$ . Let the FI matrix for multi-frame reconstruction, as derived in [27], be denoted by  $M(\bar{\mathbf{h}})$ . This derivation is based on the assumptions that the feature positions are exactly known and the motion estimates  $\{p_i, q_i\}$  are corrupted by additive white Gaussian noise. Since the estimate in this case was unbiased, the FI matrix was inverted to obtain the CRLB. This is essentially the expression of  $\mathbf{R}_{\mathbf{h}}$  as would be obtained from Corollary 1, by considering the upper  $N \times N$  matrix of  $\mathbf{R}_{\mathbf{z}}$ . Let us denote it by  $\mathbf{R}_{\mathbf{h}}^G$ , where  $G$  indicates that it is under the Gaussian assumption. Thus combining Theorems 1 and 2, we get that for the biased estimate  $\hat{\mathbf{h}}$ , the variance represented as  $\Sigma_{\hat{\mathbf{h}}}(\hat{\mathbf{h}}) = E[(\hat{\mathbf{h}} - \bar{\mathbf{h}})(\hat{\mathbf{h}} - \bar{\mathbf{h}})^T]$  must justify the following inequality (from (28)):

$$\Sigma_{\hat{\mathbf{h}}}(\hat{\mathbf{h}}) \geq b(\hat{\mathbf{h}})b^T(\hat{\mathbf{h}}) + \mathbf{R}_{\mathbf{h}}^G, \quad (29)$$

This is the minimum variance of the structure estimate obtained from a 3D reconstruction algorithm using optical flow and monocular video. A plot of the CRLB, for the data in Figure 2(a), with and without the bias factored in is shown in Figure 2(b). It shows that the percentage change in minimum possible error in reconstruction is significant.

## 6 Conclusion

In this paper, we have analyzed the propagation of errors in 3D reconstruction from monocular video with a small baseline in a statistical analysis framework. We have shown how the errors in estimating the motion between corresponding points in the video sequence are propagated through the 3D

reconstruction framework and affects the final reconstruction. Specifically, we have proved three results. The first relates the error covariance of the depth and motion estimates to the error covariance of the feature correspondence or optical flow. Secondly, we proved that the 3D reconstruction is also statistically biased and the bias is numerically significant. The two results were combined together in order to obtain a new minimum error bound on the reconstruction. These results can be used to evaluate the quality of reconstruction algorithms in real-life scenarios. The mathematical results were applied on actual video sequences in order to demonstrate their significance.

## A Outline of Proof of Theorem 2

We give a brief outline of the proof. The details can be found in [17].

Expanding  $\hat{\mathbf{h}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$  in a Taylor series around the true value  $\bar{\mathbf{h}}$ , (i.e. the noise  $N = 0$ ) and assuming the mean deviation in that region to be zero (i.e.  $E[\delta x_i] = E[\delta y_i] = E[\delta x_f] = E[\delta y_f] = 0$ ) and all the components  $\delta x_i, \delta y_i, \delta x_f, \delta y_f$  to be mutually uncorrelated, we can express

$$\begin{aligned} E[\hat{\mathbf{h}}] &\approx \bar{\mathbf{h}} + \sum_{i=1}^N \left[ \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta x_i^2} E\left[\frac{\delta x_i^2}{2}\right] + \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta y_i^2} E\left[\frac{\delta y_i^2}{2}\right] \right] \\ &+ \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta x_f^2} E\left[\frac{\delta x_f^2}{2}\right] + \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta y_f^2} E\left[\frac{\delta y_f^2}{2}\right], \end{aligned} \quad (30)$$

where all the partials are computed at  $N = 0$ . The above equation is actually a simplified version which does not take into account the errors in  $(p_i, q_i, \omega_X, \omega_Y, \omega_Z)$ . This is because  $\frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta p_i^2} = \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta q_i^2} = \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta \omega_X^2} = \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta \omega_Y^2} = \frac{\partial^2 \hat{\mathbf{h}}}{\partial \delta \omega_Z^2} = 0$ . In the absence of any measurement noise, the expected value of the estimate obtained from the least squares solution should equal the true value  $\bar{\mathbf{h}}$ . However, since there exist errors in the measurement model, the estimate is biased and the sum of the last four terms on the right hand side of (30) represents the total bias in the estimate. In order to calculate the bias, we need to compute the derivatives in (30). We can compute all the derivatives using the fact that for an arbitrary matrix  $Q$ ,  $-\frac{\partial Q^{-1}}{\partial x} = Q^{-1} \frac{\partial Q}{\partial x} Q^{-1}$  [11]. The final result can be obtained by computing the partials and substituting in (30).

## References

- [1] T.J. Broida. *Estimating the Kinematics and Structure of a Moving Object from a Sequence of Images*. PhD Thesis, 1985.

- [2] T.J. Broida and R. Chellappa. Performance bounds for estimating three-dimensional motion parameters from a sequence of noisy images. *Journal of the Optical Society of America A*, 6:879–889, 1989.
- [3] K. Daniilidis and M.E. Spetsakis. Understanding noise sensitivity in structure from motion. In *VisNav93*, 1993.
- [4] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.
- [5] O.D. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.
- [6] C. Fermuller, D. Shulman, and Y. Aloimonos. The statistics of optical flow. *Computer Vision and Image Understanding*, 82, 2001.
- [7] J. Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Transactions on Image Processing*, 5:493–506, 1996.
- [8] W.A. Fuller. *Measurement Error Models*. Wiley, 1987.
- [9] R.M. Haralick. Covariance propagation in computer vision. In *ECCV Workshop on Performance Characteristics of Vision Algorithms*, 1996.
- [10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
- [11] Thomas Kailath. *Linear Systems*. Prentice-Hall, 1980.
- [12] K.I. Kanatani. Unbiased estimation and statistical analysis of 3-d rigid motion from two views. *Pattern Analysis and Machine Intelligence*, 15(1):37–50, January 1993.
- [13] Y. Ma, J. Kosecka, and S. Sastry. Linear differential algorithm for motion recovery: A geometric approach. *International Journal of Computer Vision*, 36:71–89, January 2000.
- [14] D.D. Morris, K. Kanatani, and T. Kanade. Gauge fixing for accurate 3D estimation. In *Conference on Computer Vision and Pattern Recognition*, pages II:343–350, 2001.
- [15] Vishvijit Nalwa. *A Guided Tour of Computer Vision*. Addison Wesley, 1993.
- [16] J. Oliensis. A critique of structure from motion algorithms. Technical Report <http://www.neci.nj.nec.com/homepages/oliensis/>, NECI, 2000.
- [17] A. Roy Chowdhury. *Statistical Analysis of 3D Modeling From Monocular Video Streams*. PhD Thesis, 2002.
- [18] A. Roy Chowdhury and R. Chellappa. Stochastic approximation and rate-distortion analysis for robust structure and motion estimation. *Accepted to Int. Journal of Computer Vision*, 2003.
- [19] Jun Shao. *Mathematical Statistics*. Springer-Verlag, 1998.
- [20] S. Soatto and R. Brockett. Optimal structure from motion: Local ambiguities and global estimates. In *Conference on Computer Vision and Pattern Recognition*, pages 282–288, 1998.
- [21] S. Srinivasan. Extracting structure from optical flow using fast error search technique. *International Journal of Computer Vision*, 37:203–230, 2000.
- [22] Z. Sun, V. Ramesh, and A.M. Tekalp. Error characterization of the factorization method. *Computer Vision and Image Understanding*, 82:110–137, May 2001.
- [23] J.T. Todd. Theoretical and biological limitations on the visual perception of three-dimensional structure from motion. In *High Level Motion Processing: Computational, Neurobiological and Psychophysical Perspectives*, 1998.
- [24] S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem*. SIAM Frontiers in Applied Mathematics, 1991.
- [25] R. Walter. *Principles of Mathematical Analysis, 3rd Edition*. McGraw-Hill, 1976.
- [26] J. Weng, N. Ahuja, and T.S. Huang. Optimal motion and structure estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:864–884, September 1993.
- [27] G.S. Young and R. Chellappa. Statistical analysis of inherent ambiguities in recovering 3-d motion from a noisy flow field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14:995–1013, October 1992.
- [28] Z.Y. Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27:161–195, March 1998.