International Journal of Computer Vision 55(1), 27–53, 2003
© 2003 Kluwer Academic Publishers. Manufactured in The Netherlands.

Stochastic Approximation and Rate-Distortion Analysis for Robust Structure and Motion Estimation

AMIT K. ROY CHOWDHURY* AND R. CHELLAPPA
Center for Automation Research and Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742
amitrc@cfar.umd.edu
rama@cfar.umd.edu

Received June 21, 2001; Revised November 19, 2002; Accepted March 26, 2003

Abstract. Recent research on structure and motion recovery has focused on issues related to sensitivity and robustness of existing techniques. One possible reason is that in practical applications, the underlying assumptions made by existing algorithms are often violated. In this paper, we propose a framework for 3D reconstruction from short monocular video sequences taking into account the statistical errors in reconstruction algorithms. Detailed error analysis is especially important for this problem because the motion between pairs of frames is small and slight perturbations in its estimates can lead to large errors in 3D reconstruction. We focus on the following issues: physical sources of errors, their experimental and theoretical analysis, robust estimation techniques and measures for characterizing the quality of the final reconstruction. We derive a precise relationship between the error in the reconstruction and the error in the image correspondences. The error analysis is used to design a robust, recursive multi-frame fusion algorithm using "stochastic approximation" as the framework since it is capable of dealing with incomplete information about errors in observations. Rate-distortion analysis is proposed for evaluating the quality of the final reconstruction as a function of the number of frames and the error in the image correspondences. Finally, to demonstrate the effectiveness of the algorithm, examples of depth reconstruction are shown for different video sequences.

Keywords: structure and motion estimation, error analysis, Robbins-Monro stochastic approximation, rate distortion theory

1. Introduction

Extraction of the 3D structure of a scene from a sequence of images, termed structure from motion (SfM), has been the central problem in computer vision for the past two decades. Extensive literature on the subject can be found in Faugeras (1993), Hartley and Zisserman (2000), and Oliensis (2000), among others. While there is no doubt that immense progress has been made in the understanding of the problem, especially its geometrical aspects, many of the available algorithms perform poorly in real-life applications. This has motivated recent research on issues of sensitivity, robustness and error characterization of existing techniques (Faugeras, 1993; Kanatani, 1996; Zhang, 1998; Ma et al., 2000; Sun et al., 2001), etc.

The errors which affect the quality of SfM algorithms can be broadly classified into two groups—geometrical and statistical. The geometrical errors arise because of the well-known ambiguities (e.g. the scale ambiguity) present in the mathematical description of the problem (see Zhang and Faugeras, 1992 or Hartley and Zisserman, 2000). They can usually be handled by imposing additional constraints on the solution space. The statistical errors are a result of the poor quality of

*Partially supported by NSF ITR grant # 0086075 and DARPA/ONR grant N00014-00-1-0908.

the video sequence. They are an inherent part of the input data and need to be compensated for if the final output solution is to be robust enough for engineering applications.

SfM algorithms often make assumptions about the inputs (e.g. perfect image correspondences) that are violated in practice and lead to errors in the reconstruction. An understanding of the strengths and shortcomings of some of the existing algorithms can be found in Triggs et al. (2000). In order to make our algorithms work in the presence of these errors, we might be tempted to introduce preprocessing stages to minimize their effects (e.g. design better correspondence algorithms). However, the sources of the errors are often unknown; preprocessing stages are independent research problems in their own right (the correspondence problem is a very good example of this); and incorporating these stages adds to the total computational cost of the final system. The alternative is to understand these errors in a statistical sense and account for their influence within the structure of the main algorithm. This paper aims to achieve that goal.

1.1. Related Work

Pioneered by the seminal work of Longuet-Higgins (1981) and the eight-point algorithm developed independently by Tsai and Huang (1981), SfM has been one of the most vibrant research areas in computer vision. Most of the earlier work concentrated on developing efficient algorithms for reconstructing 3D structure from multiple frames. The use of multiple frames was motivated by the hope that the extra information will help to minimize the errors that are inevitably present in two-frame reconstructions. The problem of tracking an object across multiple frames was addressed in Gennery (1992) where a known object and its past position and velocity were used to predict its new location. Broida and Chellappa (1991) investigated the use of the extended Kalman filter for estimating motion and structure from a sequence of monocular images. Azarbayejani and Pentland (1995) extended their work to include the estimation of the focal length of the camera, along with motion and structure. Tomasi and Kanade (1992) developed an algorithm for shape and motion estimation under orthographic projection using the factorization theorem. Szeliski and Kang (1994) proposed a non-linear least squares optimization using the Levinburg-Marquardt method. Oliensis (1999) developed a multi-frame algorithm under perspective projection, which was extended recently in Oliensis and Genc (2001). Most of these multi-frame methods can be characterized as batch processing (but not necessarily sequential or progressive) which means that the problem of estimating motion and structure is formulated as one of minimizing an objective function defined as a sum of squares of the differences between the actual observed images and the projections of their estimated 3D locations, over all tracked positions and images (bundle adjustment). In contrast, Thomas and Oliensis (1999) proposed a fusion algorithm that computes the final reconstruction from intermediate reconstructions by analyzing the uncertainties in them, rather than from image data directly. Estimation of 3D motion from an overlapping image sequence was done in Weng et al. (1987) based on two-view motion analysis from either monocular or binocular image pairs.

Many researchers have analyzed the sensitivity and robustness of several existing algorithms for reconstructing a scene from its video sequence. The work of Weng et al. (1989, 1993) 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 Broida and Chellappa (1989). Young and Chellappa (1992) derived bounds on the estimation error for structure and motion parameters from two images under perspective projection using optical flow, as well as from a sequence of stereo images (Young and Chellappa, 1990). Similar results were derived in Daniilidis and Nagel (1993) and the coupling of the translation and rotation for a small field of view was studied. Daniilidis and Nagel (1990) have also shown that many algorithms for three-dimensional motion estimation, which work by minimizing an objective function leading to an eigenvector solution, suffer from instabilities. Zhang's work (Zhang, 1998) on determining the uncertainty in the estimation of the fundamental matrix is another important contribution in this area. Haralick (1996) showed how well-known estimation techniques could be used to propagate additive random perturbations through different vision algorithms. Soatto and Brockett (1998) have analyzed SfM in order to obtain provably convergent and optimal algorithms. Oliensis (2000) has emphasized the need to understand algorithm behavior and the characteristics of the natural phenomenon that is being modeled. Ma et al. (2000) also addressed the issues of sensitivity and robustness

in their motion recovery algorithm. Recently, Sun et al. (2001) have proposed an error characterization of the factorization method for 3-D shape and motion recovery from image sequences using matrix perturbation theory. Morris et al. (2000) extended the covariance-based uncertainty calculations to account for geometric indeterminacies, referred to in the literature as *gauge* freedom.

A different source of error is the bias in depth estimation. Some authors, notably (Daniilidis and Spetsakis, 1993; Kanatani, 1993), have proved that there exists a bias in the translation and rotation estimates from stereo. Recently, it has been proposed that the bias in the optical flow field can be a possible explanation for many geometrical optical illusions (Fermuller and Aloimonos, 2001). In a separate work, we have shown that the 3D reconstruction from monocular video is statistically biased and the bias is numerically significant (Roy Chowdhury and Chellappa, 2003b). The error analysis presented in this paper assumes an unbiased estimate. In Roy Chowdhury (2002), we have shown how the results presented here can be combined with the results on the bias of the estimate to obtain a generalized Crmaer-Rao lower bound for the minimum variance of an SfM estimate, thus extending the results in Young and Chellappa (1992).

1.2. Overview of Paper

In this paper we deal with the problem of 3D reconstruction from short monocular video streams. This is important in a number of applications, e.g. surveillance, where all that may be available is a short video sequence of a person's face from one view and we may want to recognize him/her from a slightly different view. Detailed error analysis is especially important for this problem because the motion between pairs of frames is small and slight perturbations in its estimates can lead to large errors in 3D reconstruction. We use the two-frame algorithm described in Srinivasan (2000), based on the optical flow equations of SfM. Our algorithm is not specific to this method; however, the algorithm described in this paper is computationally more efficient than most others and we wish to build on it to develop a fast, reliable multi-frame algorithm.

Also, this paper will concentrate on *fusion algorithms* for 3D reconstruction. By fusion, we mean a multi-frame SfM (MFSfM) algorithm that computes the final reconstruction from intermediate reconstructions (by analyzing the uncertainties in them) rather than from image data directly. An alternative approach, "integration over time," relies on updating a previous structure estimate with information contained in the new image, weighted by their respective uncertainties. However, this method is potentially unstable if the initial structure estimates are inaccurate. To their discredit, however, fusion strategies usually fail if the intermediate reconstructions are of poor quality.

We start with an analysis of the statistics of the error in two-frame depth reconstruction (Section 2). An expression relating the error covariance in the image correspondences to the error covariance in the shape and motion reconstruction is derived. The expression does not require the standard assumptions of Gaussianity of the observations and is thus a generalization of the results presented in Young and Chellappa (1992).[1] An experimental study of the properties of two-frame reconstructions allows us to choose an appropriate optimization function and a robust solution framework using *stochastic approximation* theory, which can give optimal estimates even if information on the noise statistics is incomplete (Saridis, 1974; Ljung and Soderstrom, 1987; Benveniste et al., 1987; Spall, 2000). Based on this analysis, we develop our algorithm (Section 3), which consists of two parts, a depth fusion unit using Robbins-Monro stochastic approximation (RMSA) (Robbins and Monro, 1951) and a camera motion tracking algorithm using a Kalman filter. Finally, we evaluate the quality of the multi-frame reconstruction using the rate-distortion criterion from information theory (Cover and Thomas, 1991) (Section 4). This analysis allows us to precisely understand the sensitivity of the final structure estimate to the number of frames and to errors in the input information. Our results are demonstrated using video sequences captured with an ordinary video camera (Section 5) with application to 3D face modeling. We finally conclude in Section 6 after discussing potential applications and future research directions.

2. Statistical Analysis of Two-Frame Reconstruction

We begin our study of the development of a robust framework for multi-frame SfM with an analysis of errors in two-frame reconstructions. We will start with an experimental analysis of two-frame reconstruction and then develop a theoretical framework for a precise relationship between the errors in the images and the errors in structure and motion estimates.

2.1. The Basic Equations of SfM

Given two images, I_1 and I_2, we are interested in computing the camera motion and structure of the scene from which these images were derived. 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 (under the infinitesimal motion assumption) by

$$\begin{aligned} p(x, y) &= (-v_x + x v_z) g(x, y) + xy\omega_x \\ &\quad - (1 + x^2)\,\omega_y + y\omega_z \\ q(x, y) &= (-v_y + y v_z) g(x, y) + (1 + y^2)\,\omega_x \\ &\quad - xy\omega_y - x\omega_z, \end{aligned} \tag{1}$$

where $\mathbf{V} = [v_x, v_y, v_z]$ and $\mathbf{\Omega} = [\omega_x, \omega_y, \omega_z]$ are the translational and rotational motion vectors respectively, $g(x, y) = 1/Z(x, y)$ is the inverse scene depth, and all linear dimensions are normalized in terms of the focal length f of the camera (Nalwa, 1993). The problem is to estimate $\mathbf{V}$, $\mathbf{\Omega}$ and Z given (p, q). (1) can be rewritten in a more useful form (because of the scale ambiguity (Nalwa, 1993)) as

$$\begin{aligned} p(x, y) &= (x - x_f) h(x, y) + xy\omega_x \\ &\quad - (1 + x^2)\,\omega_y + y\omega_z \\ q(x, y) &= (y - y_f) h(x, y) + (1 + y^2)\,\omega_x \\ &\quad - xy\omega_y - x\omega_z, \end{aligned} \tag{2}$$

where $(x_f, y_f) = (\frac{v_x}{v_z}, \frac{v_y}{v_z})$ is known as the *focus of expansion* (FOE) and $h(x, y) = \frac{v_z}{Z(x,y)}$. Our problem is to obtain an accurate fused estimate of the structure from multiple frames given the two-frame solution of (2). While the scene depth Z is fixed across the frames, the camera motion $(\mathbf{V}, \mathbf{\Omega})$ changes. Thus the structure is all that can be fused since it is the only thing that remains fixed from image to image. However, we need to track the camera motion across the frames in order to align the models obtained from each pair of frames.

2.2. Qualitative Analysis

Our experiments in understanding the properties of two-frame reconstructions have two parts: analyzing the statistical distribution of the intermediate depth reconstructions (we also refer to them as sub-estimates) and time-series analysis of these sub-estimates. Our experiments are conducted on two image sequences: the face sequence and the house sequence, one frame of each of which is shown in Fig. 1 (they represent indoor and outdoor video sequences respectively).

(a) (b)

Figure 1. One frame from each of the two video sequences: (a) represents an image from an indoor video sequence and (b) from an outdoor video sequence. These two sequences were used in the qualitative analysis.

2.2.1. Distribution of Depth Sub-Estimates. To obtain a good fused estimate from sub-estimates, one should know how to weigh the sub-estimates and their uncertainties in order that the final estimate accurately reflects this information. In a general fusion problem, this involves computing the likelihood function (Poor, 1988). Traditional fusion methods like Kalman filtering provide a computational method for this likelihood function and work well under Gaussian approximation. This typically happens when the sub-estimates have small uncertainties because of which a Gaussian approximation to reflect their variances is adequate. The fact that the noise in the estimates is not Gaussian has been mentioned by various authors (Chapter 12 of Kanatani, 1996) and is due to several reasons, e.g. the physical characteristics of the imaging system, the nonlinearities of the perspective projection model, etc. As pointed out in Zhang (1998) and which we have observed in our experiments as well, small errors due to localization can usually be modeled by the second order statistics, while the outliers (often due to false matches) do not lend themselves to be modeled easily by second-order statistics. Hence our analysis has two parts. We try to model the second order statistics and its propagation across the video sequence in order to compensate for the smaller errors. The larger errors are treated as outliers and rejected using the least median of squares estimator, which is known to be robust to outliers.

A standard test for Gaussianity of observations is to analyze their higher-order statistics. It is well known

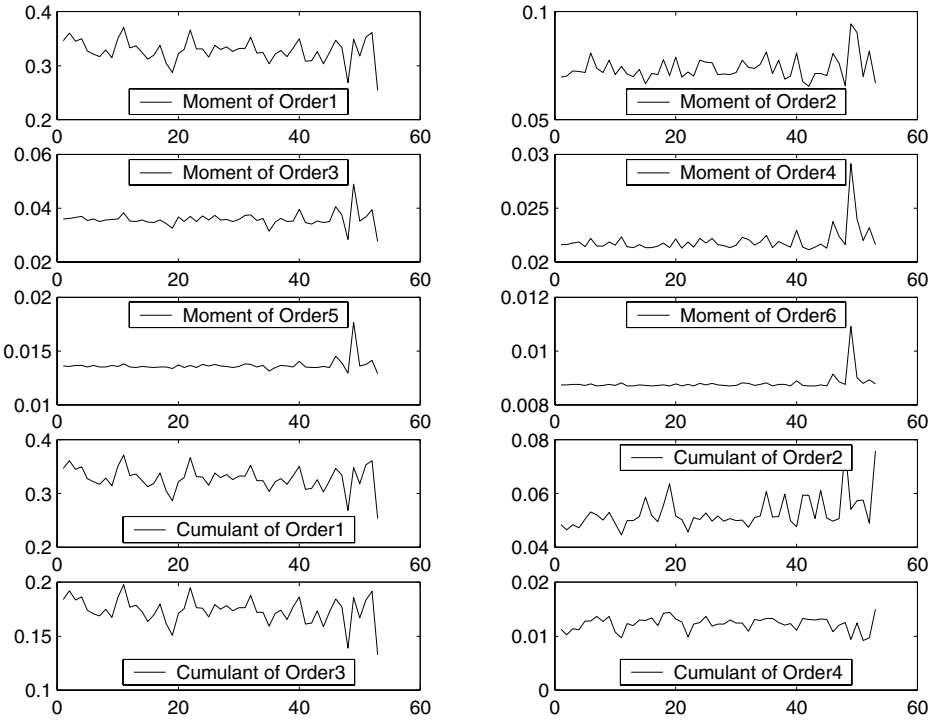

Figure 2. Plot of estimates of the moments and cumulants of the two-frame depth for the face sequence of Fig. 1(a) against the feature points. Skewness $= -0.25$; Kurtosis $= 1.9 \Rightarrow$ left skewed and flat distribution.

that for Gaussian random variables, all odd central moments are identically zero (this is actually true for any symmetric distribution) and all cumulants of order greater than two are zero (Papoulis, 1991). Figures 2 and 3 show plots of the estimates of the central moments and cumulants of two-frame depth against the feature points. Analysis of these plots reveals that there is significant non-Gaussianity in the distribution function of the depth. For the face sequence, the estimated skewness is -0.25 and the kurtosis is 1.9, while for the house sequence, the values are 1.1 and 3.2 respectively (averaged over all features). Knowing that the skewness of a standard normal distribution ($\mathcal{N}(0, 1)$) is 0 and the kurtosis is 3 (Shao, 1998), we can infer that the distribution of the depth sub-estimates for the face sequence is left skewed (negative skewness) and flat (kurtosis less than 3), while the same distribution function for the house sequence is right skewed (positive skewness) and peaked (kurtosis greater than 3). What these figures emphasize is that the distribution function of the depth sub-estimates which need to be fused is significantly non-Gaussian and it varies widely depending on the data (in fact, it is impossible to even infer whether the distribution is sub-Gaussian or super-Gaussian). However, it is not possible to infer anything more about the distribution functions, thus making it impossible to write down the likelihood function.

2.3. *Robust Estimators*

Figure 4 shows a plot of the depth values across 50 frames for four randomly chosen points in the face image sequence. It can be seen that there are isolated outliers in all four cases. It is difficult to ascertain the exact cause of the outliers; however, the general reasons for their occurrence can be inferred.[2] Application of least squares estimation techniques in the presence of such outliers will severely affect the estimates. In fact, regression analysis shows that least squares is vulnerable to outliers in both independent or *explanatory variables* as well as the observations or *response variables* (Rousseeuw and Leroy, 1987). In our case, the observations are the two-frame depth values which depend on the image correspondences (p, q), and which therefore are the explanatory variables. Thus there are outliers in both of the variables and least-squares techniques will perform poorly.

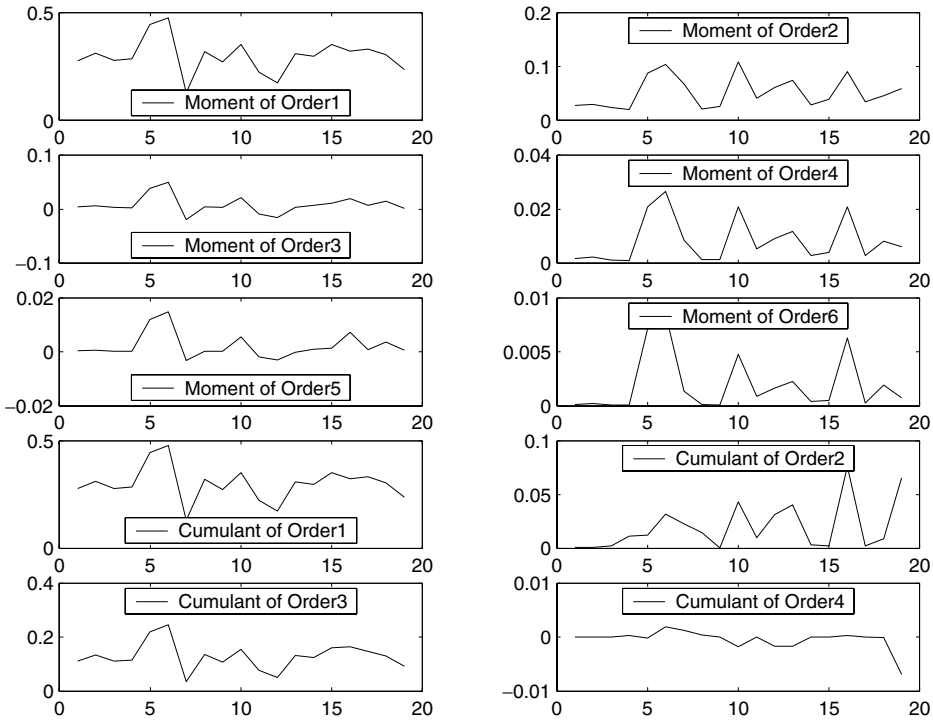

Figure 3. Plot of estimates of the moments and cumulants of the two-frame depth for the outdoor house sequence of Fig. 1(b) against the feature points. Skewness = 1.1; Kurtosis = 3.2 ⇒ right skewed and peaked distribution.

Numerous papers have been published in the statistics and signal processing literature over the last two decades on designing robust estimators (Rousseeuw and Leroy, 1987). The two most popular robust methods are *M-estimators* and the *least-median-of-squares* (LMedS) method.[3] A good review of these methods in general and as applied to vision in particular can be found in Rousseeuw (1984), Meer et al. (1992), and Black and Rangarajan (1996). In our method, we use LMedS, which estimates the parameters by solving the nonlinear minimization of the residual r_i,

$$\min_i \text{median}\, r_i^2. \tag{3}$$

The median is a preferred estimator as it has a high breakdown point. In fact, experiments prove that this method is very robust to outliers due to either bad localization or false matches (Zhang and Faugeras, 1992). However, unlike M-estimators, the LMedS problem cannot be reduced to a weighted least-squares problem, thus complicating its computation. It is also a well-known fact that the efficiency of LMedS is low in the presence of Gaussian noise (Rousseeuw, 1984).[4] As discussed before, the noise in the structure estimates deviates appreciably from Gaussianity and thus LMedS is a good choice for our application.

Given that the two-frame depth observations are non-Gaussian, a linear mean square error estimator like the Kalman filter is optimal (in the minimum variance sense) only among the class of linear estimators (for the Gaussian case, the Kalman filter is the minimum mean square estimator among all estimators). We must therefore search over a larger class of non-linear estimators. However, rather than search for a general non-linear estimator, we restrict our search to those estimators which minimize the median of squares. The question now is, is it possible to develop a recursive strategy for this optimization taking into account the statistics of the observations we mentioned previously? Before answering this question, however, we will present a quantitative analysis of the error in the reconstruction as a function of the error in the input image correspondences.

2.4. Theoretical Analysis

In our experimental studies, we found that it is not possible to postulate a distribution on the depth

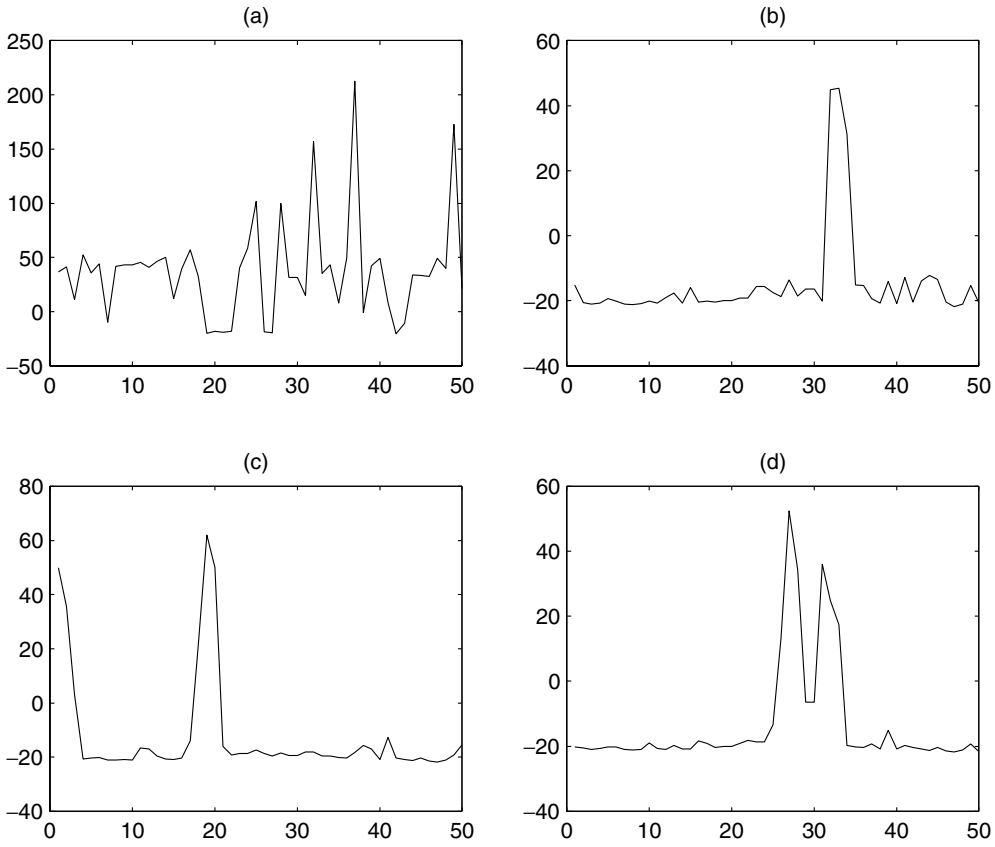

Figure 4. A plot of the depth values across 50 frames for four randomly chosen points from the face sequence. It can be seen that there are isolated outliers in all four cases.

sub-estimates, making it impossible to express the likelihood function in an analytical form. We will now show that, even though the distribution function is unknown, it is possible to infer the second-order statistics of the distribution. Specifically, we derive a closed-form expression for the covariance of structure and motion estimates as a function of the covariance of the image correspondences, which can be estimated experimentally.

Recall Eq. (2). We first consider the case where the FOE is known and then discuss the unknown FOE case.

2.4.1. Known FOE. In situations where the FOE does not change appreciably over a few frames, it is possible to estimate the FOE from the first two or three frames and assume that it remains constant for the next few frames.

In our analysis, we will follow the notation of Srinivasan (2000). Consider N points (for a sparse depth map, this denotes N feature points, while for a dense depth map it denotes the number of pixels in the image). Let us define

$$\mathbf{h} = (h_1, h_2, \ldots, h_N)^T_{N\times 1}$$

$$\mathbf{u} = (p_1, q_1, p_2, q_2, \ldots, p_N, q_N)^T_{2N\times 1}$$

$$\mathbf{r}_i = \left(x_i y_i, -\left(1 + x_i^2\right), y_i\right)^T_{3\times 1}$$

$$\mathbf{s}_i = \left(1 + y_i^2, -x_i y_i, -x_i\right)^T_{3\times 1}$$

$$\mathbf{\Omega} = (w_x, w_y, w_z)^T_{3\times 1}$$

$$\mathbf{Q} = [r_1 \quad s_1 \quad r_2 \quad s_2 \quad \ldots \quad r_N \quad s_N]^T_{2N\times 3}$$

$$\mathbf{P} = \begin{bmatrix} x_1 - x_f & 0 & \cdots & 0 \\ y_1 - y_f & 0 & \cdots & 0 \\ 0 & x_2 - x_f & \cdots & 0 \\ 0 & y_2 - y_f & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_N - x_f \\ 0 & 0 & \cdots & y_N - y_f \end{bmatrix}_{2N\times N}$$

$$\mathbf{B} = [\mathbf{P} \quad \mathbf{Q}]_{2N\times(N+3)}$$
$$\mathbf{z} = \begin{bmatrix}\mathbf{h}\\ \mathbf{\Omega}\end{bmatrix}_{(N+3)\times 1}. \tag{4}$$

Then (2) can be written as

$$\mathbf{Bz} = \mathbf{u}. \tag{5}$$

We want to compute $\mathbf{z}$ from $\mathbf{u}$. Note that for known FOE (x_f, y_f), we have linear system of equations. Let $\mathbf{z} = \psi(\mathbf{u})$. Expanding ψ in a Taylor series around $E[\mathbf{u}]$,

$$\psi(\mathbf{u}) = \psi(E[\mathbf{u}]) + D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}]) + \mathcal{O}(\mathbf{u} - E[\mathbf{u}])^2, \tag{6}$$

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,

$$\psi(\mathbf{u}) - \psi(E[\mathbf{u}]) = D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}]). \tag{7}$$

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

$$\begin{aligned}\mathbf{R_z} &= E[(\psi(\mathbf{u}) - E[\psi(\mathbf{u})])(\psi(\mathbf{u}) - E[\psi(\mathbf{u})])^T]\\ &= E[D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}])\\ &\quad \times (\mathbf{u} - E[\mathbf{u}])^T (D_\psi(E[\mathbf{u}]))^T]\\ &= D_\psi(E[\mathbf{u}])\mathbf{R_u} D_\psi(E[\mathbf{u}])^T\end{aligned} \tag{8}$$

where $\mathbf{R_u}$ is the covariance matrix of $\mathbf{u}$ and we have used the first order approximation that $E[\mathbf{z}] = \psi(E[\mathbf{u}])$. Now consider the cost function

$$\begin{aligned}C &= \frac{1}{2}\|\mathbf{Bz} - \mathbf{u}\|^2\\ &= \frac{1}{2}\sum_{i=1}^{n=2N}\left(u_i - \sum_{j=1}^{N+3} b_{ij} z_j\right)^2\\ &= \frac{1}{2}\sum_{i=1}^{n} C_i^2(u_i, \mathbf{z})\\ &= \frac{1}{2}\sum_{i=1}^{N}\left(C_{pi}^2 + C_{qi}^2\right),\end{aligned} \tag{9}$$

where C_{pi} and C_{qi} are the components of the cost function corresponding to the p and q components of the motion and b_{ij} is the (i, j)th element of $\mathbf{B}$.

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 then show how the results can be extended for the case where the FOE is unknown.

Theorem 1. *Define*

$$\begin{aligned}A_{\bar{i}p} &= \left[0\cdots 0\ -(x_{\bar{i}} - x_f)\ 0\cdots 0\ -x_{\bar{i}}y_{\bar{i}}\ \left(1 + x_{\bar{i}}^2\right)\ -y_{\bar{i}}\right],\\ &= [-(x_{\bar{i}} - x_f)\mathbf{I}_{\bar{i}}(N) | -\mathbf{r}_{\bar{i}}] = [A_{\bar{i}ph} \mid A_{\bar{i}pm}]\\ A_{\bar{i}q} &= \left[0\cdots 0\ -(y_{\bar{i}} - y_f)\ 0\cdots 0\ -\left(1 + y_{\bar{i}}^2\right)\ x_{\bar{i}}y_{\bar{i}}(N)\ x_{\bar{i}}\right],\\ &= [-(y_{\bar{i}} - y_f)\mathbf{I}_{\bar{i}}(N) \mid -\mathbf{s}_{\bar{i}}] = [A_{\bar{i}qh} \mid A_{\bar{i}qm}]\end{aligned} \tag{10}$$

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 $i = 1, \ldots, n = 2N$) and $\mathbf{I}_n(N)$ denotes a 1 in the nth 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 pth and qth components of the motion in (2). *Then*

$$\mathbf{R_z} = \mathbf{H}^{-1}\left(\sum_i \frac{\partial C_i^T}{\partial\mathbf{z}}\frac{\partial C_i}{\partial\mathbf{u}}\mathbf{R_u}\frac{\partial C_i^T}{\partial\mathbf{u}}\frac{\partial C_i}{\partial\mathbf{z}}\right)\mathbf{H}^{-T} \tag{11}$$

$$= \mathbf{H}^{-1}\left(\sum_{\bar{i}=1}^{N}\left(A_{\bar{i}p}^T A_{\bar{i}p} R_{u\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q} R_{u\bar{i}q}\right)\right)\mathbf{H}^{-T}, \tag{12}$$

and

$$\mathbf{H} = \sum_{\bar{i}=1}^{N}\left(A_{\bar{i}p}^T A_{\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q}\right). \tag{13}$$

2.5. *Proof of Error Covariance Result*

We use the implicit function theorem (Walter, 1976) to prove the above result. It has been used previously for the derivation of the uncertainty in the fundamental matrix (Faugeras, 1993) and for establishing partial results on the uniqueness of the structure and motion parameters when a long sequence is used (Broida, 1985). It was used in Fessler (1996) for error calculations in medical imaging applications. We use it here to derive explicit expressions for error covariance in terms of the parameters of (2).

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 Walter (1976).

Let $\mathbf{f}$ be a $\mathcal{C}'$ mapping of an open set $E \subset \Re^{n+m}$ into $\Re^n$, 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 $\mathbf{x} \in \Re^n$) is invertible. Then there exist open sets $U \in \Re^{n+m}$ and $W \in \Re^m$, 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

$$\mathbf{g}'(\mathbf{b}) = -(A_x)^{-1} A_y . \diamond \tag{14}$$

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

$$\phi = \frac{\partial C^T}{\partial \mathbf{z}}, \quad \text{and} \quad \mathbf{H} = \frac{\partial \phi}{\partial \mathbf{z}}. \tag{15}$$

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

$$D_\psi(\mathbf{u}) = -\mathbf{H}^{-1} \frac{\partial \phi}{\partial \mathbf{u}}. \tag{16}$$

Thus (8) becomes

$$\mathbf{R_z} = \mathbf{H}^{-1} \frac{\partial \phi}{\partial \mathbf{u}} \mathbf{R_u} \frac{\partial \phi^T}{\partial \mathbf{u}} \mathbf{H}^{-T}. \tag{17}$$

Then from (9) and (15),

$$\begin{aligned} \phi &= \frac{\partial C^T}{\partial \mathbf{z}} = \sum_i C_i \frac{\partial C_i^T}{\partial \mathbf{z}} \\ \mathbf{H} &= \frac{\partial \phi}{\partial \mathbf{z}} = \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{z}} + C_i \sum_i \frac{\partial^2 C_i^T}{\partial \mathbf{z}^2} \\ &\approx \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{z}} \\ \frac{\partial \phi}{\partial \mathbf{u}} &\approx \sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}}. \end{aligned} \tag{18}$$

Thus Eq. (17) becomes

$$\mathbf{R_z} = \mathbf{H}^{-1} \left(\sum_{ij} \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}} \mathbf{R_u} \frac{\partial C_j^T}{\partial \mathbf{u}} \frac{\partial C_j}{\partial \mathbf{z}} \right) \mathbf{H}^{-T}, \tag{19}$$

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 (9), we get

$$\frac{\partial C_i}{\partial \mathbf{z}} = \begin{cases} A_{\bar{i}p}, & i \text{ odd} \\ A_{\bar{i}q}, & i \text{ even} \end{cases}, \tag{20}$$

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

$$\begin{aligned} \frac{\partial C_i}{\partial \mathbf{u}} &= \left[\frac{\partial C_i}{\partial p_1} \; \frac{\partial C_i}{\partial q_1} \cdots \frac{\partial C_i}{\partial p_N} \; \frac{\partial C_i}{\partial q_N} \right], \\ &= \mathbf{I}_i(2N), \end{aligned} \tag{21}$$

as a $1 \times 2N$ dimensional array. Hence the Hessian in (18) becomes

$$\mathbf{H} = \sum_{\bar{i}=1}^{N} \left(A_{\bar{i}p}^T A_{\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q} \right). \tag{22}$$

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 (18), the expression for the Hessian in (22) 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.

Assuming that the feature points as well as the components of the motion vector at each feature point are uncorrelated with each other, $\mathbf{R_u} = \text{diag}[R_{u\bar{i}p}, R_{u\bar{i}q}]_{\bar{i}=1,\ldots,N}$. (Note that this condition is weaker than the one required to prove the optimality of the least squares criterion according to the Gauss-Markov theorem (Shao, 1998).) Then we can obtain a simpler relationship for the error covariances in (19):

$$\begin{aligned} \mathbf{R_z} &= \mathbf{H}^{-1} \left(\sum_i \frac{\partial C_i^T}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}} \mathbf{R_u} \frac{\partial C_i^T}{\partial \mathbf{u}} \frac{\partial C_i}{\partial \mathbf{z}} \right) \mathbf{H}^{-T} \\ &= \mathbf{H}^{-1} \left(\sum_{\bar{i}=1}^{N} \left(A_{\bar{i}p}^T A_{\bar{i}p} R_{u\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q} R_{u\bar{i}q} \right) \right) \mathbf{H}^{-T}. \end{aligned} \tag{23}$$

Equations (22) and (23) prove the statement of Theorem 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 (23) simplifies to

$$\begin{aligned}\mathbf{R_z} &= \mathbf{H}^{-1}(r^2\mathbf{H})\mathbf{H}^{-T}\\ &= r^2\mathbf{H}^{-1}.\end{aligned} \tag{24}$$

It should be noted that the assumption of uncorrelatedness of the noise in the features is invoked only at the end of the calculations. An advantage of our derivation is that we can obtain the most general expression for the covariance in (19). Thereafter the different assumptions are introduced. In practice, these assumptions can be used only if they are valid. Thus depending upon the validity of the assumptions, different expressions for the covariance in (19), (23) or (24) can be used. For a dense flow field, (23) or (24) cannot be used.

2.6. Unknown FOE

When the focus of expansion in (2) is unknown, the linear form of (5) is lost. The unknown vector $\mathbf{z} = [\mathbf{h}, x_f, y_f, \mathbf{\Omega}]^T = [\mathbf{h}, \mathbf{m}]^T$ and the cost function is $C = \frac{1}{2}\sum_{i=1}^{n} C_i^2 = \frac{1}{2}\sum_{i=1}^{n} < u_i - \hat{u}_i(\mathbf{z}), u_i - \hat{u}_i(\mathbf{z}) >$, where $\hat{u}_i$ is the estimate of the 2D motion vector obtained by projecting the reconstructed scene according to (2) and $< \cdot >$ denotes inner product. However, our method of deriving the error covariances using the implicit function theorem allows us to use the same method to derive the error covariances in this general case. The derivation presented above remains exactly the same except that we need to redefine the two vectors $A_{\bar{i}p}$ and $A_{\bar{i}q}$ as follows:

$$\begin{aligned}A_{\bar{i}p} &= [-(x_{\bar{i}} - x_f)\mathbf{I}_{\bar{i}}(N) \mid h_{\bar{i}}\ 0 - \mathbf{r}_{\bar{i}}],\\ &= [A_{\bar{i}ph} \mid A_{\bar{i}pm}],\\ A_{\bar{i}q} &= [-(y_{\bar{i}} - y_f)\mathbf{I}_{\bar{i}}(N) \mid 0\ h_{\bar{i}} - \mathbf{s}_{\bar{i}}],\\ &= [A_{\bar{i}qh} \mid A_{\bar{i}qm}]\end{aligned} \tag{25}$$

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 $h_{\bar{i}}$.

2.7. Structure of $\mathbf{R_z}$

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. From (22),

$$\mathbf{H} = \begin{bmatrix}\mathbf{H_h} & \mathbf{H_{hm}}\\ \mathbf{H_{hm}}^T & \mathbf{H_m}\end{bmatrix} \tag{26}$$

where

$$\begin{aligned}\mathbf{H_h} &= \sum_{\bar{i}=1}^{N}\left(A_{\bar{i}ph}^T A_{\bar{i}ph} + A_{\bar{i}qh}^T A_{\bar{i}qh}\right)\\ \mathbf{H_{hm}}^T &= \sum_{\bar{i}=1}^{N}\left(A_{\bar{i}pm}^T A_{\bar{i}ph} + A_{\bar{i}qm}^T A_{\bar{i}qh}\right)\\ \mathbf{H_m} &= \sum_{\bar{i}=1}^{N}\left(A_{\bar{i}pm}^T A_{\bar{i}pm} + A_{\bar{i}qm}^T A_{\bar{i}qm}\right).\end{aligned} \tag{27}$$

Thus

$$\mathbf{H_h} = \begin{bmatrix}(x_1 - x_f)^2 + (y_1 - y_f)^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & (x_N - x_f)^2 + (y_N - y_f)^2\end{bmatrix} \tag{28}$$

and

$$\mathbf{H_m} = \sum_{\bar{i}=1}^{N}\left(\mathbf{r}_{\bar{i}}^T\mathbf{r}_{\bar{i}} + \mathbf{s}_{\bar{i}}^T\mathbf{s}_{\bar{i}}\right). \tag{29}$$

Then the inverse of $\mathbf{H}$ (assuming it exists) is (Golub and Van Loan, 1989)

$$\mathbf{H}^{-1} = \begin{bmatrix}\mathbf{Q} & \mathbf{S}\\ \mathbf{S}^T & \mathbf{G}\end{bmatrix} \tag{30}$$

with

$$\begin{aligned}\mathbf{Q} &= \left(\mathbf{H_h} - \mathbf{H_{hm}}\mathbf{H_m}^{-1}\mathbf{H_{hm}}^T\right)^{-1}\\ \mathbf{G} &= \left(\mathbf{H_m} - \mathbf{H_{hm}}^T\mathbf{H_h}^{-1}\mathbf{H_{hm}}\right)^{-1}\\ \mathbf{S} &= -\mathbf{Q}\mathbf{H_{hm}}\mathbf{H_m}^{-1}.\end{aligned} \tag{31}$$

From (23)

$$\begin{aligned}
&\sum_{\bar{i}=1}^{N} \left(A_{\bar{i}p}^T A_{\bar{i}p} R_{u\bar{i}p} + A_{\bar{i}q}^T A_{\bar{i}q} R_{u\bar{i}q}\right) \\
&\quad = \begin{bmatrix} \sum_{\bar{i}=1}^{N} \left(A_{\bar{i}ph}^T A_{\bar{i}ph} R_{u\bar{i}p} + A_{\bar{i}qh}^T A_{\bar{i}qh} R_{u\bar{i}q}\right) & \sum_{\bar{i}=1}^{N} \left(A_{\bar{i}ph}^T A_{\bar{i}pm} R_{u\bar{i}p} + A_{\bar{i}qh}^T A_{\bar{i}qm} R_{u\bar{i}q}\right) \\ \sum_{\bar{i}=1}^{N} \left(A_{\bar{i}pm}^T A_{\bar{i}ph} R_{u\bar{i}p} + A_{\bar{i}qm}^T A_{\bar{i}qh} R_{u\bar{i}q}\right) & \sum_{\bar{i}=1}^{N} \left(A_{\bar{i}pm}^T A_{\bar{i}pm} R_{u\bar{i}p} + A_{\bar{i}qm}^T A_{\bar{i}qm} R_{u\bar{i}q}\right) \end{bmatrix} \\
&\quad \triangleq \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix}
\end{aligned} \tag{32}$$

with

$$\begin{aligned}
&\sum_{\bar{i}=1}^{N} \left(A_{\bar{i}ph}^T A_{\bar{i}ph} R_{u\bar{i}p} + A_{\bar{i}qh}^T A_{\bar{i}qh} R_{u\bar{i}q}\right) \\
&\quad = \begin{bmatrix} (x_1 - x_f)^2 \sigma_{p1}^2 + (y_1 - y_f)^2 \sigma_{q1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (x_N - x_f)^2 \sigma_{pN}^2 + (y_N - y_f)^2 \sigma_{qN}^2 \end{bmatrix},
\end{aligned} \tag{33}$$

where $\sigma_{p\bar{i}}^2$ and $\sigma_{q\bar{i}}^2$ are the variances of the p and q motion components for the i-th feature point (i.e. $R_{u\bar{i}p} = \sigma_{p\bar{i}}^2$ and $R_{u\bar{i}q} = \sigma_{q\bar{i}}^2$). Then substituting (28) and (32) into (23), we obtain a partition for $\mathbf{R_z}$ as

$$\mathbf{R_z} = \begin{bmatrix} \mathbf{R_h} & \mathbf{R_{hm}} \\ \mathbf{R_{hm}}^T & \mathbf{R_m} \end{bmatrix} \tag{34}$$

$$= \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{G} \end{bmatrix}^T. \tag{35}$$

Under the simplifying assumptions of Eq. (24), the partition of $\mathbf{R_z}$ can be obtained from the partition of $\mathbf{H}$ directly. Thus

$$\mathbf{R_h} = r^2 \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{G} \end{bmatrix}. \tag{36}$$

This is precisely the expression for the covariance and Cramer-Rao lower bound (CRLB) derived in Young and Chellappa (1992) under an IID Gaussian noise assumption. This should be the case since the least squares technique is optimal under these conditions (the Gauss-Markov theorem (Shao, 1998)).

2.8. Estimating the Covariance of the Feature Points

The covariance of the feature points is in principle a function of the tracking algorithm, its parameters and the image intensity function in the neighborhood of the tracked points. We try to estimate the covariance of the feature points due to measurement errors caused primarily due to localization of the points. We use 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 (Sun et al., 2001). If $\mathbf{x} = [u(i, j), v(i, j)]^T$ represents the motion estimate in the x and y directions respectively at a point (i, j), then the error covariance at that point can be estimated by the inverse of the Hessian matrix as

$$\mathbf{R_u} = \begin{bmatrix} \dfrac{\partial^2 I(i,j)}{\partial x^2} & \dfrac{\partial^2 I(i,j)}{\partial x \partial y} \\ \dfrac{\partial^2 I(i,j)}{\partial x \partial y} & \dfrac{\partial^2 I(i,j)}{\partial y^2} \end{bmatrix}^{-1}, \tag{37}$$

where $I(i, j)$ is the intensity at the point (i, j).

3. Fusion Algorithm Using Stochastic Approximation

3.1. Problem Formulation and Notation

Figure 5 shows a block-diagram schematic of the complete multi-frame fusion algorithm. The input is a video sequence of a static scene captured by a moving camera.

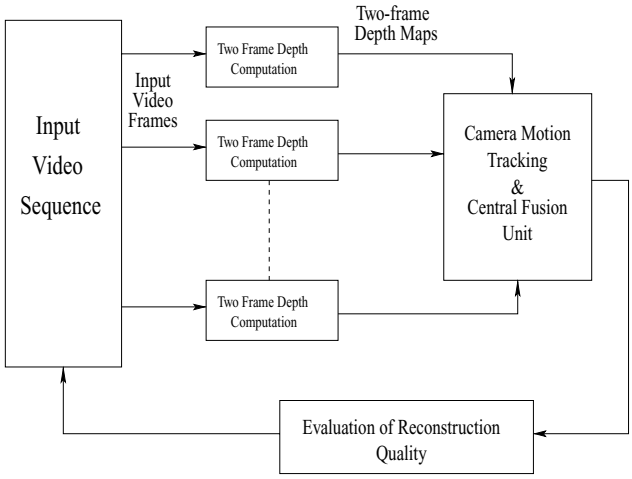

Figure 5. Block diagram of the multi-frame fusion system.

We choose an appropriate two-frame depth reconstruction strategy. The two frames maybe adjacent ones, or maybe a few frames farther apart. However, the constraint of small motion in optical flow estimation needs to be borne in mind. For simplicity of mathematical notation, we will use the case of adjacent pair of frames for explaining our algorithm. Our problem is to design an efficient algorithm to align the depth maps onto a single frame of reference (since the camera is moving), fuse the aligned depth maps in an appropriate way and evaluate the quality of the final reconstruction in order to optimize the fusion strategy and design a stopping criterion. All this needs to be done after due consideration is given to the possible sources of errors and their effects as outlined in the previous section.

Since we are dealing with multiple points and multiple frames, it is worthwhile to present our notational conventions to avoid confusion. Subscripts will refer to feature points and superscripts will refer to frame numbers. Thus x_i^j refers to the variable x for the i-th feature point in the j-th frame. When either of them is omitted, it means that the expressions are valid irrespective of the omitted feature point or frame number. As an example, when we referred to p_i previously, it meant the horizontal component of the image velocity at the i-th feature point, and is true for any frame. Similarly, when we use $\mathbf{s}^i$ in the analysis below, it represents the structure at the i-th frame, and the expression is true for every feature point.

The problem can be stated as follows. Let $\mathbf{s}^i \in \mathbf{R}^3$ represent the structure, computed for a particular point, from i-th and $(i+1)$-st frame, $i = 1, \ldots, K$, where the total number of frames is $K+1$. Let the fused structure sub-estimate at the i-th frame be denoted by $\mathbf{S}^i \in \mathbf{R}^3$. Let $\mathbf{\Omega}^i$ and $\mathbf{V}^i$ represent the rotation and translation of the camera between the i-th and $(i+1)$-st frames. Note that the camera motion estimates are valid for all the points in the object in that frame. The 3×3 rotation matrix $\mathbf{P}^i$ describes the change of coordinates between times i and $i+1$, and is orthonormal with positive determinant. When the rotational velocity $\mathbf{\Omega}$ is held constant between time samples, $\mathbf{P}$ is related to $\mathbf{\Omega}$ by $\mathbf{P} = e^{\hat{\mathbf{\Omega}}}$.[5]The fused sub-estimate $\mathbf{S}^i$ can now be transformed as $T^i(\mathbf{S}^i) = \mathbf{P}^i\mathbf{S}^i + \mathbf{V}^{i\,T}$. But in order to do this, we need to estimate the motion parameters $\mathbf{V}$ and $\mathbf{\Omega}$. Since we can determine only the direction of translational motion $(v_x/v_z, v_y/v_z)$, we will represent the motion components by the vector $\mathbf{m} = [\frac{v_x}{v_z}, \frac{v_y}{v_z}, \omega_x, \omega_y, \omega_z]$. To keep the notation simple, m will be used to denote each of the components of $\mathbf{m}$. Thus, the problems at stage $(i+1)$ will be to i) reliably track the motion parameters obtained from the two-frame solutions, and ii) fuse $\mathbf{s}^{i+1}$ and $T^i(\mathbf{S}^i)$. If $\{l^i\}$ is the transformed sequence of inverse depth values with respect to a common frame of reference, then the optimal value of the depth at the point under consideration is obtained as

$$u^* = \arg\min_u \text{median}_i \left(w_l^i (l^i - u)^2 \right), \qquad (39)$$

where $w_l^i = (\bar{R}_{\mathbf{h}}^i(l))^{-1}$, with $\bar{R}_{\mathbf{h}}^i(l)$ representing the covariance of l^i (which can be obtained from (35)).

However, since we will be using a recursive strategy, it is not necessary to align all the depth maps to a common frame of reference a priori. We will use a Robbins-Monro stochastic approximation (RMSA) algorithm (refer to Appendix) where it is enough to align the fused sub-estimate and the two-frame depth for each pair of frames and proceed as more images become available.

3.2. *The Fusion Algorithm*

For each feature point, we compute $X^i(u) = w_l^i(l^i - u)^2, u \in \mathcal{U}$. Our aim is to compute the median (say θ) of $X^0, \ldots, X^K$ i.e. to obtain θ such that $g(\theta) = F_X(\theta) - 0.5 = 0$, where $F_X(\theta)$ is the distribution function of θ. Define $Y^k(\hat{\theta}^k) = p^k(\hat{\theta}^k) - 0.5$, where $p^k(\hat{\theta}^k) = \mathbf{I}_{[X^k \le \hat{T}^k(\hat{\theta}^k)]}$ ($\mathbf{I}$ represents the indicator function, $\hat{T}^k$ is the estimate of the camera motion and $\hat{\theta}^k$ is the estimate obtained at the kth stage). Then

$$\begin{aligned} E[Y^k(\hat{\theta}^k) \mid \hat{\theta}^k] &= E[p^k(\hat{\theta}^k) \mid \hat{\theta}^k] - 0.5 \\ &= E\big[\mathbf{I}_{[X^k \le \hat{T}^k(\hat{\theta}^k)]}\big] - 0.5 \\ &= P(X^k \le \hat{T}^k(\hat{\theta}^k)) - 0.5 \\ &= F_X(\hat{\theta}^k) - 0.5 = g(\hat{\theta}^k). \end{aligned}$$

Then the RM recursion for the problem is (Robbins and Monro, 1951)

$$\hat{\theta}^{k+1} = \hat{T}^k(\hat{\theta}^k) - a^k(p^k(\hat{\theta}^k) - 0.5), \qquad (40)$$

where a^k is determined by (53). When $k = K$, we obtain the fused inverse depth $\hat{\theta}^{K+1}$, from which we can get the fused depth value $\mathbf{S}^{K+1}$.

3.2.1. Motivation to use SA. "Stochastic approximation ... may be considered as a recursive estimation method, updated by an appropriately weighted, arbitrarily chosen error corrective term, with the only requirement that in the limit it converges to the true parameter value sought" (Saridis, 1974). Recall that in Section 2 we outlined the difficulty of choosing an appropriate distribution of the noise in the depth sub-estimates due the multiplicity of sources of error that combine in a complicated manner and the danger of assuming a statistical model that is incorrect and will produce erroneous reconstructions. Stochastic Approximation (SA) provides an elegant tool to deal with such problems since we do not need to know the distribution function of the error. Besides, it provides a recursive algorithm and guarantees local optimality of the estimate, which can be non-linear. On the other hand, the Kalman filter is optimal only among the class of linear filters (in the mean square sense) for any noise distribution. For the Gaussian distribution, it is an optimal filter in the mean square sense. Since LMedS is a non-linear estimator and the distribution of the depth sub-estimates is unknown, SA is used to obtain an optimal solution based on the method of calculating the quantile of any distribution recursively, proposed originally by Robbins and Monro in their seminal paper (Robbins and Monro, 1951). The issues of convergence and optimality of RMSA have been studied in depth and we direct the interested reader to some excellent references in Appendix.

3.3. *Optimal Camera Motion Estimation*

Since depth and rotational motion are dependent on each other, there is every reason to be suspicious of the camera motion values also. However, experimental analysis has shown that the camera motion is less prone to outliers than the depth estimates. A possible reason for this is that the camera motion is obtained using a larger number of feature points in the image and thus is less susceptible to input errors in some of the features. Our camera motion estimator is a smoothing filter which tracks the motion across the frames and removes any sharp unwanted variations. The discrete-time dynamical model of the camera motion is

$$\begin{aligned} \mathbf{m}^i &= \mathbf{m}^{i-1} + \mathbf{w}^i, \\ \mathbf{y}^i &= \mathbf{m}^i + \mathbf{v}^i. \end{aligned} \qquad (41)$$

$\mathbf{w}$ is a zero-mean white noise process with $E[\mathbf{w}^i \mathbf{w}^j] = \mathbf{Q}^i \delta(i, j)$. The observations $\mathbf{y}^i$ of the camera motion (output of the two-frame algorithm) are corrupted by a zero-mean noise process $\mathbf{v}^i$ with a diagonal covariance matrix $\mathbf{V}^i$. $\mathbf{v}$ and $\mathbf{w}$ are assumed to be mutually uncorrelated across all instants of time, i.e. $E[\mathbf{v}^i \mathbf{w}^j] = 0$ for all (i, j), and are also independent of the parameter $\mathbf{m}^i$ at all time instants. We are interested in designing a linear mean square error (LMSE) estimator of the camera motion $\mathbf{m}^t$ based on the observations $\mathbf{y} = [\mathbf{y}^t, \mathbf{y}^{t-1}, \ldots, \mathbf{y}^{t-k+1}]'$. Let $\hat{\mathbf{m}}^{t|s}$ denote the estimate of $\hat{\mathbf{m}}^t$ based on the observations $[\mathbf{y}^1, \ldots, \mathbf{y}^s]$ and $\Sigma^{t|s} = E[(\mathbf{m}^t - \hat{\mathbf{m}}^{t|s})(\mathbf{m}^t - \hat{\mathbf{m}}^{t|s})']$. Then the LMSE

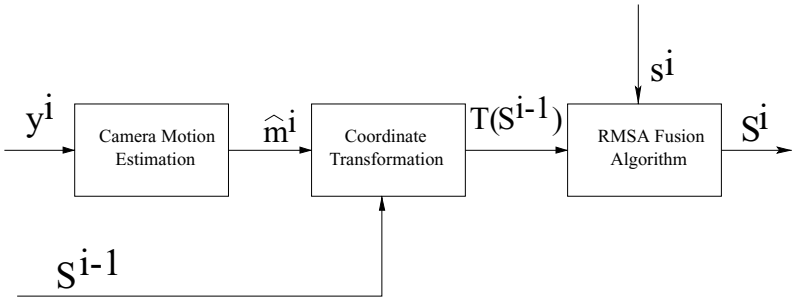

Figure 6. Block diagram of the multi-frame fusion algorithm.

estimate can be obtained from the Kalman filtering algorithm as follows. Re-indexing the observation vector $\mathbf{y}$ as $[\mathbf{y}^k, \ldots, \mathbf{y}^1]$, the Kalman filter is given by the following recursion (Poor, 1988):

$$\begin{aligned} \hat{\mathbf{m}}^{k\,|\,k} &= \hat{\mathbf{m}}^{k\,|\,k-1} + K^k(y^k - \hat{\mathbf{m}}^{k\,|\,k}) \\ \hat{\mathbf{m}}^{k\,|\,k-1} &= \hat{\mathbf{m}}^{k-1\,|\,k-1} \\ K^k &= \Sigma^{k\,|\,k-1}[\mathbf{V}^k + \Sigma^{k\,|\,k-1}]^{-1} \\ \Sigma^{k\,|\,k-1} &= \Sigma^{k-1\,|\,k-1} + Q^k. \end{aligned} \tag{42}$$

Then $\Sigma_{\mathbf{y}^k} = E[(\mathbf{y}^k - E[\mathbf{y}^k])(\mathbf{y}^k - E[\mathbf{y}^k])'] = E[(\mathbf{m}^k + \mathbf{v}^k - \mu_{\mathbf{m}})(\mathbf{m}^k + \mathbf{v}^k - \mu_{\mathbf{m}})'] = E[(\mathbf{m}^k - \mu_{\mathbf{m}})(\mathbf{m}^k - \mu_{\mathbf{m}})'] + \mathbf{V}^k = \mathbf{R}^k_{\mathbf{m}}$, where $\mu_{\mathbf{m}} = E[\mathbf{m}^i] = E[\mathbf{m}^{i-1}] = E[\mathbf{y}^i]$. Thus the observation noise covariance can be estimated from (35) and the camera motion filter is derived.

3.3.1. Why Kalman Filter? Since the system dynamics of the camera motion are time-varying, SA techniques are not guaranteed to converge. One heuristic that is commonly applied is to choose the step size a_k in (52) to be a small positive number as a trade-off between tracking capability and noise sensitivity (Chapter 2 of Ljung and Soderstrom, 1987). In our experiments, we found that it was possible to make the step-size small and constant and make the algorithm converge. However, for different kinds of camera motions, this constant had to be different. This is a problem as it would require tuning the parameter every time the camera motion changed, without any guarantee of optimality of performance. Hence, we settled for a tracking algorithm using the Kalman filter and analytically computing the error covariances. Also, the presence of outliers in two-frame camera motion estimates is less pronounced than in the depth sub-estimates; hence least squares is a good criterion for tracking camera motion. The difficulty of incorporating time-varying dynamics into the SA approach, coupled with the suitability of a least squares criterion, dictates the choice of the Kalman filter for camera motion estimation.

3.4. The Algorithm

Assume that we have the fused 3D structure $\mathbf{S}^i$ obtained from i frames and the 2-frame depth map $\mathbf{s}^{i+1}$ computed from the i-th and $(i+1)$-st frames. Figure 6 shows a block diagram of the multi-frame fusion algorithm. The main steps of the algorithm are:

Track Estimate the camera motion according to the optimal linear camera motion estimation filter of (42).

Transform Transform the previous model $\mathbf{S}^i$ to the new reference frame.

Update Update the transformed model using $\mathbf{s}^{i+1}$ to obtain $\mathbf{S}^{i+1}$ from (40).

Evaluate Reconstruction Compute a performance measure for the fused reconstruction (explained in the next section).

Iterate Decide whether to stop on the basis of the performance measure. If not, set $i = i + 1$ and go back to Track.

4. Information Theoretic Analysis for Multi-Frame Algorithms

In this section, we introduce two different measures of performance for multi-frame algorithms. The first relates to the error covariance in the structure estimate as a function of the number of frames, while the second deals with the information content in the

estimate. Either one of them or their combination can be used in practice depending on the application. The first is applicable to a small number of frames, while the second criterion can be used if the number of frames is sufficiently large to produce a reliable estimate of the distribution of the two-frame depth values. The second method, however, gives a running estimate of the performance of the algorithm, as we shall explain.

A point that needs to be noted here is that we implicitly assume in our multi-frame fusion algorithm that most of the 2-frame reconstructions are of reasonable quality. If all the intermediate reconstructions are of extremely poor quality, no amount of processing will lead to a final solution which is acceptable. A very interesting question to address would be to identify such situations automatically. We are not aware of any previous work on multi-frame SfM that tries to automatically recognize situations where the quality of the intermediate estimates is so poor that it is extremely difficult, if not impossible, to obtain a reasonable final estimate. We use ideas from information theory to address this problem. Due to space constraints, we cannot include them here, but the interested reader can access a short version of it in Roy Chowdhury and Chellappa (2002).

4.1. Rate-Distortion Analysis

Rate-distortion is used to obtain an idea of the quality of the reconstruction as a function of the number of frames. Other methods, like estimating confidence regions, are based on numerical simulations (Cho et al., 1997). In many applications, this is the only way possible because of a lack of analytical structure in the basic problem formulation. We show here that for the problem of 3D reconstruction from optical flow, the error estimates can be obtained analytically without resorting to simulations. While confidence regions, estimated by methods like bootstrapping could be used, the ability to obtain analytical closed form expressions has the added advantage of analyzing the effects of different parameters on the quality of reconstruction.

Let the two-frame inverse depth values for a particular feature point be denoted by $X^1, X^2, \ldots, X^N$. For the purposes of this analysis, we will assume that the sequence of depth values does not have any outliers (it is handled by the LMedS estimator) and hence the sample mean $\bar{X}$ is an unbiased estimate of the true value X^*. 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 $E[\bar{X}] = \frac{1}{N}\sum_{i=1}^{N} E[X^i]$ and $\mathrm{Cov}[\bar{X}] = E[(\bar{X} - X^*)^2] = E[\bar{X}^2] - E[\bar{X}]^2$.

$$E[\bar{X}]^2 = \left(E\left[\frac{1}{N}\sum_{i=1}^{N} X^i\right]\right)^2 = \frac{1}{N^2}\left(\sum_{i=1}^{N} E[X^i]\right)^2 = \frac{1}{N^2}\left[\sum_{i=1}^{N} E[X^i]^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} E[X^i]E[X^j]\right], \quad i \neq j \tag{43}$$

and

$$E[\bar{X}^2] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} X^i\right)^2\right] = \frac{1}{N^2} E\left[\sum_{i=1}^{N} {X^i}^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} X^i X^j\right], \quad i \neq j = \frac{1}{N^2}\left[\sum_{i=1}^{N} E\left[{X^i}^2\right] + \sum_{i=1}^{N}\sum_{j=1}^{N} E[X^i X^j]\right], \quad i \neq j \tag{44}$$

which yields the expression for the covariance of the estimator as

$$\mathrm{Cov}[\bar{X}] = \frac{1}{N^2}\left[\sum_{i=1}^{N}\mathrm{Cov}[X^i] + \sum_{i=1}^{N}\sum_{j=1}^{N}(E[X^i X^j] - E[X^i]E[X^j])\right], \quad i \neq j. \tag{45}$$

The first summation, $\mathrm{Cov}[X^i]$, is the variance of the two-frame depth estimates obtained from $\mathbf{R}_{\mathbf{z}}^i$ in (24). Under the assumption of independence of the two-frame observations, the second term of (45) vanishes and we obtain a closed-form expression for the variance of the estimator for the N-frame SfM algorithm.The covariance of the estimate of the j-th feature point is

$$\mathrm{Cov}[\bar{X}_j] = \frac{1}{N^2}\left[\sum_{i=1}^{N} \mathbf{R}_{\mathbf{h}}^i(j, j)\right], \tag{46}$$

where $\mathbf{R}_{\mathbf{h}}^i(j, j)$ is the j-th diagonal term obtained from (35) for the i-th and $(i + 1)$-st frames. Under

the assumption of IID Gaussian noise of Young and Chellappa (1992) for the two-frame algorithm, (46) simplifies to the following form:

$$\begin{aligned}\mathrm{Cov}[\bar{X}_j] &= \frac{1}{N^2}\left[\sum_{i=1}^{N} {r^i}^2 \mathbf{Q}^i(j,j)\right], \\ &= \frac{1}{N^2}\left[\sum_{i=1}^{N} {r^i}^2 \left(\mathbf{H}^i_{\mathbf{h}} - \mathbf{H}^i_{\mathbf{hm}}\mathbf{H}^{i^{-1}}_{\mathbf{m}}\mathbf{H}^{iT}_{\mathbf{hm}}\right)^{-1}\right],\end{aligned} \tag{47}$$

where the terms are defined in (30) and (31). The expressions are valid for both the known and unknown FOE cases with $A_{\bar{i}p}$ and $A_{\bar{i}q}$ appropriately defined.

The average distortion in the reconstruction over M feature points is

$$\begin{aligned}&E_{M,N}[(\bar{X} - E[\bar{X}])^2] \\ &\quad = E_M[E_N[(\bar{X} - E[\bar{X}])^2 \mid \bar{X} = \bar{X}_j]] \\ &\quad = \frac{1}{MN^2}\sum_{j=1}^{M}\sum_{i=1}^{N} \mathbf{R}^i_{\mathbf{h}}(j,j) \\ &\quad = \frac{1}{MN^2}\sum_{i=1}^{N} \mathrm{trace}\left(\mathbf{R}^i_{\mathbf{h}}\right).\end{aligned} \tag{48}$$

Figure 7 plots the covariance of the estimator for the inverse depth as a function of frame number using (48) for the two video sequences of our experiments (Section 2.2). A few interesting observations regarding these curves can now be made.

- In the traditional rate-distortion function used in source coding (Cover and Thomas, 1991), the distortion of a signal X from its n-bit representation X_n is plotted as a function of n, distortion usually being defined in a mean square sense. In our application, we analyze the precision of the reconstruction with increasing number of frames; hence the analogy of rate with the number of frames. (Of course, the true error will be different from the plots, because of the approximations in the derivation; however, they are a very good approximation.) We will call these curves the *video rate-distortion (VRD) curves*.
- Given a specific tolerable level of distortion, each of these curves specifies the minimum number of frames necessary to achieve that level of distortion.
- The errors in SfM are due to a number of reasons, the effects of which are impossible to quantify separately. These curves give a compact representation for understanding the effects of these various sources of errors on the final estimate.
- Each curve identifies an operating point of a MFSfM algorithm as a trade-off between tolerable reconstruction error and the computational cost of considering more frames.
- The curves depend on the covariance of the image correspondences only, if the FOE is known. In situations where the FOE does not change appreciably over the image sequence of interest, it is possible to plot these curves after the first pair of frames itself (after estimating the FOE).
- Though the average distortion over all features is plotted here, the curves can also be obtained for each individual feature point using (46). Since the uncertainty of the depth estimates is a function of the feature point (as the variance of the image correspondences will depend on the particular feature), the curves can be used to identify points which are more prone to reconstruction errors and thus would require greater numbers of frames to achieve tolerable distortion.
- The nature of the plots for the unknown FOE case will remain the same, with $A_{\bar{i}p}$ and $A_{\bar{i}q}$ defined appropriately as in (25) (and now depending on h). However, they can no longer be computed without first estimating h. Hence the distortion in (48) needs to be estimated as the algorithm progresses.

4.2. Relative Information

We now introduce a second measure of fidelity for multi-frame structure reconstruction based on the information content rather than the error characteristic. If $\hat{f}$ is the density of an estimate, its Shannon, or differential, entropy is defined as Cover and Thomas (1991)

$$H(\hat{f}) = -\int \hat{f}(\eta)\ln\hat{f}(\eta)d\eta. \tag{49}$$

If $\hat{f}^1, \ldots, \hat{f}^\alpha, \ldots, \hat{f}^N$ is a time sequence of the distribution of the estimates, then the incremental differential entropy

$$\Delta I_\alpha = H(\hat{f}^{\alpha-1}) - H(\hat{f}^\alpha) \tag{50}$$

is a measure of the decrease in differential entropy. Since entropy is a measure of uncertainty, we will call (50) as the relative information (Goodman et al., 1997).

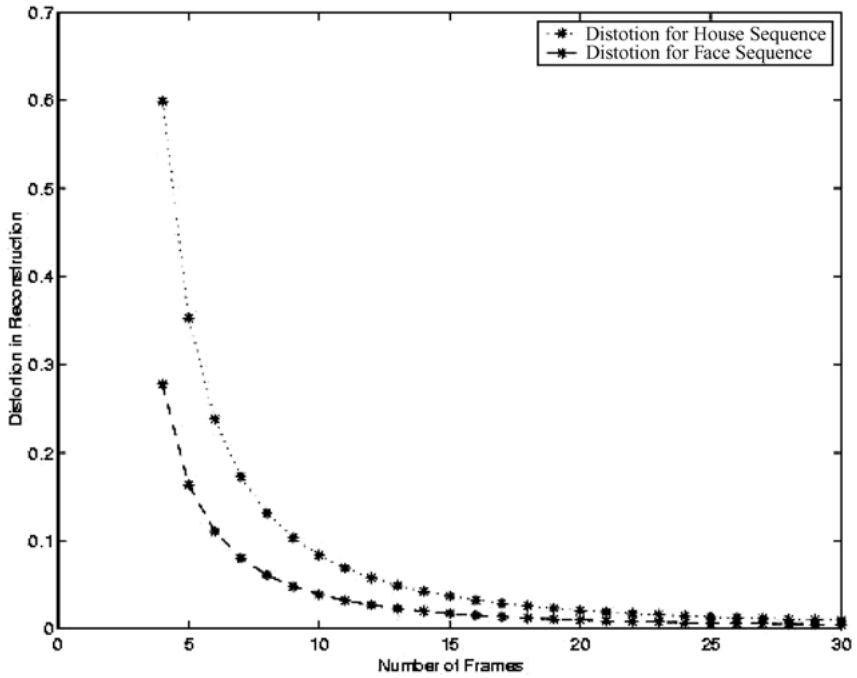

Figure 7. Plot of average distortion in reconstruction as a function of the number of frames for the two video sequences. The vertical axis is scaled down by a factor of 10^3.

Computation of the differential entropy may be difficult in practice, as we need to know the distribution of the estimates. The distribution can be estimated using histogram techniques; however, if the number of frames is small the method can be inaccurate. Our experimental results, however, show that with more than twenty frames, the results obtained by this method are similar to those for the other criterion explained above. The intuitive understanding of the method is that with more frames, the distribution of the inverse depth estimates converges, and so does the differential entropy at two consecutive time instants.

Figure 8 shows plots of the estimates of the relative information against the frame number for a random selection of nine feature points. To estimate the density functions, standard histogram techniques were used. Since this estimate is usually unreliable for a few samples, the values for the first fifteen frames are neglected. Some interesting observations can be made from these plots. As expected, the relative information converges toward zero as the number of frames increases, indicating a decrease in uncertainty. The sudden peaks or dips (plots (b), (e), (f)) correspond to outliers in the observation sequence. These values suddenly perturb the histogram estimate; however, as the number of observations increases these isolated effects die down. If the relative information converges to zero and then diverges (plots (c), (e)), it possibly means that the later values are erroneous and because of this the information content tends to increase. If after a sufficient number of frames the relative information plot does not approach zero, it implies that the information content in the observations is still significant and more observations may be required. Thus continuous monitoring of the relative information can give important clues regarding the convergence of the MFSfM algorithm.

5. Experimental Results

We now present the results of our algorithm. The main application of our work was on 3D face modeling from short monocular video sequences. We first present some simulation results with synthetic data to demonstrate the accuracy of our algorithms. Then, we present our results on the Yosemite sequence. Finally, we consider a real-life scenario of 3D face modeling and present a detailed analysis of our method. For this problem, the input video sequences were captured from

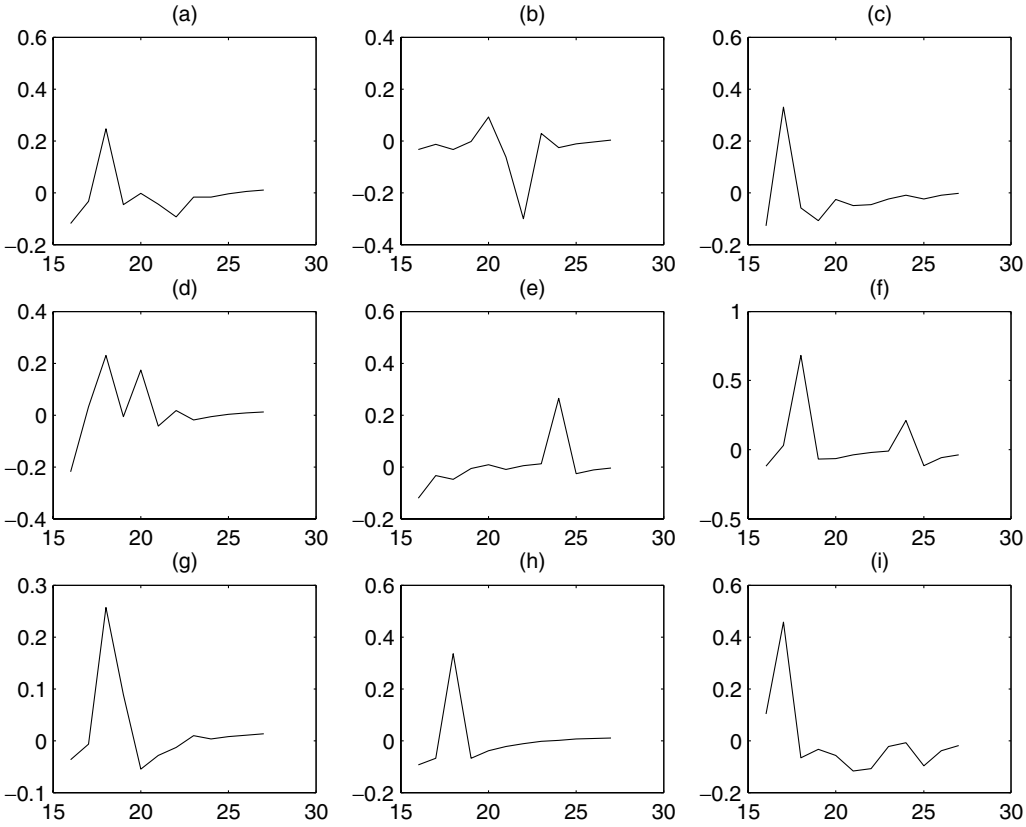

Figure 8. Plots of the relative information for a random selection of feature points. The vertical axis represents the value of the relative information and the horizontal axis represents the frame number. Histogram techniques were used to compute the probability density functions; hence the differential entropy was not computed for the first 15 frames.

a hand-held or tripod-mounted video camera. The output was a 3D model of the scene. A C and MATLAB implementation of the multi-frame fusion algorithm is available. We also have an end-to-end system for 3D reconstruction on a Pentium PC for demonstrations. Videos of all the original and reconstructed scenes presented in this paper are available from the authors.

5.1. Estimating the Statistics of Feature Position Errors

The first step in implementing our algorithm is to obtain the statistics of the feature correspondences. Given a sequence of images, the motion between the images was estimated using optical flow. In our method, although we obtain a dense flow field in order to obtain a dense depth map, we compute the statistics for a subset of points and assume them to be spatially wide-sense stationary in a small region around that point. There are many ways in which the error covariance of the individual feature points may be computed. Sun et al. (2001) proposes computing the second partial derivatives of the image intensities in order to obtain the error covariance of a point. We propose combining this with resampling techniques like bootstrapping for obtaining more robust estimates for the error covariances (Cho et al., 1997). The variance in the image correspondence for each feature was computed for the horizontal and vertical components using the technique in Sun et al. (2001) and repeated for 200 bootstrap samples and 50 initial frames (Efron and Tibshirani, 1993).[6] We conducted experiments on the two sequences that we have used throughout this paper, the face sequence and the outdoor house sequence. Figures 9 and 10 depict the estimated variance of the feature points for the horizontal and vertical components of the motion of the features. The diameters of the circles are proportional to the

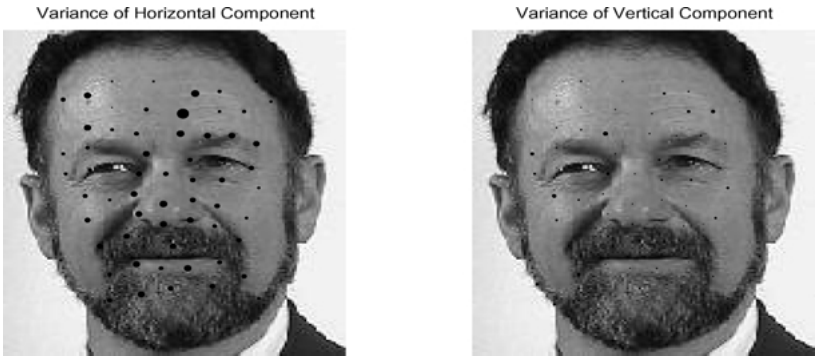

Figure 9. Plot of the variances of the feature correspondences in a face sequence. The variance is represented at the corresponding feature point in the image. The diameter of the circle is proportional to the variance for that feature point. One of the plots is for the variance in the horizontal component of the motion for that feature, while the other is for the vertical component.

variance at that point. The variances in the inverse depth estimates as obtained from (23) is also shown in Figs. 12 and 13.[7]

5.2. *Overview of Implementation Strategy*

Figure 5 represents the overall 3D reconstruction strategy based on our multi-frame fusion algorithm. The optical flow computed for every pair of images was given as the input to the two-frame SfM algorithm described in Srinivasan (2000). The output was the depth at these points and the motion of the camera between these frames. For each pair of frames, the covariance of the error in the structure and motion estimates was computed according to (23). The Kalman filter based camera motion estimator was implemented according to (42). The depth maps from pairs of frames (consecutive or a few frames distant) were aligned on the basis of the camera motion estimates. The aligned depth maps were then fused using the recursive RMSA fusion algorithm. The multi-frame distortion curve for the entire video sequence was computed at each step for the individual feature points using (46) and for their average representation using (48). When the distortion was below an accepted level, the computation was terminated and the computed depth estimate accepted as the final value. The final distortion for each feature was used to decide whether to include it in model building or not. In all the experiments the FOE was estimated from the first two or three frames and assumed constant thereafter. A discussion on the validity of this assumption can be found in Srinivasan (2000). In our experiments, we aim to reconstruct from a few

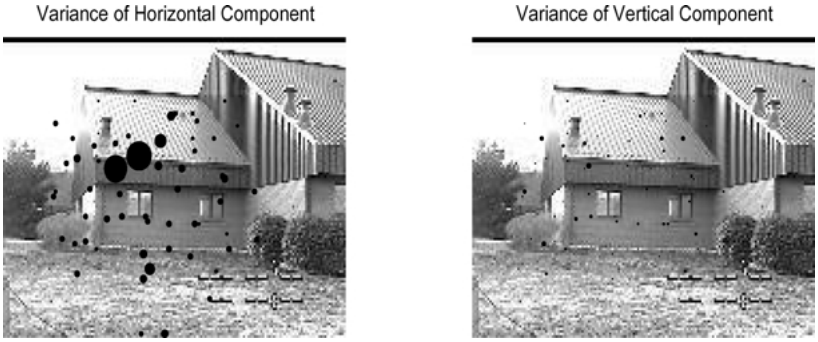

Figure 10. Plot of the variances of the feature correspondences in an outdoor sequence. The variance is represented at the corresponding feature point in the image. The diameter of the circle is proportional to the variance for that feature point. One of the plots is for the variance in the horizontal component of the motion for that feature, while the other is for the vertical component.

frames of the video sequence where the motion between the start and end of the sequence is relatively small. Hence the assumption of the constant FOE is valid. Comparison with the estimates obtained from every adjacent pair of frames showed that this was a justified assumption. The depth map obtained at this stage was used to build a 3D model using the Graphics toolbox of MATLAB. The set of feature points were used to create a Delaunay triangulation. The depth values were assigned to each of the vertices of the triangle in order to create a mesh to which the texture was mapped to create the final 3D model. The method of building the 3D model of the scene is simplistic; it is used only as a means to represent the results of the algorithm. Advanced techniques in computer graphics can definitely produce much better models of the scene.

5.3. Synthetic Data

In order to analyze the numerical accuracy of our method, we conducted some experiments with synthetically generated data. We generated a set of fifty 3D points and obtained their projections at different camera positions, where the camera motion is small enough for the optical flow equations in (2) to be valid. Thus, we have a sequence of fifty frames, each with fifty points. A subset of this was treated as input to our method. Random noise was added to the point positions. We considered the following three cases for the experimental analysis:

- The Kalman filtering approach, similar to Broida and Chellappa (1989), but assuming that the focus of expansion is known. We consider two cases where (i) the noise covariance in the feature positions is estimated, and (ii) the actual noise covariance is used.
- Our stochastic approximation algorithm after estimating the noise in feature point positions.
- Our stochastic approximation algorithm with the actual noise in feature point positions.

The results obtained using the noise estimation procedure outlined in this paper are appreciably better, as shown in Table 1, where the root mean square (RMS) error between the estimated 3D depths and the true ones is reported as a percentage of the true depth value. The trend in the results proves the efficieny of our method for 3D reconstruction with unknown noise statistics. Since the noise distribution is non-Gaussian, the Kalman filter is optimal (in the minimum mean square error sense) only within the class of linear estimators (Poor, 1988). RMSA does better since it searches over a larger class of estimators.

Table 1. The RMS error in the depth estimates as a percentage of the true depth value.

Method	Percentage error
1 Kalman filter, estimated noise	8.6
2 RMSA, estimated noise	4.7
3 Kalman filter, actual noise	3.4
4 RMSA, actual noise	2.1

5.4. Yosemite Sequence

Figure 11 shows the reconstruction of the 3D scene from the Yosemite video sequence. Fifteen frames were used for this reconstruction. The size of each image in the video was 316×252 pixels. (a) and (b) represent two frames from the original sequence. The depth map was reconstructed using the method described above and the 3D model was constructed from these values. (c) to (i) represent views of the 3D model from different angles. The 3D model is flipped with respect to the original sequence in order to obtain a better depiction of the computed depth.

5.5. Face Modeling

In the set-up for conducting these experiments in our lab, a commonly available video camera is mounted on a tripod. A person is asked to sit in front of the camera and move his/her head slowly in any desired manner. The video sequence is captured and given as an input to the MFSfM algorithm, which then reconstructs the 3D model automatically (including motion estimation). Each of the frames of the video sequence was 140×162 pixels in size. We present our detailed results on one particular video sequence. They would be similar for other sequences. The first and last frames on this sequence is shown in Fig. 14(a) and (b). From these two images it is clear that we are concerned with reconstructing the 3D structure from a short monocular video sequence. The motivation for this problem comes from the fact that the 3D model reconstructed in this

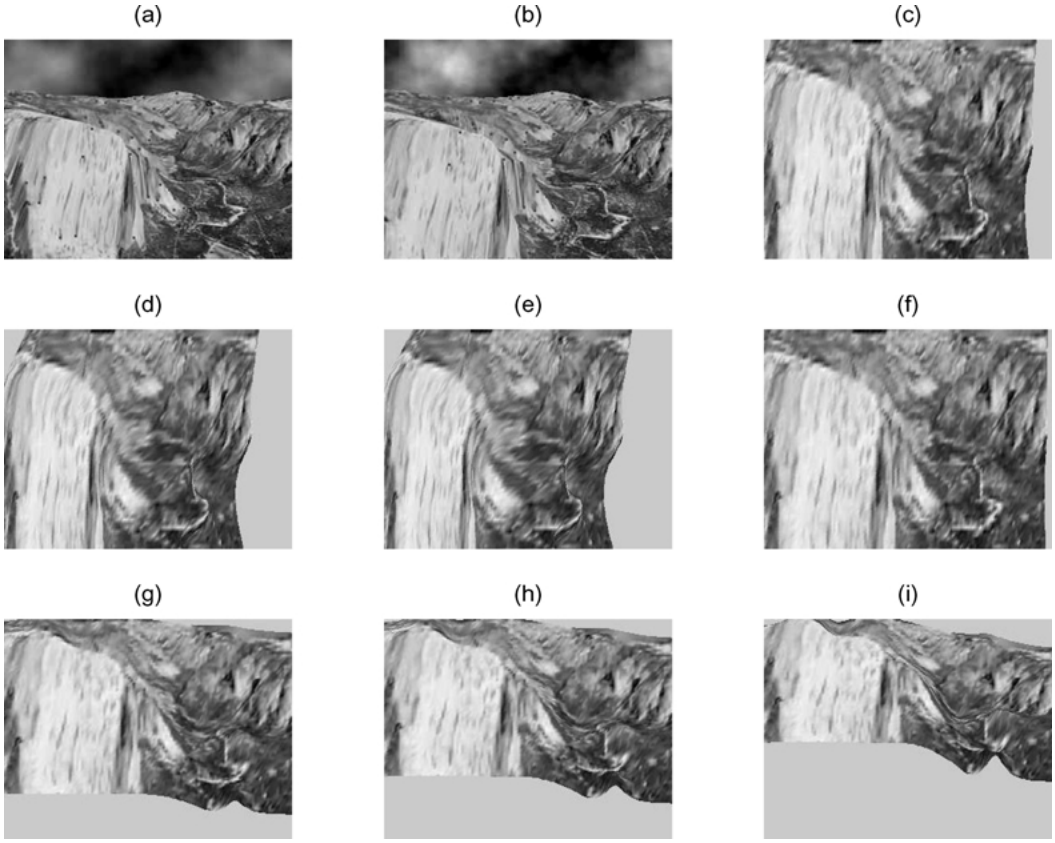

Figure 11. (a) and (b) represent two images from the Yosemite sequence for which the depth was computed. The remaining figures (c)–(i) are results of 3D reconstruction from 15 frames for different viewing angles.

way can be used for face recognition purposes across different poses by taking 2D projections at different viewing angles.

The variance of the errors in the feature positions is used to calculate the variance in the inverse depth estimates. The variance of the inverse depth estimates for each of the feature points, as computed from (23), is shown in Fig. 12. The variance is proportional to the diameter of the circle at that particular point. For the flow based method, only Z is estimated from (2) and $X = xfh, Y = yfh$ are obtained from the perspective projection model, where $h = 1/Z$ and f is the focal length of the camera. For a calibrated system (which we assume), the covariance of X and Y is proportional to the covariance of h, given a particular point (x, y) in the image. This is unlike some other methods; e.g. in the uncertainty analysis of the factorization method (Sun et al., 2001) 3D ellipsoids need to be used to depict the uncertainty in reconstruction since the 3D point (X, Y, Z) on the object is estimated. The diagonal elements of the covariance matrix $\mathbf{R_h}$ are also plotted in the same figure. From these plots, it is clear that the uncertainty in the estimates is a function of the particular feature point. It is important to understand these errors before creating the 3D model. One single point in error can seriously affect the quality of the entire model because of the interpolation techniques which are inherent in the modeling process. The video rate-distortion curve (VRD) for this sequence is shown in Fig. 7. From this plot, it is clear that about 20 frames are sufficient to obtain a reconstruction with a small level of distortion. Finally, Fig. 14 represents the 3D model generated for this video sequence using the multi-frame algorithm; a depth map image with intensity corresponding to depth and views from different camera positions not part of the original sequence are presented in order to show the effectiveness of the reconstruction.

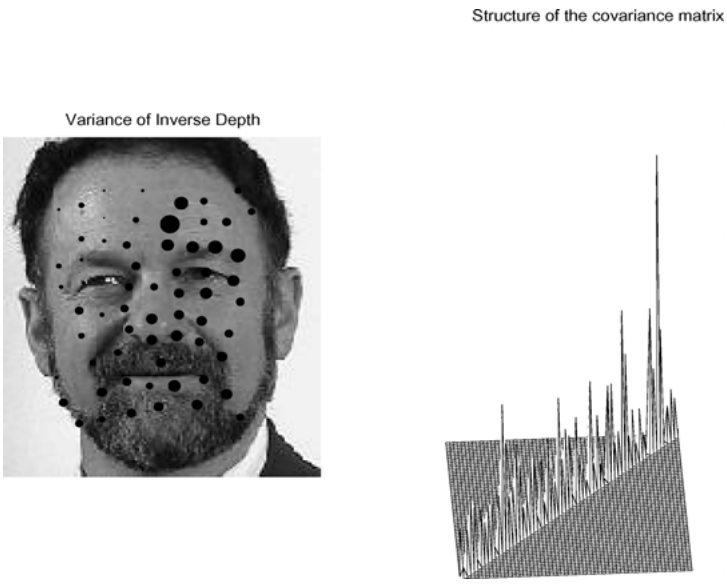

Figure 12. Plot of the variances of the inverse depth for different features in a face sequence. The variance is represented at the corresponding feature point in the image. The diameter of the circle is proportional to the variance for that feature point. In the second plot, the diagonal elements of $\mathbf{R_h}$ are shown.

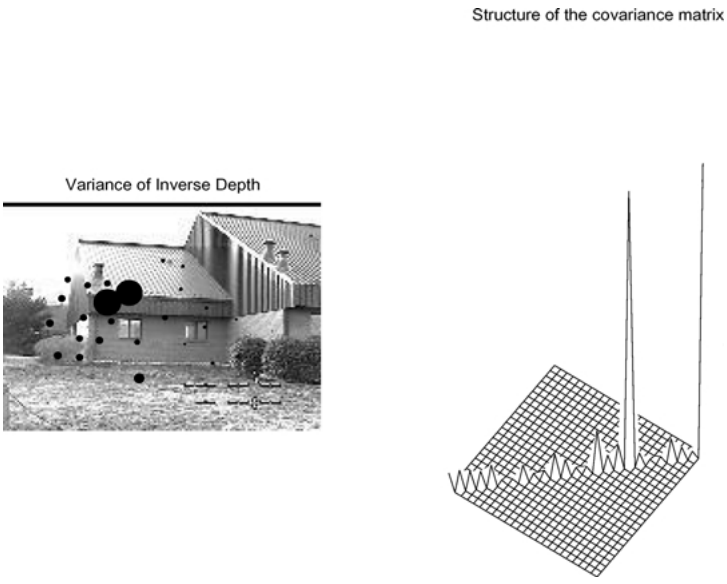

Figure 13. Plot of the variances of the inverse depth for different features in an outdoor image sequence. In the first plot, the variance is represented at the corresponding feature point in the image. The diameter of the circle is proportional to the variance for that feature point. In the second plot, the diagonal elements of $\mathbf{R_h}$ are shown.

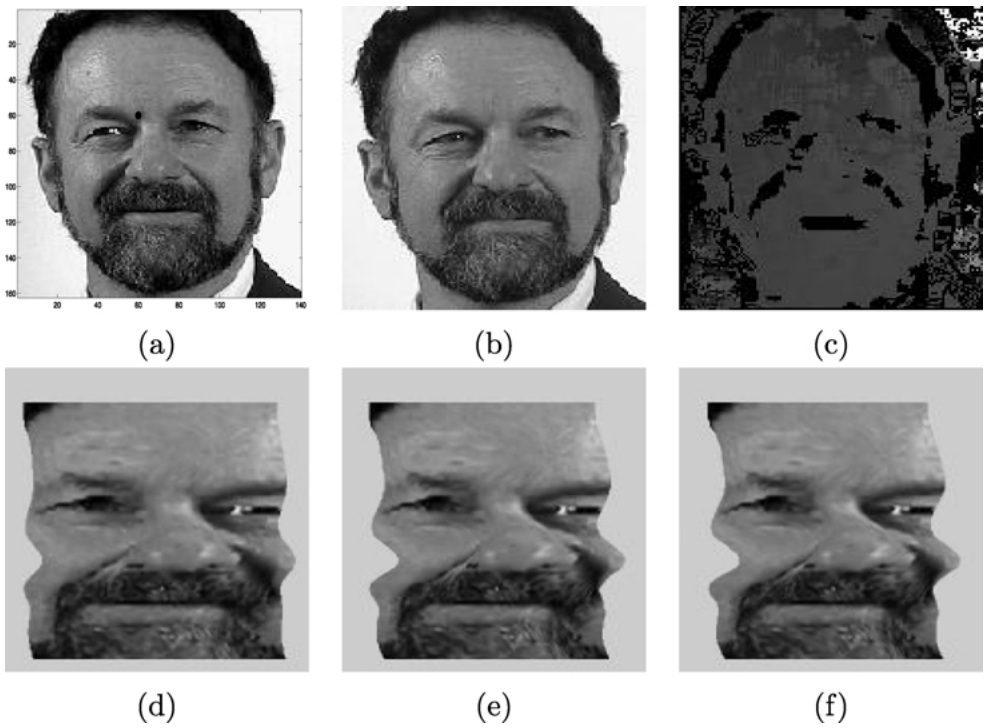

Figure 14. (a) and (b) represent two images from a face video sequence. The FOE is marked on the image in (a) (above the left eye). (c) represents the depth map, with intensity corresponding to depth. The remaining figures (d)–(f) are results of 3D reconstructions from 15 frames for different viewing angles using the RMSA algorithm.

The algorithm that we have presented in this paper is general enough to be applied to a large class of video sequences. The aim of this paper is to demonstrate how different estimation and information theoretic criteria can be used to enhance the robustness and accuracy of multi-frame SfM algorithms. However, the main application focus of our work has been on modeling 3D faces. Since the analysis and algorithm presented here would finally become part of the 3D face modeling solution, we feel it necessary to discuss it very briefly. Space constraints prohibit a more detailed description. It can however be found in another paper (Roy Chowdhury et al., 2003a).

3D modeling of faces from a video sequence usually take advantage of the fact that most faces have a similar average representation, usually termed as a generic face. In Shan et al. (2001) and Fua (2000), the authors used a generic model to initialize the SfM algorithm. While this often gives very good results, it suffers from a disadvantage that the optimization can converge to a solution very close to the initial point, resulting in a reconstruction that bears the characteristics of the generic model rather than the particular face which we are trying to reconstruct. In Roy Chowdhury et al. (2003a), we have shown that it is possible to overcome this problem by introducing the generic model at a later stage of the algorithm. This is possible because of the detailed statistical analysis that we perform in our algorithm and which has been the focus of this paper. The output of our algorithm produces a reasonably good reconstruction, and the generic model can be then be used to correct for some of the local errors which persist in order to produce a smooth reconstruction. In Fig. 15, we depict some of the views from the 3D reconstruction obtained using the multi-frame algorithm described in this paper and the generic model.

5.5.1. Accuracy of Face Modeling Algorithm. In order to analyze the accuracy of the 3D face reconstruction, we require the ground truth of the 3D models. We experimented with a publicly available database of

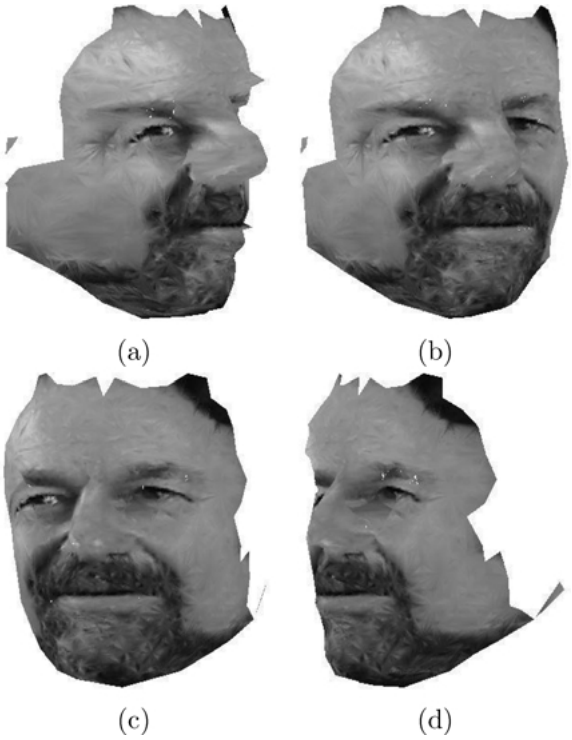

Figure 15. Different views of the 3D model after texture mapping.

3D models obtained from a Minolta 700 range scanner. The data is available on the World Wide Web at http://sampl.eng.ohio-state.edu/sampl/data/3DDB/RID/minolta/faces-hands.1299/index.html. We will report numerical results from our algorithm on some of the data available here, though we will not publish the images or 3D models of the subjects. We will use the same convention as on the website of referring to the subjects as "frame001" to "frame005."

In order to perform an accurate analysis of face-modeling method, we require a video sequence of the person and the 3D depth values. This, however, is not available on this particular database or on any other that we know of. Thus we had to generate a sequence of images in order to apply our algorithm. This was done using the 3D model and the texture map provided on the web-site. We considered the error in the 3D estimate compared to the actual 3D values. The percentage RMS error of the 3D models with respect to the true values is tabulated in Table 2.

Table 2. Average percentage error of the 3D models

Subject index	Percentage error
1 (frame 001)	3.6
2 (frame 002)	3.3
3 (frame 003)	3.2
4 (frame 004)	3.4
5 (frame 005)	3.0

6. Conclusions and Future Work

Our study provides a framework for error characterization and incorporation of robust statistics into the SfM problem. While many algorithms exist for computing the scene structure, their sensitivity to practical conditions is still an open question which we have tried to address. We developed a robust estimation theoretic framework for structure and motion computation from short, monocular video sequences and proposed a solution methodology using stochastic approximation, since the noise statistics are largely unknown. A non-linear estimate, which asymptotically converges to the true value, is obtained. A closed-form expression for the error covariance of the motion and structure estimates as a function of the error in the image correspondences is derived, without taking recourse to the standard assumptions of Gaussian noise. Propagation of the two-frame error to multi-frame reconstruction and its dependence on the number of frames is studied and a criterion based on rate-distortion theory is proposed. Experimental results of scene (especially 3D face) reconstruction, along with visualization of the errors, is presented. The extension of this work to face recognition across pose and illumination variation is being currently studied.

Appendix: Robbins-Monro Stochastic Approximation

The method of stochastic approximation (SA) is useful for certain sequential parameter estimation problems (Ljung and Soderstrom, 1987; Spall, 2000; Benveniste et al., 1987). Let $\{e(k)\}$ be a sequence of random variables with the same distribution indexed by a discrete time variable k. A function $Q(x, e(k))$ is given such that

$$E[Q(x, e(k))] = g(x) = 0 \tag{51}$$

where E denotes expectation over e. The distribution of $e(k)$ is not known; the exact form of the function

$Q(x, e)$ may also be unknown, though its values are observed and it can be constructed for any chosen x. The problem is to determine the solution of $g(x) = 0$. Robbins and Monro (RM) suggested the following scheme for solving (51) recursively as time evolves (Robbins and Monro, 1951):

$$\hat{x}(k) = \hat{x}(k-1) + a_k Q(\hat{x}(k-1), e(k)) \tag{52}$$

where the gain sequence $\{a_k\}$ must satisfy the following conditions (Benveniste et al., 1987; Ljung and Soderstrom, 1987; Spall, 2000):

$$a_k \geq 0, \quad a_k \to 0, \quad \sum_{k=1}^{\infty} a_k = \infty, \quad \sum_{k=1}^{\infty} a_k^2 < \infty. \tag{53}$$

A popular choice of the gain sequence, which was used in our experiments also, is $a_k = a/(k+1)^{0.501}$. It can be shown that the estimate obtained from SA is unbiased, consistent and asymptotically normal, and in many cases, also efficient (Ljung and Soderstrom, 1987; Poor, 1988).

Acknowledgments

The authors would like to thank Prof. Azriel Rosenfeld for his comments and suggestions in the preparation of the manuscript.

Notes

1. We have subsequently come to know that a somewhat similar method was applied for error calculations in medical imaging applications (Fessler, 1996).
2. Computation of the depth requires point correspondences between the frames using either flow-based or feature-based methods, usually with some heuristics. Zhang and Faugeras (1992) cite two sources for outliers in matching feature points across images. One of the causes is mislocation of features from their exact pixel positions and the other is mismatched feature correspondences.
3. The general viewpoint is that M-estimators are usually robust to outliers due to bad localization but not to false matches (Zhang and Faugeras, 1992).
4. The efficiency of an estimator is defined as the ratio of the lowest achievable variance of the estimated parameters (obtained from the inverse of the Fisher information matrix) and the actual variance obtained from the given method.
5. For any vector $\mathbf{a} = [a_1, a_2, a_3]$, there exists a unique skew-symmetric matrix

$$\hat{a} = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}. \tag{38}$$

   The operator $\hat{a}$ performs the vector product on $\mathbf{R}^3$: $\hat{a}X = \boldsymbol{a} \times X, \forall X \in \mathbf{R}^3$.

   With an abuse of notation, the same variable is used for the random variable and its realization.
6. As pointed out in Cho et al. (1997), 200 bootstrap samples and more than 20 measurements suffice to produce a good estimate.
7. For some points with relatively smooth texture, the variance is small, which is counter-intuitive. However, on close scrutiny, it becomes clear that these regions have better illumination.

References

Azarbayejani, A. and Pentland, A. 1995. Recursive estimation of motion, structure, and focal length. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17:562–575.

Benveniste, A., Metivier, M., and Priouret, P. 1987. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag.

Black, M. and Rangarajan, A. 1996. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19:57–91.

Broida, T. 1985. *Estimating the Kinematics and Structure of a Moving Object from a Sequence of Images*. Ph.D. Thesis.

Broida, T., Chandrashekhar, S., and Chellappa, R. 1990. Recursive estimation of 3-D kinematics and structure from a noisy monocular image sequence. *IEEE Trans. on Aerospace and Electronic Systems*, 26:639–656.

Broida, T. and Chellappa, R. 1989. 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.

Broida, T. and Chellappa, R. 1991. Estimating the kinematics and structure of a rigid object from a sequence of monocular images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13:497–513.

Cho, K., Meer, P., and Cabrera, J. 1997. Performance assessment through bootstrap. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:1185–1198.

Cover, T. and Thomas, J. 1991. *Elements of Information Theory*. John Wiley and Sons.

Daniilidis, K. and Nagel, H. 1990. Analytic results on error sensitivity of motion estimation from two views. *Image and Vision Computing*, 8(4):297–303.

Daniilidis, K. and Nagel, H. 1993. The coupling of rotation and translation in motion estimation of planar surfaces. In *Conference on Computer Vision and Pattern Recognition*, pp. 188–193.

Daniilidis, K. and Spetsakis, M. 1993. Understanding noise sensitivity in structure from motion. In *VisNav93*.

Efron, B. and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Chapman and Hall.

Faugeras, O. 1993. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press.

Fermuller, C. and Aloimonos, Y. 2001. Statistics explains geometrical optical illusions. In *Foundations of Image Understanding*, Chap. 14.

Fessler, J. 1996. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Transactions on Image Processing*, 5:493–506.

Fua, P. 2000. Regularized bundle-adjustment to model heads from image sequences without calibration data. *International Journal of Computer Vision*, 38(2):153–171.

Gennery, D. 1992. Visual tracking of known three-dimensional objects. *International Journal of Computer Vision*, 7(3):243–270.

Golub, G. and Van Loan, C. 1989. *Matrix Computations*. Johns Hopkins University Press.

Goodman, I., Mahler, R., and Nguyen, H. 1997. *Mathematics of Data Fusion*. Kluwer Academic Publishers.

Haralick, R. 1996. Covariance propagation in computer vision. In *ECCV Workshop on Performance Characteristics of Vision Algorithms*.

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

Kanatani, K. 1993. Unbiased estimation and statistical analysis of 3-D rigid motion from two views. *Pattern Analysis and Machine Intelligence*, 15(1):37–50.

Kanatani, K. 1996. *Statistical Optimization for Geometric Computation: Theory and Practice*. North-Holland.

Ljung, L. and Soderstrom, T. 1987. *Theory and Practice of Recursive Identification*. MIT Press.

Longuet-Higgins, H. 1981. A computer algorithm for reconstructing a scenes from two projections. *Nature*, 293:133–135.

Ma, Y., Kosecka, J., and Sastry, S. 2000. Linear differential algorithm for motion recovery: A geometric approach. *International Journal of Computer Vision*, 36:71–89.

Meer, P., Mintz, D., and Rosenfeld, A. 1992. Analysis of the least median of squares estimator for computer vision applications. In *Conference on Computer Vision and Pattern Recognition*, pp. 621–623.

Morris, D., Kanatani, K., and Kanade, T. 2000. 3D model accuracy and gauge fixing. Technical Report, Carnegie-Mellon University, Pittsburgh.

Nalwa, V. 1993. *A Guided Tour of Computer Vision*. Addison Wesley.

Oliensis, J. 1999. A multi-frame structure-from-motion algorithm under perspective projection. *International Journal of Computer Vision*, 34:1–30.

Oliensis, J. 2000. A critique of structure from motion algorithms. Technical Report http://www.neci.nj.nec.com/homepages/oliensis/, NECI.

Oliensis, J. and Genc, Y. 2001. Fast and accurate algorithms for projective multi-image structure from motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):546–559.

Papoulis, A. 1991. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.

Poor, H. 1988. *An Introduction to Signal Detection and Estimation*. Springer-Verlag.

Robbins, H. and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Rousseeuw, P. 1984. Least median of square regression. *Journal of the American Statistical Association*, 79:871–880.

Rousseeuw, P. and Leroy, A. 1987. *Robust Regression and Outlier Detection*. John Wiley and Sons.

Roy Chowdhury, A. 2002. *Statistical Analysis of 3D Modeling From Monocular Video Streams*. Ph.D. Thesis, University of Maryland, College Park.

Roy Chowdhury, A. and Chellappa, R. 2002. Towards a criterion for evaluating the quality of 3D reconstructions. In *International Conference on Acoustics, Speech and Signal Processing*.

Roy Chowdhury, A. and Chellappa, R. 2003a. Face reconstruction from monocular video using uncertainty analysis and a generic model. *Accepted to Computer Vision and Image Understanding*.

Roy Chowdhury, A. and Chellappa, R. 2003b. Statistical error propagation in 3D modeling from monocular video. In *CVPR Workshop on Statistical Analysis in Computer Vision*.

Saridis, G. December 1974. Stochastic approximation methods for identification and control—A survey. *IEEE Trans. on Automatic Control*, 19.

Shan, Y., Liu, Z., and Zhang, Z. 2001. Model-based bundle adjustment with application to face modeling. In *International Conference on Computer Vision*. pp. 644–651.

Shao, J. 1998. *Mathematical Statistics*. Springer-Verlag.

Soatto, S. and Brockett, R. 1998. Optimal structure from motion: Local ambiguities and global estimates. In *Conference on Computer Vision and Pattern Recognition*, pp. 282–288.

Spall, J. 2000. Preprint of *Introduction to Stochastic Search and Optimization*. Wiley.

Srinivasan, S. 2000. Extracting structure from optical flow using fast error search technique. *International Journal of Computer Vision*, 37:203–230.

Sun, Z., Ramesh, V., and Tekalp, A. 2001. Error characterization of the factorization method. *Computer Vision and Image Understanding*, 82:110–137.

Szeliski, R. and Kang, S. 1994. Recovering 3D shape and motion from image streams using non-linear least squares. *Journal of Visual Computation and Image Representation*, 5:10–28.

Thomas, J. and Oliensis, J. 1999. Dealing with noise in multiframe structure from motion. *Computer Vision and Image Understanding*, 76:109–124.

Tomasi, C. and Kanade, T. 1992. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9:137–154.

Triggs, B., Zisserman, A., and Szeliski, R. 2000. *Vision Algorithms: Theory and Practice*. Springer.

Tsai, R. and Huang, T. 1981. Estimating 3-D motion parameters of a rigid planar patch: I. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 29:1147–1152.

Walter, R. 1976. *Principles of Mathematical Analysis*, 3rd Edition. McGraw-Hill.

Weng, J., Ahuja, N., and Huang, T. 1993. Optimal motion and structure estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:864–884.

Weng, J., Huang, T., and Ahuja, N. 1987. 3-D motion estimation, understanding, and prediction from noisy image sequences. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9:370–389.

Weng, J., Huang, T., and Ahuja, N. 1989. Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Trans. on Pattern Analysis and Machine*

Intelligence, 11(5):451–476.

Young, G. and Chellappa, R. 1990. 3-D motion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness results. *Pattern Analysis and Machine Intelligence*:12(8):735–759.

Young, G. and Chellappa, R. 1992. 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.

Zhang, Z. 1998. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27:161–195.

Zhang, Z. and Faugeras, O. 1992. *3D Dynamic Scene Analysis*. Springer-Verlag.