Uncertainty in 3D Shape Estimation

Hui Ji and Cornelia Fermüller
Center for Automation Research
University of Maryland
College Park, MD 20742-3275, USA
{jihui, fer}@cfar.umd.edu

Abstract

This paper analyses the uncertainty in the estimation of shape from different cues, specifically motion, stereo, and texture. It is shown that there are computational limitations of a statistical nature that previously have not been recognized. Because there is noise in all the input parameters, we cannot avoid bias. The analysis of shape from multiple views rests on a new constraint which relates image lines and rotation to shape. Because the human visual system has to cope with bias as well, it makes errors. This explains the underestimation of slant found in computational and psychophysical experiments, and demonstrated here for an illusory display. We discuss properties of the best known estimators with regard to the problem, as well as possible avenues for visual systems to deal with the bias. Finally, we show experiments that confirm the theoretical analysis.

1. Introduction

At the apogee of visual recovery are shape models of the scene. Cues such as motion, texture, shading, and contours encode information about the scene surfaces. By inverting the image formation process (optical and geometrical) it is possible to recover three dimensional information about the scene in view. However, despite tremendous progress there are still many difficulties. For example, when structure is reconstructed from multiple views (e.g. from motion or stereo), and even when the 3D viewing geometry is estimated correctly, the shape often is incorrect. Similarly, the processes of shape from texture, even when the underlying assumptions hold, are known to give distorted shape reconstruction.

Why? Is there some fundamental reason that this happens, or is it due to the inadequacy and lack of sophistication of our computational models? The literature in psychophysics reports that humans also experience difficulties in computing 3D shape and this has been demonstrated by many experiments. For a variety of conditions and from a number of cues the mis-estimation is an underestimation of slant. For example, planar surface patches estimated from texture [1, 6], contour [14], stereopsis [7], and motion of various parameters [16] have been found to be estimated with smaller slant, that is, closer in orientation to a front-parallel plane than they actual are.

In this paper we investigate the problem of 3D shape estimation using the cues of motion, stereo and texture. We show that there exist inherent computational limitations. These result from the well known statistical dilemma. Shape estimation processes are estimation processes. But because there is noise in the image data, and because the complexity of the visual computations does not allow to accurately estimate the noise parameters, there is bias in the estimation. Thus, we find, that one of the reasons for inaccuracy in shape estimation, is systematic error, i.e. bias.

The paper accomplishes three things. (a) We introduce a new constraint for shape from multiple views (motion and stereo) which relates shape and rotation to image lines. This constraint makes it possible to: (b) provide a statistical analysis of shape from motion, which reveals an underestimation of slant as experienced by humans. The same applies to shape from texture constraints. An understanding of the bias allows us to create displays that give rise to illusory erroneous depth perception. Since we understand the parameters involved in the bias we can set them such that the bias is very large causing mis-perception. **(c)** We discuss and implement the statistical procedures which are best for shape estimation. We found that the best we can do with these techniques is slightly reduce the bias. We then suggest that we may do better in structure from motion by employing in addition to the local shape computations a new global constraint on the shape.

2 Overview and the main concepts

The idea underlying the statistical analysis is simple. The constraints in the recovery of shape can be formulated as linear equations in the unknown parameters. Thus the problem is reduced to finding the "best" solution to an over-determined equation system of the form $A'u' = b'$ where $A' \in R^{N\times K}$ and $b' \in R^{N\times 1}$ and $N \geq K$. The observations A' and b' are always corrupted by the errors, and in addition there is system error. We are dealing with what is called the errors-in-variable (EIV) model in statistical regression, which is defined as:

Definition 1 *(Errors-In-Variable Model)*

$$\begin{aligned} b &= Au + \epsilon \\ b' &= b + \delta_b \\ A' &= A + \delta_A \end{aligned}$$

u are the true but unknown parameters. A' and b' are observations of the true but unknown values A and b. δ_A, δ_b are the measurement errors and ϵ is the system error which exists if A and b are not perfect related.

The most common choice to solving the system is by means of LS (least squares) estimation. However, it is well known, that the LS estimator u_l, whose solution is characterized by $u_l = (A'^T A')^{-1} A'^T b'$, generally is biased [18].

Consider the simple case where all elements in δ_A and δ_b are i.i.d random variables with zero mean and variance σ^2. Then

$$\lim_{n\to\infty} E(u_l - u) = -\sigma^2 (\lim_{n\to\infty} (\frac{1}{n} A^T A))^{-1} u, \tag{1}$$

which implies that u_l is asymptotically biased. Large variance in δ_A, ill-conditioned A or an u which is oriented close to the eigenvector of the smallest singular value of A all could increase the bias and push the LS solution u_l away from the real solution. Generally it leads to an underestimation of the parameters.

Using the bias from least squares we analyze in Sections 3 the estimation of shape from motion and extend the analysis in Section 4 to stereo. Section 5 discusses shape estimation from texture.

The skeptical reader may argue that the bias is only an artifact of the particular estimator, that is least squares. Aren't there other estimators we could use to eliminate the bias? Our answer to this comes in 6. In a nutshell, although there are estimators with provably no bias (under certain conditions), in order to apply them in a way that they correct the bias, we would need to have information about the noise parameters, which in general we cannot obtain. Section 6 describes properties of the best estimators with respect to our problem. Section 7 shows experiments, and Section 8 summarizes the study and discusses possible avenues to deal with the bias.

Many previous studies analysed the statistics of visual processes. In particular, bias was shown for 2D feature estimation and optical flow [13, 19]. In [9] bias was discussed for a number of visual recovery processes, and many studies analysed the statistics of structure from motion [2]. However, these analyses stayed at the general level of parameter estimation; no one has shown before the effects on the estimated shape.

Shape from motion, or in general shape from multiple view is an active research area, and many research groups are involved in extracting 3D models on the basis of multiple view geometry [8]. The theory proceeds by first solving for camera geometry (where are the cameras?). After the cameras are placed, the structure of the scene is obtained by extending the lines from the camera centers to corresponding features; their intersections provide points in 3D space which make up a 3D model.

Thus, the structure of the scene requires both the translation and the rotation between views. But the structure can be viewed as consisting of two components: (a) the shape, i.e. the normals to the surface and (b) the (scaled) depth. It is quite simple to show that shape depends only on the rotation between two views, while depth depends also on the translation. This new constraint, which is explained next, allows us to perform an uncertainty analysisx of the estimation of shape from motion and deduce the underestimation of slant.

3 Shape from Motion

3.1 Formulation of the constraint

Consider the scene to be a textured plane with surface normal n. The texture is described by the lines on the plane.

A line L in 3d space is described by Plücker coordinates $L = (L_d, L_m)$, where

$$\begin{cases} L_d = P_1 - P_2; \\ L_m = L_d \times P = P_2 \times P_1. \end{cases}$$

for any points P, P_1, P_2 on the line. L_d denotes the direction of the line in space, and L_m its moment. The length of L_m is the distance from the origin to the line. L_d and L_m are perpendicular, that is $L_d \cdot L_m = 0$. The projection of the 3D line L on the image is just L_m and normalized to have the third coordinate 1, it is:

$$\ell = \frac{1}{\hat{z} \cdot L_m} L_m.$$

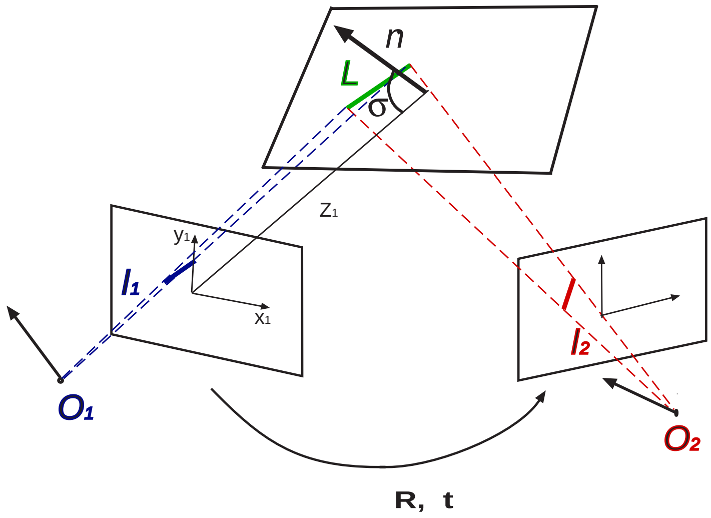


Figure 1. Two views of a planar patch containing a line.

Let us first describe the intuition: The camera undergoes a rigid motion described by a translation vector T and a rotation matrix R, as shown in Fig. 1. Let subscripts $_1$ and $_2$ denote quantities at time instances t_1 and t_2. Using projective coordinates l_1 and l_2 represent the normals to the planes defined by the camera centers (O_1 and O_2) and the projections of the line in space on the images. L is the intersection of these two planes. Thus the crossproduct of l_1 and l_2 is parallel to L_d. Writing this equation in the first coordinate system, we obtain

$$l_1 \times R^T l_2 = k L_d \tag{2}$$

with k a scalar. (Note: We can find using rotation only a line parallel to L, but not the line L. To find L we need translation.) The immediate consequence is that from two corresponding lines in two views we can find the shape of the patch containing the lines. Since L_d is perpendicular to the surface normal, n, we obtain

$$(l_1 \times R^T l_2) \cdot n = 0.$$

We model here differential motion, that is a point in space has velocity $\dot{P} = t + \omega \times P$, in which case we have that

$$\begin{cases} \dot{L}_d = \dot{P}_1 - \dot{P}_2 = \omega \times (P_1 - P_2) = \omega \times L_d \\ \dot{L}_m = \dot{P}_2 \times P_1 + P_2 \times \dot{P}_1 = t \times L_d + \omega \times L_m \end{cases}$$

Hence

$$\dot{\ell} = \frac{\dot{L}_m}{\hat{z}L_m} - \frac{\hat{z}\dot{L}_m}{\hat{z}L_m}\frac{L_m}{\hat{z}L_m} = \frac{1}{\hat{z}L_m} t \times L_d + \omega \times \ell + \frac{\hat{z}\dot{L}_m}{\hat{z}L_m}\ell,$$

and the constraint in (2) takes the form

$$\ell \times (\dot{\ell} - \omega \times \ell) = \frac{t \cdot \ell}{\hat{z}L_m} L_d. \tag{3}$$

Thus, if the 3D line is on the plane with normal vector n, its image ℓ must obey the following constraint

$$\begin{aligned} n \cdot (\ell \times (\dot{\ell} - \omega \times \ell)) &= 0 \quad \text{or} && (4) \\ n \cdot e &= 0 && (5) \end{aligned}$$

with $e = (\ell \times (\dot{\ell} - \omega \times \ell))$

3.2 Error analysis

Let $n = (n_1, n_2, 1)$ be the surface normal, and let $\{\ell_i = (a_i, b_i, 1)\}$ denote the lines on the plane, and $\{\dot{\ell}_i = (\dot{a}_i, \dot{b}_i, 0)\}$ denote the motion parameters of the lines ℓ_i. We estimate the orientation of the plane using LS estimation.

From (5) we know, that n in the ideal case should satisfy equation,

$$(e_{1_i}, e_{2_i}) \cdot \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = -e_{3_i}, \tag{6}$$

where

$$\begin{cases} e_{1_i} = -\dot{b}_i + (-(1+b_i^2)\omega_1 + a_i b_i \omega_2 + a_i \omega_3) \\ e_{2_i} = \dot{a}_i + (a_i b_i \omega_1 - (1+a_i^2)\omega_2 + b_i \omega_3) \\ e_{3_i} = (\dot{a}_i b_i - \dot{b}_i a_i) + (a_i \omega_1 + b_i \omega_2 - (a_i^2 + b_i^2)\omega_3). \end{cases}$$

There is noise in the measurements of the line locations and the measurements of line motion. For simplicity of notation, let us ignore here the error in the estimates of the rotation parameters. Throughout the paper let primed letters denote estimates, unprimed letters denote real values, and δ's denote the errors. That is, $\delta\dot{a}_i = \dot{a}'_i - \dot{a}_i$ and $\delta\dot{b}_i = \dot{b}'_i - \dot{b}_i$ with expected value 0, variance δ_1^2; $\delta a_i = a'_i - a_i$ and $\delta b_i = b'_i - b_i$ with expected value 0 and variance δ_2^2. Then we have

$$(e_1 + \delta e_1)n'_1 + (e_2 + \delta e_2)n'_2 = -(e_3 + \delta e_3).$$

Let E and δE denote the $N \times 2$ matrices and G denote the $N \times 1$ matrix as follows,

$$\begin{aligned} E &= (e_{1_i}, e_{2_i})_n, & \delta E &= (\delta e_{1_i}, \delta e_{2_i})_n, \\ G &= (-e_{3_i})_n, & \delta G &= (-\delta e_{3_i})_n. \end{aligned}$$

Then the estimation $u' = (n'_1, n'_2)$ is obtained by solving the equation,

$$(E + \delta E)^T (E + \delta E) u' = (E + \delta E)^T (G + \delta G).$$

Let M denote $E^T E$. Assuming that the errors are much smaller than the real values, we develop the LS solution of u' in a Taylor expansion and obtain as an approximation for the estimate:

$$\begin{aligned} u' &= u - \sum_i \sum_{t_i \in V} \delta t_i^2 \Big(M^{-1} \begin{pmatrix} \frac{\partial^2 e_{1_i}^2}{\partial t_i^2} & \frac{\partial e_{1_i} e_{2_i}}{\partial t_i^2} \\ \frac{\partial e_{1_i} e_{2_i}}{\partial t_i^2} & \frac{\partial^2 e_{2_i}^2}{\partial t_i^2} \end{pmatrix} u \\ &\quad + M^{-1} \begin{pmatrix} \frac{\partial^2 e_{1_i} e_{3_i}}{\partial t_i^2} \\ \frac{\partial^2 e_{2_i} e_{3_i}}{\partial t_i^2} \end{pmatrix} \Big), \end{aligned}$$

where V is the set of all variables $\{a_i, b_i, \dot{a}_i, \dot{b}_i\}$.

For the simplicity of expression, we consider a_i and b_i to be independent random variables which are symmetric to the center of the image coordinate system; in other words, $E(a_i^k) = E(b_i^k) = 0, k = 1, 2$. Then with enough equations, the expected value for the LS solution u' is well approximated by

$$E(u') = u - M^{-1}(\delta_1^2 D + \delta_2^2 F)u - M^{-1}\delta_2^2 H, \tag{7}$$

where

$$\begin{aligned} D &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix}; \quad H = \omega_3 \sum_i^N \begin{pmatrix} \omega_1 b_i^2 & 0 \\ 0 & a_i^2 \omega_2 \end{pmatrix}, \\ F &= \sum_i^N \begin{pmatrix} 4b_i^2\omega_1^2 + c_i\omega_2^2 + \omega_3^2 & c_i\omega_1\omega_2 \\ c_i\omega_1\omega_2 & c_i\omega_1^2 + 4a_i^2\omega_2^2 + \omega_3^2 \end{pmatrix} \end{aligned}$$

where $c_i = a_i^2 + b_i^2$

3.3 The effects on slant

The slant σ is the angle between the surface normal and the negative Z-axis (0^o slant corresponds to a plane parallel to the image plane, 90^o slant corresponds to a plane that contains the optical axis) and the tilt τ is the angle between the direction of the projection of the surface normal onto the XY-plane and the X-axis. Using these coordinates $\frac{n}{\|n\|} = (\cos\tau\sin\sigma, \sin\tau\sin\sigma, \cos\sigma)$.

For the case when rotation around the Z-axis can be ignored (i.e, $\omega_3 = 0$) equation (7) simplifies to

$$E(u') = (I - \delta_A)u = (I - M^{-1}(\delta_1^2 D + \delta_2^2 F))u.$$

Since D and F are positive defined matrices, so is δ_A. And usually the perturbation δs are small. Then the eigenvalues of $(I - \delta_A)$ are between zero and one, which leads to the Rayleigh quotient inequality:

$$\frac{E(u')^T E(u')}{u^T u} \leq \|I - \delta_A\| < 1.$$

Since $\sigma = \cos^{-1}(1 + u^T u)$ is a strictly increasing function, by linear approximation, we have

$$E(\sigma') < \sigma,$$

which shows that slant is underestimated. The degree of underestimation highly depends on the structure of matrix M; the inverse of M is involved in equation (7). Thus, the smaller the determinant of matrix M, the larger the bias in the estimation. The velocity of rotation also contributes to the magnitude of the bias as can be seen from matrix F; larger velocity more bias.

We can say more about the dependence of slant estimation on the texture distribution. Recall from equation (3) that

$$e = \frac{(t \cdot \ell)}{\hat{z} L_m} L_d,$$

Let us consider a slanted plane whose texture only has two major directional components. Let the directional components be $L_{d_1} = (\cos\tau_1 \sin\sigma_1, \sin\tau_1 \sin\sigma_1, \cos\sigma_1)$ and $L_{d_2} = (\cos\tau_2 \sin\sigma_2, \sin\tau_2 \sin\sigma_2, \cos\sigma_2)$. Then we have

$$\begin{aligned} M &= E^T E = \begin{pmatrix} \sum e_{1_i}^2 & \sum e_{1_i} e_{2_i} \\ \sum e_{1_i} e_{2_i} & \sum e_{2_i}^2 \end{pmatrix} \\ &= \sum(\frac{T \cdot \ell_i}{\hat{z} L_{m_i}})^2 \sin^2\sigma_1 \begin{pmatrix} \cos^2\tau_1 & \sin\tau_1\cos\tau_1 \\ \sin\tau_1\cos\tau_1 & \sin^2\tau_1 \end{pmatrix} \\ &+ \sum(\frac{T \cdot \ell'_i}{\hat{z} L'_{m_i}})^2 \sin^2\sigma_2 \begin{pmatrix} \cos^2\tau_2 & \sin\tau_2\cos\tau_2 \\ \sin\tau_2\cos\tau_2 & \sin^2\tau_2 \end{pmatrix} \end{aligned}$$

and the determinant $det(M)$ of M amounts to

$$\begin{aligned} det(M) &= [(\frac{1}{N}\sum(\frac{t \cdot \ell_{1_i}}{\hat{z} \cdot L_{M_{1_i}}})^2)^{\frac{1}{2}}(\frac{1}{N}\sum(\frac{t \cdot \ell_{2_i}}{\hat{z} \cdot L_{M_{2_i}}})^2)^{\frac{1}{2}} \\ &\quad \sin\sigma_1 \sin\sigma_2 \sin(\tau_1 - \tau_2)]^2. \end{aligned}$$

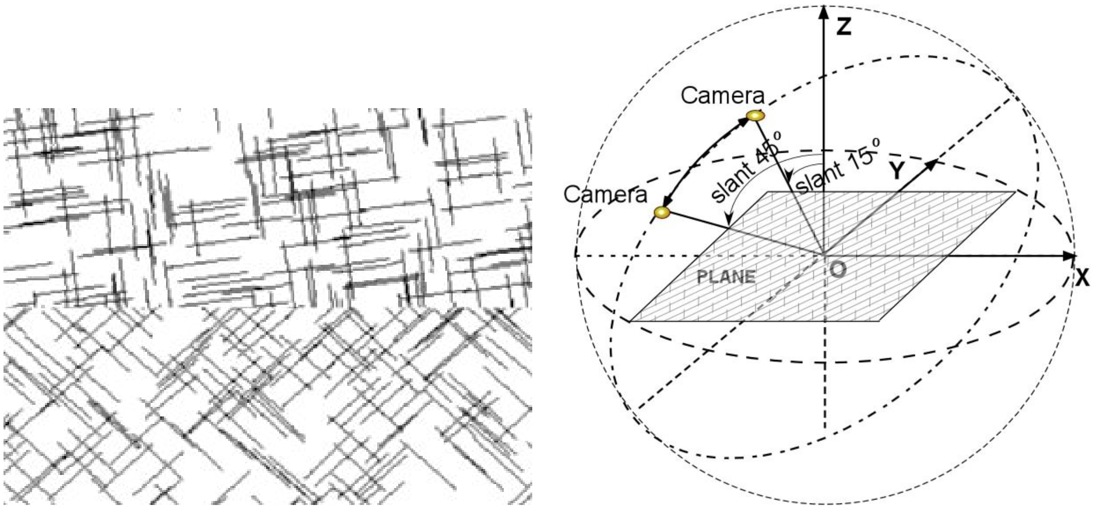
The smaller $det(M)$, the larger the underestimation. Using our model we can predict the findings from experiments in the psychological literature ([16]). For example, it has been observed in [16], that an increase in the slant of a rotating surface causes increased underestimation of the slant. By our formula, it is easy to see that $det(M)$ has a factor $\sin(\sigma_1)\sin(\sigma_2)$, where σ_1 and σ_2 are the the angles between the directions of the line in space and the negative Z-axis. Unless, they are 0 degree, these values decrease with an increase of the slant of the plane, and this leads to a smaller $det(M)$. Hence, we get a larger error towards underestimation of the slant.

To demonstrate the predictive power of the model we created two illusory displays. In the first one, the scene consists of a plane with two textures, one in the upper half, the other in the lower half. Figure 2a shows the plane when it is parallel to the screen. The texture in the upper part consists of two line clusters with slope 8^o and 98^o. The lower part has two lines clusters with slope 45^o and 135^o. A video was created for the camera orbiting the sphere along a great circle in the YZ plane as shown in Figure 2b – that is the camera translates and rotates such that it keeps fixating at the center. At the beginning of the motion, the slant of the plane with respect to the camera is 15^o, at the end it is 45^o. The image sequence can be seen at [4] As can be experienced, it creates the perception of the plane to be segmented into two parts, with the upper part having a much smaller slant.

This is predicted by the biases in the different textures. For the upper texture the bias is much larger, thus producing larger underestimation of the slant, and the underestimation gets worse as the slant increases. The ratio of the determinants of the upper and lower texture is a good measure. For the given scene it takes values between 0.08 (for 15^o slant) and 0.25. (for 45^o slant). In a second display the plane is divided into multiple segments with two alternating textures. In every other segment there is large bias, and this gives rise to the perception of the plane folding as if it were a staircase.

4 Shape from stereo

Here we adopt the symmetric stereo setting. That is, the coordinate system is in between the two cameras whose rotations with respect to this coordinate system are described by the rotation matrices R



(a) (b)

Figure 2. (a) The plane in view (b) Scene geometry in the shape from motion demonstration.

and R^T. We obtain the the linear system

$$kL_d = (R\ell_1) \times (R^T \ell_2)$$

The transformation between the two views is a translation in the XZ plane and a rotation around the Y-axis with angle $2t$. By the same notion as in the previous section, we have as the prime equation for $n = (n_1, n_2, 1)$: $(e_{1_i}, e_{2_i}) \cdot \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = -e_{3_i}$, where

$$\left\{ \begin{array}{lcl} e_{1_i} & = & b_{1_i}(a_{2_i} \sin t + \cos t) - (-a_{2_i} \sin t + \cos t) b_{1_i} \\ e_{2_i} & = & -(a_{1_i} \sin t + \cos t)(a_{2_i} \cos t - \sin t) \\ & & -(a_{1_i} \cos t + \sin t)(a_{2_i} \sin t + \cos t) \\ e_{3_i} & = & -(a_{1_i} \cos t + \sin t) b_{2_i} + (a_{2_i} \cos t - \sin t) b_{1_i} \end{array} \right.$$

The measurement errors in the line locations are assumed to be i.i.d. with zero-mean and covariance δ_2. Let $E = (e_{1_i}, e_{2_i})_n$. Under the small baseline assumption and some alignment of the image, we obtain as approximation for the expected value of the LS solution:

$$E(u') = u - nM^{-1}Fu,$$

where $M = E^t E$ and asymptotically

$$F = \delta_2 \begin{pmatrix} G(a_{1_i}^2 + a_{2_i}^2 + b_{1_i}^2 + b_{2_i}^2, t) & 0 \\ 0 & G(a_{1_i}^2 + a_{2_i}^2, 2t) \end{pmatrix},$$

where $G(x,t) = E(x) \sin^2 t + \cos^2 t$. By the same arguments as in the case of shape from motion , the slant is underestimated.

5 Shape from texture

There are three effects on the distortion of image texture: the perspective scaling, the foreshortening and the position effect [6].

We illustrate here the underestimation of the slant for the foreshortening cue, which assumes isotropic texture, i.e that there is no directional bias in the distribution of the texels on the surface. Then the directional statistics of image texels can be related to the the orientation of the surface.

Let the directional statistics of the edges map the angles, α, of the edges to the unit circle by the mapping $F : \alpha \longrightarrow (\cos 2\alpha, \sin 2\alpha)$. is the angle of the edges. Then the center of mass (C, S), defined by

$$C = \int_0^\pi \cos 2\alpha dF(\alpha), \quad S = \int_0^\pi \sin 2\alpha dF(\alpha),$$

provides information about the slant and the tilt of the surface plane. We have that the slant σ increases monotonically as a function of the quantity $Q = \sqrt{C^2 + S^2}$. Given the measurements of the edge angles α_i', Q' is computed as

$$Q' = \frac{1}{n}\sqrt{(\sum_{i=1}^{n} \cos 2\alpha_i')^2 + (\sum_{i=1}^{n} \sin 2\alpha_i')^2}.$$

There are two types of errors, the measurement error in the angle α_i' and the modeling error due to the fact that the real edge distribution is not exactly isotropic. Now consider that the errors in the angle measurements of the edges are i.i.d. random variables with expected value 0 and variance δ^2. Let Q denote the actual value of the quantity Q'. It can be shown that the expected value of Q' amounts to

$$E(Q') = (1 - 4\delta^2)Q + O(\frac{1}{n}).$$

Hence asymptotically we have $E(Q') < Q$, and thus the expected value of the slant $E(\sigma') < \sigma$, in other words, the slant is underestimated. Similar results can be obtained for the modeling error. Let $\tilde{\alpha}_j$ denote the angles of the edgels from another distribution and $\tilde{Q}$ denote the norm mass center. Then we can show that

$$E(Q') \leq \frac{\sqrt{n^2Q^2 + m^2\tilde{Q}^2}}{m + n}.$$

Since Q is the major pattern, we expect that m is much smaller than n, and so we have, unless the slant is close to 90^o, that $E(Q') < Q$. Thus, in general, this error also leads to an underestimation of slant. A similar conclusion can be obtained as well for the position cue and the scaling cue.

6 Statistical Alternatives

The statistical model that describes the data in visual estimation processes is the errors-in-variable model (Definition 1). The main problem with Least squares (LS) estimation is that it does not consider errors in the explanatory variables, that is δ_A. The obvious question thus arises: Are there better alternatives that reduce the bias? Clearly, bias is not the only thing that matters. There is a trade-off between bias and variance. Generally an estimator correcting for bias increases the variance while decreasing the bias.

Next we discuss well known approaches from the literature and what the difficulties are in applying them to the vision problem. We conclude that the most promising approach is the technique of instrumental variables.

CLS (Corrected Least Squares) estimation is the classical way to correct for the bias. If the variance of the error is known, it gives asymptotically unbiased estimation. The problem is that accurate estimation of the variance of the error is a challenging task if the sample size is small. For small amounts of data the estimation of the variance has high variance itself. Consequently this leads to higher variance for CLS.

Usually the mean squared error (MSE) is used as a criterion for the performance of an estimator. It amounts to the sum of the square of the bias plus the variance of the estimator. According to this criterion the best linear estimation (linear in b) should be a partial correction using the CLS; the smaller the variance the larger the correction.

TLS (Total Least Square) is a nonlinear estimation that recently has attracted a lot of attention [18] The basic idea is to deal with the errors in A' and b' symmetrically. The TLS estimator seeks to

$$\text{minimize}_{\hat{A},\hat{b},u} \|[A';b'] - [\hat{A},\hat{b}]\|_F^2 \quad \text{subject to} \quad \hat{b} = \hat{A}u, \tag{8}$$

where $\|\cdot\|_F$ is the Frobenius norm. If all errors δ_A, δ_b are i.i.d., then TLS estimation is asymptotically unbiased. In the case they are not, one would need to whiten the data. But this requires the estimation of the ratio of the error variances δ_A and δ_b, which is at least as hard as obtaining the variance of δ_b. An incorrect value of the ratio often results in an unacceptably large *over correction* for the bias. However, the main problem for TLS is system error. We can have multiple tests to obtain the measurement error, like re-measuring or re-sampling; but unless we know the exact parameters of the model, we can't test the system error. Hence, unless the system error is small and accurate estimation of the ratio of variances can be obtained accurately, TLS will not be unbiased, and its variance will be larger than the variance of LS. In comparison, CLS has better robustness on model assumptions than TLS and system error doesn't affect it much.

Resampling techniques, such as bootstrap and Jacknife have been discussed for bias correction. These techniques can correct for the error term which is inverse proportional to the number of data points (i.e. $O(\frac{1}{n})$), and thus they can improve the estimate of the mean for unbiased estimators. However, these techniques cannot correct for the bias [3]. They are useful for estimating the variance in order to provide confidence intervals ([12]).

Technique of instrumental variables: All other estimation techniques require knowledge of the error covariance, which is hard to obtain, because its estimation is highly unstable and requires a large sample size. An attractive alternative is the technique of instrumental variables which deals with the errors in the explanatory variables but does not require the error variance as a priori. Its definition is as follows.

We denote the K-dimensional row vector of A' by A'_i. Let W_i be a q-dimensional row vector with $q \geq K$. Let W_i be such that

i) $E(W_i^T(\delta_{A_i}, \delta_{b_i})) = (\mathbf{0}, \mathbf{0})$;

ii) The rank of $(\sum_{i=1}^N W_i^T W)^{-1} \sum_{i=1}^N W_i^T A'_i$ is full ,

then $W = (W_i)$ are called the instrumental variables of A'. For example, we could have two methods to measure A'. If the errors in the measurements of the two methods can be treated as independent, then the measurements of one method could be considered as the instrumental variables of the measurements of the other method.

Then we have an unbiased estimator of u by solving the replaced equations system

$$(W^T A')u = W^T b',$$

which by standard least square method amounts to

$$u = (A'^T W W^T A')^{-1}(A'^T W W^T b').$$

This estimator is asymptotically unbiased [5] with the variance close to the variance of the CLS estimator.

We found this method to be most successful in dealing with our problem. The technique of instrumental variables is highly robust to improper error modeling. It could be realized by using multiple edge detections, fitting schemes, and difference operators to obtain multiple measurements of the explanatory variables. We will not be able to achieve complete independence of the measurement errors. But the worst that can happen is that the different measuring methods have the exact same measurement error, in which case the method reduces to LS estimation.

6.1 Discussion of the errors in shape from motion

The measurements are the line parameters $\{a_i, b_i\}$, and the image motion parameters of the lines, $\{\dot{a}_i, \dot{b}_i\}$. We can expect four types of noise:

Sensor noise: effects the measurements of image intensity $I(x, y, t)$. It seems reasonable to approximate the sensor noise as i.i.d.. But we have to consider dependences when the images are smoothed.

Fitting error: Estimating the line parameters a_i, b_i amounts to edge detection. Clearly there are errors in this process. Longer edgels are associated with smaller errors and shorter edgels with larger errors.

Discretization Error: Derivatives are computed using difference operators, which have truncation errors associated with them. The magnitude of the error depends on the smoothness and the frequency of the texture.

System error: When computing the motion of lines, we assume that the image intensity is constant between frames. Significant errors occur at specular components. We use first order expansions when deriving velocities. Thus, errors are expected for large local velocities. Furthermore, the modeling of the scene as consisting of planar patches is an approximation to the actual surface of the scene.

Among the errors above, sensor noise has been considered in a number of papers in structure from motion ([10, 13]). Other errors have hardly been mentioned or have been simply ignored. But actually other errors could contribute much more to the error than the sensor noise. Furthermore, the sensor characteristics may stay fixed. But other noise components do not. They change with the lighting conditions, the physical properties of the scene being viewed, and the orientation of the viewer in 3D space.

Considering all the errors, the errors δA_i and δb_i are due to a number of different components and cannot be assumed to be independent and identical. This makes the estimation of the variance unreliable. Thus CLS and TLS are not useful for correcting the bias. The technique of instrumental variables still can handle this model. Our experiments showed that this method resulted in a minor improvement.

7 Experiments

We compared the different regression methods for the estimation of slant from motion using simple textured planes as in the illusory video. TLS estimation was implemented by assuming all errors to be i.i.d.. CLS was implemented by assuming the errors in e_1 and e_2 to be i.i.d.. The variance of the errors was estimated by the SVD method, that is by taking the smallest singular value of the matrix $[A;b]$ as the estimation of the variance. We implemented two types of bootstrap methods. In one method, the samples (e_{1_i}, e_{2_i}) were bootstrapped, in the other the residuals $e_{1_i}n_1 + e_{2_i}n_2 - e_{3_i}$. We also implemented a simple version of the instrumental variable method by using three differently sized Gaussian filters to obtain three samples for each image gradient.

We generated data sets of random textures with sparse line elements as in the videos. In texture set No.1 the lines have dominant directions 10^o and 100^o; in texture set No.2 the dominant directions are 45^o and 135^o. We tested for two slants, 45^o and 60^o. The motion in the sequences was only translation, thus there is nor error due to rotation. The tables below show the average estimated value of the slant for the four data sets.

Experiments with the slant 45^o

No.	LS	CLS	TLS	Jack	Boot 1	Boot 2	Inst. Var.
1	41.8675	37.5327	54.8778	44.4426	43.0787	41.5363	43.0123
2	39.8156	40.8279	42.7695	39.0638	40.4007	40.1554	41.9675

Experiments with the slant 60^o

No.	LS	CLS	TLS	Jack	Boot 1	Boot 2	Inst. Var.
1	45.7148	46.0307	46.6830	45.9929	46.2710	45.6726	49.3678
2	42.5746	44.4127	43.3031	47.1324	45.5572	42.8377	48.1202

The experiments demonstrate that LS tends to underestimate the parameters. TLS tends to give larger estimattes than LS, but sometimes it overestimates the parameters, that is, it tends to over-correct the bias. CLS corrects the bias little. The reason could be either that the estimation of the variance is not trustable, or that the assumption that the measurement errors are independent is not correct. The performance of Bootstrap and Jacknife is not much better. The bias hardly gets corrected. The instrumental variable method seems a bit better than the other methods, but it still only corrects the bias by a small amount.

8 Conclusions and Discussion

This paper analyzed the statistics of shape estimation. We showed that bias is a serious problem. We analyzed the bias for least squares estimation and we showed that it predicts the underestimation of slant, which is known from computational and psychophysical experiments.

One may question that LS estimation is a proper model for human vision. We discussed and showed experimentally that most elaborate estimators (CLS (unless one largely overestimates the variance) Bootstrap, Instrumental Variables) also have bias which is qualitatively of the same form as the one of LS. Thus these estimators, too, would lead to an underestimation of slant. TLS, depending on the ratio of variances, may give the opposite result.

Our analysis of shape from motion was possible because of a new constraint which relates shape and rotation only to image features. The reader interested in structure from motion may argue that using non-linear techniques we may estimate iteratively the parameters of the errors as well as the parameters of motion and structure. If we knew the error model and had a lot of data available, theoretically it would be possible to correct. However, we usually don't have enough data to obtain the errors, which depend on many factors. Furthermore, we don't know the exact error model.

Bundle adjustment techniques [17] have become very popular in structure from motion. These techniques actually are closely related to the idea of TLS regression. The formulation for the problem of structure from motion in the discrete case is

$$\min_{\hat{P}_k,\hat{R}_i} \sum \sum \|r_{k,i} - \hat{P}_k \hat{R}_i\|^2.$$

where P_k is the projection matrix for the k-th view, and $r_{k,i}$ are the image points corresponding to a 3D point R_i in the k-th view. Rewritting the TLS estimator (equation 8) as follows:

$$\min_{x,\hat{A}_i} \sum \|[A_i, b_i] - [\hat{A}_i, \hat{A}_i x]\|^2,$$

the resemblance is easy to see. In this case, since the assumption that the errors of r_{k_i} are i.i.d seems reasonable, and since the system error should not be large (the epipolar constraint is exact) we could expect bundle adjustment to refine the solution from LS. The problem with bundle adjustment, of course, is that the starting point needs to be fairly close to the real solution.

However, in the differential case of structure from motion, bundle adjustment cannot correct the bias. First, the assumption of i.i.d. errors is not valid. We need to estimate flow. But the errors in the spatial and the errors in the temporal derivatives of the intensity are different. Second, there is significant system error. The constraint underlying image motion – the brightness constancy constraint – is an approximation, considering only the first order derivatives, neglecting higher order terms. This introduces system error that is not significantly smaller than the measurement errors. Thus, based on the arguments on TLS estimation, bundle adjustment wouldn't successfully correct the bias and it could bring unexpected numerical instability.

The question, thus, for computational vision arises: How can we deal with the bias? Clearly, bias is not confined to shape estimation only. Other processes of visual reconstruction are estimation processes as well and thus will suffer from the same problem. Since better estimation techniques are not the answer, we have to use the data such that bias does (mostly) not effect the goal, that is what we want to do with the data. First, we should use the data selectively. Since we understand how the the different parameters influence the bias, we can choose data that is not effected much by the bias. For example in computing shape from motion we can avoid patches with textures corresponding to a badly conditioned matrix M. Second, we should use when possible, the data globally. Large amounts of data usually are not directionally biased, and thus the bias in estimation will be small. For example, when estimating 3D motion from image motion we should make use of all the data from the whole image. The same applies to shape estimation. Third, above we discussed that the statistics of structure from motion is easier for the discrete case than the continuous case. Thus, it is advantageous to estimate shape from views far apart. Of course, using far away views we run into the difficulty of finding good correspondence. The way to address structure from motion then is to use continuous motion to obtain a preliminary estimate of the 3D motion and shape, and subsequently use these estimates to obtain shape from views far apart. There is a new constraint which can help us improve the estimation. Shape only depends on rotation and thus the surface normals of all the patches in multiple views can be shown to lie in a three-dimensional linear space. (It is known that there is a rank 4 constraint on multiple planes [15, 11], but shape is even lower dimensional.)

Consider a plane with normal vector $\vec{n}$ and the camera undergoing a rigid motion with rotation R and translation T. The corresponding plane after the motion has a surface normal $\vec{n}' = R\vec{n}$. For multiple planes with normals n_i with $i = 0, 1, \cdots, N$ let us define the $3 \times N$ matrix

$$A = (n_1, n_2, n_3, \cdots, n_N).$$

We then have after the motion

$$A' = (n'_1, n'_2, n'_3, \cdots, n'_N) = R(n_1, n_2, n_3, \cdots, n_N) = RA$$

Considering M views for the same set of planes, we then have for every view j

$$A_j = (n_{1_j}, n_{2_j}, \cdots, n_{N_j} = R_j(n_1, n_2, \cdots, n_{N_j}) = R_j A$$

Thus we can combine all the surface normals in a $3M \times N$ matrix H of rank 3 , which has the following structure:

$$H = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_M \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_M \end{pmatrix} A.$$

A possible way to utilize this constraint is as follows: First, estimate the shape from multiple frames by means of a structure from motion algorithms using continuous motion. Next, track a set of image patches over the sequence – we only need to track the pataches not the points. Then choose a number of frames separated by significant baseline, and enforce the above constraint to improve the shape and rotation estimates. Lastly, compensate for rotation and shape, that is, transform the patches to a fronto-parallel view and compute the corresponding translation by means of phase correlation. Possibly, one then could refine the solution by iterating over the different steps.

References

[1] G. J. Andersen, M. L. Braunstein, and A. Saidpour. The perception of depth and slant from texture in three-dimensional scenes. *Perception*, 27:2635–2656, 1998.

[2] K. Daniilidis and H.-H. Nagel. Analytical results on error sensitivity of motion estimation from two views. *Image and Vision Computing*, 8:297–303, 1990.

[3] B. Efron and R. Tibshirani. *An introduction to the boostrap*. Chapman & Hall, 1993.

[4] C. Fermüller. http://www.optical-illusions.org, 2003.

[5] W. Fuller. *Measurement Error Models*. Wiley, New York, 1987.

[6] J. Garding. Shape from texture and contour by weak isotropy. *J. of Artificial Intelligence*, 64:243–297, 1993.

[7] R. Goutcher and P. Mamassian. A ground plane preference for stereoscopic slant. Poster, European Conference on Vision Perception, 2002.

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[9] K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier, Amsterdam, 1996.

[10] N. Lydia and S. Victor. Errors-in-variables modeling in optical flow estimation. *IEEE Trans. on Image Processing*, 10:1528–1540, 2001.

[11] L.Zelnik-Manor and M. Irani. Multi-frame alignment of planes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1151–1156, 1999.

[12] P. Meer and B. Matei. Bootstrapping errors-in-variables models. In *Vision algorithms: Theory and Practice*, volume Lecture notes in Computer Science, pages 236–252. Springer, 2000.

[13] H. Nagel. Optical flow estimation and the interaction between measurement errors at adjacent pixel positions. *International Journal of Computer Vision*, 15:271–288, 1995.

[14] J. Perrone. Slant underestimation: A general model. *Perception*, 11:641–654, 1982.

[15] A. Shashua and S. Avidan. The rank constraint in multiple (¿=3) view geometry. In *ECCV*, pages 196–206, 1996.

[16] J. Todd and V. J. Perotti. The visual perception of surface orientation from optical motion. *Perception & Psychophysics*, 61:1577–1589, 1999.

[17] B. Triggs, P.McLauchlan, R. Hartley, and A.Fitzgibbon. *Vision Algorithms: Theory and Practice*, chapter Bundle adjustment - a modern synthesis. Springer Verlag, 2000.

[18] S. van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, 1991.

[19] Y. Weiss and E. H. Adelson. Slow and smooth, a Bayesian theory for the combination of local motion signals in human vision. AI Memo 1616, MIT, 1998.