An Information Theoretic Criterion for Evaluating the Quality of 3-D Reconstructions From Video

Amit K. Roy-Chowdhury and Rama Chellappa

***Abstract*—Even though numerous algorithms exist for estimating the three-dimensional (3-D) structure of a scene from its video, the solutions obtained are often of unacceptable quality. To overcome some of the deficiencies, many application systems rely on processing more data than necessary, thus raising the question: how is the accuracy of the solution related to the amount of data processed by the algorithm? Can we automatically recognize situations where the quality of the data is so bad that even a large number of additional observations will not yield the desired solution? Previous efforts to answer this question have used statistical measures like second order moments. They are useful if the estimate of the structure is unbiased and the higher order statistical effects are negligible, which is often not the case. This paper introduces an alternative information-theoretic criterion for evaluating the quality of a 3-D reconstruction. The accuracy of the reconstruction is judged by considering the change in mutual information (MI) (termed as the incremental MI) between a scene and its reconstructions. An example of 3-D reconstruction from a video sequence using optical flow equations and known noise distribution is considered and it is shown how the MI can be computed from first principles. We present simulations on both synthetic and real data to demonstrate the effectiveness of the proposed criterion.**

***Index Terms*—Entropy, error analysis, information theory, mutual information (MI), structure from motion.**

I. Introduction

RECONSTRUCTING a three-dimensional (3-D) model of a scene from a video sequence is an important problem for applications in multimedia, recognition, medical imaging etc. There are different methods for estimating the 3-D structure of a scene from both still and moving images [1]. One of the well known strategies of reconstructing a scene from a video sequence is the structure from motion (SfM) algorithm, which works by computing the motion between corresponding points in an image sequence and then estimating the 3-D structure and the motion of the camera. The accuracy of SfM solutions is limited by various factors which can be broadly classified as inherent geometric indeterminacies [2], [3] and statistical inaccuracies [4]–[6]. This paper deals with the statistical aspect of the error in the 3-D estimates.

3-D reconstructions using SfM obtained from a sequence of images are often of unacceptable quality. The main reason for this is the poor quality of input images and lack of robustness in reconstruction algorithms to deal with this issue [7], [8]. Therefore, many application systems process more images than necessary, hoping to minimize the effect of the errors because of the redundancy in the processed input data. For such cases, in order to obtain an optimal 3-D reconstruction system, it is important to understand how the quality of the 3-D estimates is affected by the number of images processed. Is it possible to obtain a quantitative measure of the quality as a function of the number of images and to recognize situations where the input data is so poor that it is not possible to obtain a 3-D estimate of the desired fidelity?

This is the precise question this paper addresses. We pose the SfM problem in the classical information theoretic framework and propose a cost function for quality evaluation based on computing the mutual information (MI) between the scene structure and its estimates. We track the change in MI, which we term as the incremental MI (IMI), with increasing number of input images. The underlying idea is the following: as more images are considered, the change in the MI between the estimate obtained from these images and the scene structure decreases. The method does not depend on any particular algorithm, though the estimation of the IMI can be optimized for a particular method. We propose methods for estimating the MI using statistical sampling techniques. Using an example of reconstructing a scene from video using optical flow [8], [9] and Gaussian noise distribution, we show how the IMI can be computed from first principles in terms of the input parameters.

The paper is organized as follows. We start with an overview of error analysis methods in SfM and a brief survey of the use of information theory in computer vision. We then provide a motivation for the use of an information theoretic criterion. Section IV provides a formal problem description. Section V introduces the IMI criterion and analyzes some of its properties and its applicability to our problem. We also show how it can be computed in the most general setting. In Section VI, we consider an example of reconstructing a 3-D scene with video corrupted by Gaussian noise and derive the IMI from first principles. Finally, in Section VII, we provide detailed experimental results using both simulated and real data.

II. Related Work

We will briefly survey existing work in the two areas which this paper deals with, namely the error analysis for 3-D recon-

Manuscript received August 20, 2002; revised November 5, 2003. This work was supported in part by the National Science Foundation ITR under Grant 0086075. The work was completed when the author was a graduate student in the Electrical and Computer Engineering Department and the Center for Automation Research, University of Maryland, College Park, MD 20742. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tamas Sziranyi.

A. K. Roy-Chowdhury is with the University of California, Riverside, CA 92521 USA (e-mail: rama@cfar.umd.edu).

R. Chellappa is with the Department of Electrical and Computer Engineering and the Center for Automation Research, University of Maryland, College Park, MD 20742-3275 USA (e-mail: rama@cfar.umd.edu).

Digital Object Identifier 10.1109/TIP.2004.827240

structions and the role of information theoretic concepts in video processing.

A. Error Analysis for 3-D Reconstruction

Many researchers have analyzed the sensitivity and robustness of many of the existing algorithms. The work of Weng *et al.* [10] 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 [11]. Young and Chellappa derived bounds on the estimation error for structure and motion parameters from two images under perspective projection as well as from a sequence of stereo images [5]. Similar results were derived by Daniilidis and Nagel in [12] and the coupling of the translation and rotation for a small field of view was studied. They also proved that many algorithms for 3-D motion estimation, that work by minimizing an objective function, suffer from instabilities, and examined the error sensitivity in terms of translation direction, viewing angle and distance of the moving object from the camera. Zhang's work [8] on determining the uncertainty in estimation of the fundamental matrix is another important contribution in this area. Chiuso *et al.* [13] and Soatto and Brockett [14] have analyzed SfM in order to obtain provably convergent and optimal algorithms. Oliensis emphasized the need to understand algorithm behavior and the characteristics of the natural phenomenon that is being modeled [7]. Ma *et al.* [15] also addressed the issues of sensitivity and robustness in their motion recovery algorithm. Sun *et al.* [16] 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.* [17] analyzed the nontrivial effects of unknown scale factor, referred to in the literature as *gauge* freedom, on the covariance calculations in SfM. In [18] and [19], we showed that it is possible to analytically compute the error covariance of 3-D reconstruction as a function of the error covariance of the optical flow estimates. Using the implicit function theorem [20], we proved that such a result could be derived without strong statistical assumptions. In [21], the authors showed that the statistical bias in the optical flow could be used to explain certain geometrical optical illusions. We have extended their work to prove that the 3-D estimate from SfM using optical flow is also significantly statistically biased [18], [22].

B. Information Theoretic Concepts in Image and Video Processing

Recently, information theoretic concepts have been used in various problems in image processing and computer vision, like image registration [23], object recognition [24], [25] and feature extraction and clustering [26], [27]. In [23], the authors propose a method for aligning two images by maximizing the MI between them and use a stochastic optimization algorithm to perform the optimization. The underlying continuous pdfs (probability distribution functions) were represented using Parzen window densities [28]. In [24], the MI (termed "transinformation") was used to optimally place receptive fields over the object of interest. This was extended to include sequential decision processes in [25]. A slightly different technique using the "average loss of entropy" was used in [29], [30] for viewpoint selection. In the area of feature extraction, an information theoretic approach using Fano's inequality for the error rate in classification was proposed in [26]. Information theory was used in clustering and other pattern recognition problems by Watanbe [31], [32] and a few other authors [27], [33]. In [27], the authors developed a clustering algorithm based on a sample-by-sample estimate of Renyi's entropy [34].

We are not aware of any previous work on the use of information theoretic ideas for the quality evaluation of 3-D reconstruction algorithms from video. The closest reference we can draw to our work is the geometric information criterion (GIC) of Kanatani [35], which deals with model selection for geometric data. We will show later that our criterion, the IMI, for evaluating the quality of 3-D reconstructions is related to the idea of reducing the uncertainty in the reconstructions, which, in turn, is conceptually related to the MDL principle [36].

III. Motivation for an Information Theoretic Criterion

As is evident from the literature survey above, the statistical quality analysis of 3-D reconstruction algorithms has been studied quite extensively. However, most of the methods have relied on computing the second order statistical moments, like covariance of the estimate. The covariance is a preferred measure because of its relation to the Cramer–Rao lower bound (CRLB), which dictates the minimum variance that an estimator can achieve [37]. If the variance of a sequence of estimates of the 3-D structure tends toward the CRLB, then the estimate is said to be asymptotically efficient. However, computation of the CRLB often assumes that the estimate is unbiased (see [5]). This is because, computing the bias of an estimator is not an easy task. Hence, even though expressions exist for the CRLB of a biased estimator (known as the generalized CRLB), it is rarely used. The other main objection to the use of variance as a measure of quality is that it neglects the effect of higher order statistics. This is often a major approximation because the outliers, which are the source of many problems in SfM, are often not modeled accurately by second order statistics.

Recent work [21], [22] has shown that the depth estimates obtained from SfM algorithms are statistically biased, and the bias is significant. Also, as we have shown in [19], the noise in the SfM estimates is significantly non-Gaussian. Hence we propose an information theoretic criterion which works by estimating the probability distribution function of the concerned physical quantities (i.e., the depth), rather than concentrate on certain moments only. This method does not depend on the particular algorithm used for reconstructing the 3-D scene. The major limitation of an information theoretic criterion is its efficient, robust and accurate estimation. This is because it is often difficult, and computationally expensive, to estimate the probability density functions of the parameters of interest. However, estimation of MI has received some attention among researchers in signal processing and information theory [38]. It is our hope that such information theoretic criteria, as proposed in this paper, will become practically applicable as progress is made on robustly estimating them.

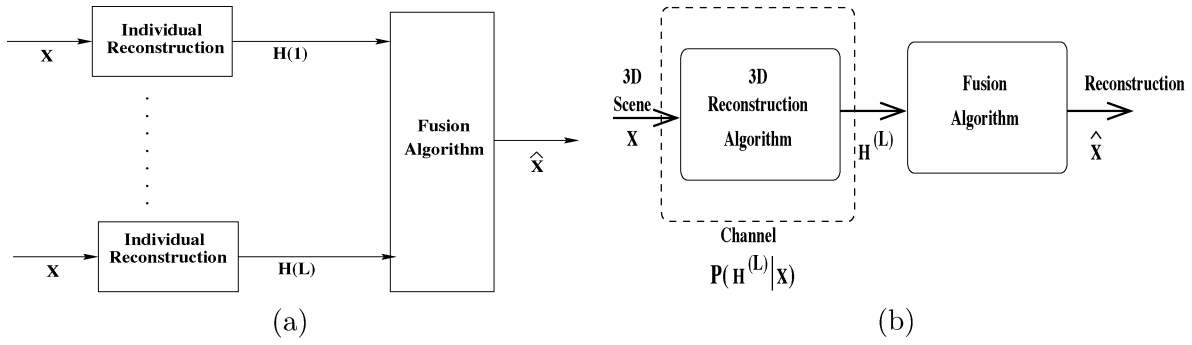


Fig. 1. (a) Block diagram representation of the reconstruction framework. $\mathbf{X}$ is the inverse depth that we want to estimate, $(\mathbf{H}(1), \ldots, \mathbf{H}(L))$ are the intermediate reconstructions (e.g., from pairs of frames), and $\hat{\mathbf{X}}$ is the final fused estimate. (b) A channel model representation of the 3-D reconstruction framework in (a). The channel is characterized by the probability distribution function $P(\mathbf{H}^{(L)}|X)$.

One of the common measures used for computing the accuracy of 3-D reconstruction is the reprojection error, which is the mismatch between an image and the projected model [39]. Since comparison is done in the image domain, it is difficult to infer precisely that the cause of a large reprojection error is the 3-D model. A bad 3-D estimate will lead to large reprojection errors. However, a small reprojection error may be obtained in spite of an inaccurate 3-D model. We provide an alternate way of computing the accuracy of 3-D models. Another point is that reprojection errors are often computed using squared differences between the image and the projection. This leads to undue importance on the second order moments.

IV. Problem Formulation

Theoretically speaking, it is possible to solve for the scene structure and camera motion from two images of the scene [1]. For N corresponding points in two frames, we can write $2N$ equations relating the horizontal and vertical components of the image plane motion for each point with the depth at the point and the camera motion between two frames. The number of unknowns is $N + 5$: the depth at N points, three camera rotation parameters and two camera translation parameters (since we can get only the translation direction because of the scale ambiguity [1]). Thus it is possible to solve for the unknowns from the motion equations in a least squares framework. Details of this can be found in our previous work [19], where we considered the statistical uncertainties in the 3-D estimate. However, the solution obtained from a least-squares procedure is not satisfactory in many practical examples. In this paper, we focus on evaluating the quality of the reconstruction from video sequences with a small baseline.

Since the motion between nearby frames of a video sequence is usually small, the SfM equations based on motion estimates from optical flow [9] is typically valid. However, since the motion is small, even a small amount of error in motion estimates can lead to large errors in structure estimates. This is the classical low signal-to-noise ratio case in signal processing. In our experiments, we have observed that the error can often be as large as (sometimes even greater than) the actual motion between two corresponding points. Hence, in order to obtain accurate solutions to the 3-D structure estimation problems, it is necessary to understand the nature of these errors and their effects. It has been shown by many authors [7], [40] that one of the ways to reduce the effects of these errors is to integrate the estimates over the entire video sequence. In this paper, we try to understand how the quality of the final reconstruction is affected by the number of images in the video sequence in an algorithm-independent manner. We pose the SfM reconstruction problem in an information theoretic framework and use the MI between the unknown scene structure and the 3-D estimate to get a precise idea of the quality of the reconstruction.

A. Notational Convention

Fig. 1 is a block-diagram representation of the 3-D structure estimation algorithm. $\{\mathbf{H}(i), i = 1, \ldots, L\}$ represents the inverse depth[1] from individual reconstructions, which in our case are the structure estimates from pairs of frames from the video sequence (may or may not be adjacent ones). We assume that all the depth values are aligned to a common frame of reference. Feature points will be represented by subscripts, separate reconstructions will be within parenthesis. Thus $H_k(i)$ represents the estimate of the k^{th} feature point for the i^{th} reconstruction.[2] Unless required, the subscript will often be omitted from the notation. The vector of estimates of the inverse depth $[H_k(1), \ldots, H_k(L)]'$ will be denoted by $\mathbf{H}_k^{(L)}$. The boldface notation $\mathbf{H}(i)$ will represent all the features in the i^{th} reconstruction. The final estimate $\hat{\mathbf{X}}$ of $\mathbf{X} = [X_1, \ldots, X_M]'$ is obtained by fusing the individual reconstructions $(\mathbf{H}(1), \ldots, \mathbf{H}(L))$. Our analysis will assume that the feature points are independent and each of them will be treated separately. Hence, we will use the notation $\mathbf{H}^{(L)}$ to denote all the reconstructions for a particular feature point, which we do not represent explicitly. Similarly, X will represent the inverse depth at a particular unspecified point.

[1]The inverse depth is used throughout this paper since it is the quantity that is estimated from the SfM equations for reconstruction from optical flow and its statistics can be obtained in an analytic form more easily than for the depth.

[2]The reconstruction can be obtained from any two pairs of frames, which are not necessarily adjacent, as long as the assumptions of optical flow computation are not violated.

B. System Model

We assume that the individual estimates are corrupted by additive noise, i.e.

$$H(i) = X + V(i) \tag{1}$$

where X is the inverse depth value of the particular feature. A more abstract representation of Fig. 1(a) is shown in Fig. 1(b), where the 3-D reconstruction strategy is represented in a channel model. The input to the channel is the unknown 3-D scene in the form of a video sequence. The output is the sequence of the inverse depths of the scene (aligned to a particular frame of reference) represented as $\mathbf{H}^{(L)}$. The channel is a conceptual representation of the 3-D reconstruction strategy comprising of the video sequence, the correspondence algorithm and the two-frame SfM algorithm. It is characterized by the probability distribution function $P(\mathbf{H}^{(L)}|X)$, which is assumed to be known or can be estimated. If the components of $\mathbf{H}^{(L)}$ are statistically independent, $P(\mathbf{H}^{(L)}|X) = \prod_{i=1}^{L} P(H(i)|X)$. In a later section, we will show how the channel characteristic can be estimated in terms of known parameters of the input video sequence.

The fusion algorithm is treated as a post-processing stage, separate from the channel. From Fig. 1(b), it is clear that X, $\mathbf{H}^{(L)}$ and $\hat{X}$ form a Markov chain, i.e., $X \rightarrow \mathbf{H}^{(L)} \rightarrow \hat{X}$. Representing by $I(X,Y)$, the MI between two random variables X and Y, we can use the data processing inequality [41] and obtain

$$I(X,\hat{X}) \leq I\left(X,\mathbf{H}^{(L)}\right). \tag{2}$$

This allows us to use the MI between the unknown scene structure and its intermediate estimates as a criterion for evaluating the reconstruction quality, since we are assured that the MI of the final reconstruction with the actual scene depth will always be lower or equal. An efficient fusion algorithm should be such that $I(X,\hat{X})$ is as close as possible to $I(X,\mathbf{H}^{(L)})$.

V. Incremental Mutual Information

Consider the channel model representation of the reconstruction strategy in Fig. 1(b) and the data processing inequality of (2). A typical representation of the MI $I(X,\hat{X})$ and $I(X,\mathbf{H}^{(L)})$ is shown in Fig. 2, which is a diagrammatic representation of the data processing inequality as a function of the number of frames, n.

The data processing inequality allows us to evaluate the quality of reconstruction even before the final estimate, $\hat{X}$, has been obtained. This enables us to understand the effect of intermediate reconstructions and the fusion strategy separately. Since our evaluation criterion is based on $I(X,\mathbf{H}^{(L)})$, we can decide whether considering more images from the video sequence will add to the quality of the final reconstruction. Thus, it is possible to monitor the progress of a multi-frame 3-D reconstruction algorithm as it processes more and more video frames.

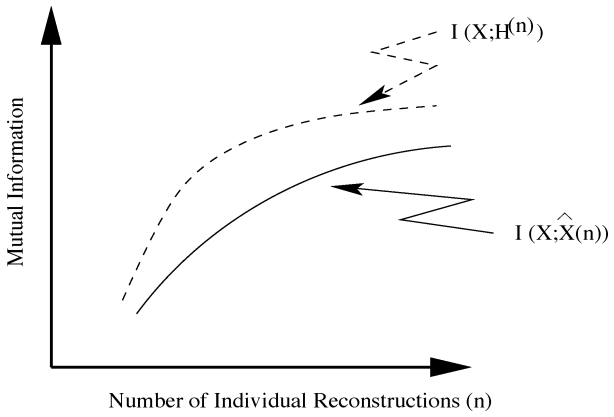


Fig. 2. Typical plot of the MI in the data processing inequality of (2).

Our criterion for evaluating the quality of reconstruction depends on estimating the difference in MI for the two sets of observations, $\mathbf{H}^{(L)}$ and $\mathbf{H}^{(L-1)}$. We term this as the IMI, i.e.

$$\Delta I(L) = I\left(X,\mathbf{H}^{(L)}\right) - I\left(X,\mathbf{H}^{(L-1)}\right). \tag{3}$$

The term gives us an idea of the contribution of the L^{th} observation to the reconstruction strategy with respect to the previous $(L-1)$ observations. As the number of observations increase, the effect of an additional observation decreases and approaches zero in the limit. In order to be assured that the reconstruction quality is actually improving, we need to consider only those situations where the MI $I(X,\mathbf{H}^{(L)})$ is nondecreasing. This ensures that we remove cases where the reconstruction is actually getting worse, and further observations are not improving it any more.

Using the relationship between MI and entropy, it is possible to obtain a different interpretation of the IMI. Denoting by $h(X)$ the entropy of the random variable X, we know that [41] $I(X;Y) = h(X) + h(Y) - h(X,Y)$. Thus $\Delta I(L)$ in (3) can be written as

$$\begin{aligned}\Delta I(L) &= I\left(X;\mathbf{H}^{(L)}\right) - I\left(X;\mathbf{H}^{(L-1)}\right)\\ &= h\left(X|\mathbf{H}^{(L-1)}\right) - h\left(X|\mathbf{H}^{(L)}\right).\end{aligned} \tag{4}$$

The quantity defined as the IMI can also be referred to as the incremental conditional entropy. Since entropy of a random variable is a measure of its uncertainty, ΔI measures the reduction in the uncertainty as we add an extra observation. Since the IMI tends to zero in the limit, the difference in the conditional entropy also approaches zero. Thus we will consider more and more images from the video sequence till the uncertainty in the final structure estimate can be reduced no further. This is the intuitive idea behind our criterion in (3).

The rate at which the IMI decreases is also an important measure of the progress of the algorithm. An extremely slow rate of fall indicates that more images will be necessary to achieve an acceptable level of quality. Since there is motion between adjacent frames of the video, a particular point will move out of the field of view of the camera after a certain amount of time. A very slow rate of fall of ΔI might mean that the quality of

the reconstruction is not good enough even when the point is no longer visible. The rate of change of ΔI can be obtained as

$$\begin{aligned}\Delta^2 I(L) &= \Delta I(L) - \Delta I(L-1) \\ &= I\left(X, \mathbf{H}^{(L)}\right) + I\left(X, \mathbf{H}^{(L-2)}\right) \\ &\quad - 2I\left(X, \mathbf{H}^{(L-1)}\right). \end{aligned} \tag{5}$$

Combining (3) and (5), we can state that an acceptable reconstruction quality has been achieved when $I(X, \mathbf{H}^{(L)})$ is nondecreasing **and** the following conditions are satisfied simultaneously

$$\begin{aligned}\Delta^2 I(L) &\leq 0, \quad \forall L > L_0 \\ \Delta I(L) &< \tau \end{aligned} \tag{6}$$

where L_0 is a constant and τ is a threshold defining an acceptable quality of reconstruction. Since $\Delta I(L)$ is monotone nonincreasing for $L > L_0$ and is bounded below by zero, the monotone convergence theorem [20] applied to (4) implies that $h(X|\mathbf{H}^{(L-1)}) \rightarrow h(X|\mathbf{H}^{(L)}) \rightarrow h_0$ for some $L > L_0$. Thus, h_0 is the minimum level of uncertainty in a scene described by L observations.

Since the criterion does not depend on how the intermediate reconstructions are obtained, it is, in principle, independent of the 3-D reconstruction strategy. However, the procedure for estimation of IMI may be optimized for a particular algorithm.

A. Estimating the MI

We now turn our attention to estimating the IMI from the data. This requires a knowledge of the probability density functions of the random variables, which we do not know a priori and have to estimate from samples. The entropy of a random variable z, with pdf p, can be expressed as

$$h(z) = E_z\left[-\log p(z)\right]. \tag{7}$$

Thus, if we can estimate the probability densities, we can obtain the IMI using (4).

We assume that the channel characteristic, $P(\mathbf{H}^{(L)}|X)$, is known. Using the observation model of (1) and assuming that the noise process $\{V(i)\}_{i=1}^{L}$ is independent of X, we can write

$$P\left(\mathbf{H}^{(L)}|X\right) = P\left(\{X + V(i)\}_{i=1}^{L}|X\right) = P\left(\mathbf{V}^{(L)}\right). \tag{8}$$

Thus, knowledge of the channel characteristic implies that we know the joint distribution of the noise process. If $\{V(i)\}$ is an independent sequence of random variables, the joint distribution is simply the product of the noise distributions in the individual reconstructions. In the next section, we show by an example how the channel characteristic can be estimated from first principles starting with the basic equations of SfM from optical flow. Alternatively, the noise process can be assumed stationary and the probability distribution estimated from the initial few frames using histogram techniques. A method of estimating the probability distributions and MI using statistical sampling techniques can be found in [38].

Once $P(\mathbf{H}^{(L)}|X)$ is known, we can obtain

$$\begin{aligned}P\left(\mathbf{H}^{(L)}\right) &= \int_{\mathcal{X}} P\left(\mathbf{H}^{(L)}|X\right) p_X(x)dx \\ &= \sum_{x_i \in \mathcal{X}} P\left(\mathbf{H}^{(L)}|x_i\right) p_X(x_i) \end{aligned} \tag{9}$$

where $p_X(x_i)$ is the probability that the random variable $X = x_i$. Knowing $p_X(x)$ implies that we have an a priori statistical model on the scene structure X.

Expressing the MI in terms of the entropies, we can write

$$I\left(X, \mathbf{H}^{(L)}\right) = h\left(\mathbf{H}^{(L)}\right) - h\left(\mathbf{H}^{(L)}|X\right). \tag{10}$$

Using $P(\mathbf{H}^{(L)}|X)$ and $P(\mathbf{H}^{(L)})$, we can compute (10) by estimating the entropies using the law of large numbers [42]. The expected value of a random variable $f(Z)$ (in this case it is the entropy function) can be computed by sampling z_i from the distribution $P(z)$ and computing

$$E_Z\left[f(Z)\right] = \frac{1}{n}\sum_{i=1}^{n} f(z_i). \tag{11}$$

This can be used to compute the entropies from (7).

VI. A Case Study: Reconstructing in the Presence of Gaussian Noise

In this section, we consider the special case of Gaussian noise. As explained in Section III, the bias in the structure estimates is one of the reasons for using the IMI under Gaussian noise assumptions. For this particular case we can derive a closed form expression for the IMI, as opposed to the Monte Carlo simulations necessary for the general case, which is dealt with in Sections V-A and VII-A-3. Since the motion between nearby frames in a video sequence is usually small, we will adopt the optical flow framework for reconstructing the structure [1].

Consider a coordinate frame O-XYZ attached rigidly to a camera with the origin at the center of perspective projection and the Z-axis perpendicular to the image plane o-xy. Assume that the camera is in motion with respect to the rigid body imaged scene with translational velocity $\mathbf{V} = [v_X, v_Y, v_Z]$ and rotational velocity $\mathbf{\Omega} = [\omega_X, \omega_Y, \omega_Z]$. It is assumed that the coordinate frame is attached rigidly to the camera with the origin at the center of perspective projection and the z-axis perpendicular to the image plane. Using the small-motion approximation to the perspective projection model for motion field analysis, and denoting by $p(x, y)$ and $q(x, y)$, the horizontal and vertical velocity fields of a point (x, y) in the image plane, we can write the equations relating the object motion and scene depth as [1]

$$\begin{aligned}p(x,y) &= (x - fx_f)h(x,y) + \frac{1}{f}xy\omega_X \\ &\quad - \left(f + \frac{1}{f}x^2\right)\omega_Y + y\omega_Z \\ q(x,y) &= (y - fy_f)h(x,y) + \left(f + \frac{1}{f}y^2\right)\omega_X \\ &\quad - \frac{1}{f}xy\omega_Y - x\omega_Z \end{aligned} \tag{12}$$

where f is the focal length of the camera, $(x_f, y_f) = ((v_X/v_Z), (v_y/v_z))$ is known as the *focus of expansion* (FOE), and $h(x,y) = (v_Z/z(x,y))$ is the scaled inverse scene depth. We will assume that the FOE is known over a few frames of the video sequence. Under the assumption that the motion between adjacent frames in a video is small, we compute the FOE from the first two or three frames and then keep it constant over the next few frames [43]. For N corresponding points, using subscript k to represent the above defined quantities at the k^{th} point and scaling all linear dimensions with respect to the focal length, we define (similar to [43])

$$\begin{aligned}
\mathbf{h} &= [h_1, h_2, \ldots, h_N]^T_{N\times 1}\\
\mathbf{u} &= [p_1, q_1, p_2, q_2, \ldots, p_N, q_N]^T_{2N\times 1}\\
\mathbf{r}_i &= \left[x_i y_i, -\left(1+x_i^2\right), y_i\right]^T_{3\times 1}\\
\mathbf{s}_i &= \left[1+y_i^2, -x_i y_i, -x_i\right]^T_{3\times 1}\\
\mathbf{\Omega} &= [w_X, w_Y, w_Z]^T_{3\times 1}\\
\mathbf{S} &= [r_1 \quad s_1 \quad r_2 \quad s_2 \quad \ldots \quad r_N \quad s_N]^T_{2N\times 3}\\
\mathbf{P} &= \begin{bmatrix} x_1 - x_f & 0 & \cdots & 0\\ y_1 - y_f & 0 & \cdots & 0\\ 0 & x_2 - x_f & \cdots & 0\\ 0 & y_2 - y_f & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & x_N - x_f\\ 0 & 0 & \cdots & y_N - y_f \end{bmatrix}_{2N\times N}\\
\mathbf{A} &= [\mathbf{P} \quad \mathbf{S}]_{2N\times(N+3)}\\
\mathbf{z} &= \begin{bmatrix}\mathbf{h}\\ \mathbf{\Omega}\end{bmatrix}_{(N+3)\times 1}.
\end{aligned} \tag{13}$$

Then (12) can be written as

$$\mathbf{A}\mathbf{z} = \mathbf{u}. \tag{14}$$

Our aims are to compute $\mathbf{z}$ from $\mathbf{u}$ and to obtain a quantitative idea of the accuracy of the 3-D reconstruction $\mathbf{z}$ as a function of the uncertainty in the motion estimates $\mathbf{u}$. Let us denote by $\mathbf{R_u}$ the covariance matrix of $\mathbf{u}$ and by C the cost function that minimizes the reprojection error (i.e., bundle adjustment)

$$\begin{aligned}
C &= \sum_{i=1}^{N}\left[(p_i - \hat{p}_i)^2 + (q_i - \hat{q}_i)^2\right] = \frac{1}{2}\|\mathbf{A}\mathbf{z} - \mathbf{u}\|^2\\
&= \frac{1}{2}\sum_{k=1}^{n=2N}\left(u_k - \sum_{l=1}^{N+3} a_{kl} z_l\right)^2 = \frac{1}{2}\sum_{k=1}^{n=2N} C_k^2(u_k, \mathbf{z})
\end{aligned} \tag{15}$$

where u_k, a_{kl} and z_l represent elements of $\mathbf{u}$, $\mathbf{A}$ and $\mathbf{z}$, respectively, and $(\hat{p}_i, \hat{q}_i)$ are the projections of the depth and motion estimates, $\mathbf{z}$, onto the image plane and are obtained from the right hand side of the (12). In [18], [19], using the implicit function theorem (Chapter 9 of [20]), we proved the following result.

Theorem 1: Define

$$\begin{aligned}
\mathbf{A}_{\bar{k}p} &= [0 \quad \cdots \quad 0 \quad -(x_{\bar{k}} - x_f) \quad 0 \quad \cdots \quad 0\\
&\qquad -x_{\bar{k}} y_{\bar{k}} \quad \left(1 + x_{\bar{k}}^2\right) \quad -y_{\bar{k}}],\\
&= [-(x_{\bar{k}} - x_f)\mathbf{I}_{\bar{k}}(N) | -\mathbf{r}_{\bar{k}}] = [\mathbf{A}_{\bar{k}ph} | \mathbf{A}_{\bar{k}pm}]\\
\mathbf{A}_{\bar{k}q} &= [0 \quad \cdots \quad 0 \quad -(y_{\bar{k}} - y_f) \quad 0 \quad \cdots \quad 0\\
&\qquad -\left(1 + y_{\bar{k}}^2\right) \quad x_{\bar{k}} y_{\bar{k}}(N) \quad x_{\bar{k}}]\\
&= [-(y_{\bar{k}} - y_f)\mathbf{I}_{\bar{k}}(N) | -\mathbf{s}_{\bar{k}}] = [\mathbf{A}_{\bar{k}qh} | \mathbf{A}_{\bar{k}qm}]
\end{aligned} \tag{16}$$

where $\bar{k} = \lceil k/2 \rceil$ is the ceiling of k ($\bar{k}$ 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 n^{th} position of the array of length N and zeros elsewhere. The subscript p in $\mathbf{A}_{\bar{k}p}$ and q in $\mathbf{A}_{\bar{k}q}$ denotes that the elements of the respective vectors are derived from the p^{th} and q^{th} components of the motion in (12). Then

$$\mathbf{R_z} = \mathbf{J}^{-1}\left(\sum_k \frac{\partial C_k^T}{\partial \mathbf{z}}\frac{\partial C_k}{\partial \mathbf{u}}\mathbf{R_u}\frac{\partial C_k^T}{\partial \mathbf{u}}\frac{\partial C_k}{\partial \mathbf{z}}\right)\mathbf{J}^{-T} \tag{17}$$

$$= \mathbf{J}^{-1}\left(\sum_{\bar{k}=1}^{N}\left(\mathbf{A}_{\bar{k}p}^T\mathbf{A}_{\bar{k}p}R_{u\bar{k}p} + \mathbf{A}_{\bar{k}q}^T\mathbf{A}_{\bar{k}q}R_{u\bar{k}q}\right)\right)\mathbf{J}^{-T} \tag{18}$$

and

$$\mathbf{J} = \sum_{\bar{i}=1}^{N}\left(\mathbf{A}_{\bar{k}p}{}^T\mathbf{A}_{\bar{k}p} + \mathbf{A}_{\bar{k}q}{}^T\mathbf{A}_{\bar{k}q}\right) \tag{19}$$

where $\mathbf{R_u} = \text{diag}[R_{u1p}, R_{u1q}, \ldots, R_{uNp}, R_{uNq}]$.

This theorem gives an expression for the covariance of the inverse depth estimate as a function of the covariance of the noise in the two-dimensional (2-D) image-plane motion estimates. Because of the partitioning of $\mathbf{z}$ in (13), we can write

$$\mathbf{R_z} = \begin{bmatrix}\mathbf{R_h} & \mathbf{R_{h\Omega}}\\ \mathbf{R_{h\Omega}^T} & \mathbf{R_\Omega}\end{bmatrix} \tag{20}$$

where $\mathbf{R_h}$ and $\mathbf{R_\Omega}$ are the covariances of $\mathbf{h}$ and $\mathbf{\Omega}$ and $\mathbf{R_{h\Omega}}$ is the cross-covariance between $\mathbf{h}$ and $\mathbf{\Omega}$.

Recall our previous formulation in (1), where X was the unknown true inverse depth of a particular point. Assume that $X \sim \mathcal{N}(0, \sigma_x^2 = Q_X)$, i.e., the mean of X is subtracted out. Also, $\{V(i), i = 1, \ldots, L\}$ is a sequence of independent random variables distributed as $\mathcal{N}(0, \sigma_{V(i)}^2)$, representing the noise distribution in the L two-frame inverse depth estimates. Let $\mathbf{Q}_V = \text{diag}[Q_V(i)]_{i=1,\ldots,L} = \text{diag}[\sigma_{V(1)}^2, \ldots, \sigma_{V(L)}^2]$.[3] Since $\mathbf{R_h}$ in (20) is the error covariance of the two-frame depth estimate, we get $\text{Cov}[H(i)|X] = \mathbf{R_h}(i)$. Thus $\mathbf{Q}_V = \text{diag}[\mathbf{R_h}(1), \ldots, \mathbf{R_h}(L)]$, where $\mathbf{R_h}(i)$ is the value of $\mathbf{R_h}$ at a particular point for the inverse depth obtained from the i and $(i+1)^{\text{st}}$ frames.

From (1), $E[H(i)] = 0$ and

$$E[H(i)H(j)] = E[(X + V(i))(X + V(j))] = Q_X + Q_V(i)\delta_{ij} \tag{21}$$

[3]By $\text{diag}[a_1, \ldots, a_N]$ or $\text{diag}[a_i]_{i=1,\ldots,N}$, we mean a diagonal matrix of size $N \times N$.

where δ_{ij} is a Kronecker delta function. Thus the covariance of $\mathbf{H}^{(L)}$ is $\mathbf{Q}_{\mathbf{H}^{(L)}} = \mathbf{Q}_{V^{(L)}} + \mathbf{1}_L Q_X \mathbf{1}_L^T$, where $\mathbf{1}_L$ is a vector of L ones. Using the fact that the entropy (differential) of a Gaussian random variable $Z \sim \mathcal{N}(0, \Sigma)$ is $(1/2)\log(2\pi e|\Sigma|)$ [41] ($|.|$ denotes the determinant), the MI between X and $H(i)$

$$I(X; H(i)) = h(H(i)) - h(H(i)|X) = \frac{1}{2}\log\left(1 + \frac{Q_X}{Q_V(i)}\right). \tag{22}$$

Next, consider the MI between the unknown X and the vector of observations $\mathbf{H}^{(L)}$. We will denote by $|K|$ the determinant of a matrix K

$$\begin{aligned} I\left(X; \mathbf{H}^{(L)}\right) &= h\left(\mathbf{H}^{(L)}\right) - h\left(\mathbf{H}^{(L)}|X\right) \\ &\overset{(a)}{=} h\left(\mathbf{H}^{(L)}\right) - \sum_{i=1}^{L} \frac{1}{2}\log\left(2\pi e Q_V(i)\right) \\ &\overset{(b)}{=} \frac{1}{2}\log\left(\frac{\left|\mathbf{Q}_V + \mathbf{1}_L Q_X \mathbf{1}_L^T\right|}{|\mathbf{Q}_V|}\right) \end{aligned} \tag{23}$$

where (a) is a result of applying the chain rule of entropy and substituting the expression for the differential entropy of a Gaussian random variable; (b) is due to the fact that $|Q_V| = \prod_{i=1}^{L} Q_V(i) = \prod_{i=1}^{L} \sigma_{V(i)}^2$. Using the method of induction and the properties of determinants, it can be shown that $|\mathbf{Q}_V + \mathbf{1}_L Q_X \mathbf{1}_L^T| = \prod_{i=1}^{L} \sigma_{V(i)}^2 + \sigma_x^2 \sum_{i=1}^{L} \prod_{\substack{j=1 \\ j\neq i}}^{L} \sigma_{V(j)}^2$ (see Appendix A). Then from (23), the expression for the MI becomes

$$I\left(X; \mathbf{H}^{(L)}\right) = \frac{1}{2}\log\left(1 + \sum_{i=1}^{L} \frac{\sigma_X^2}{\sigma_{V(i)}^2}\right). \tag{24}$$

Thus, the IMI $\Delta I(L)$ is

$$\begin{aligned} \Delta I(L) &= I\left(X; \mathbf{H}^{(L)}\right) - I\left(X; \mathbf{H}^{(L-1)}\right) \\ &= \frac{1}{2}\log\left(\frac{\left|\mathbf{Q}_{V^{(L)}} + \mathbf{1}_L Q_X \mathbf{1}_L^T\right|}{\left|\mathbf{Q}_{V^{(L-1)}} + \mathbf{1}_{L-1} Q_X \mathbf{1}_{L-1}^T\right|} \cdot \frac{|\mathbf{Q}_{V^{(L-1)}}|}{|\mathbf{Q}_{V^{(L)}}|}\right) \\ &= \frac{1}{2}\log\left(\frac{\prod_{i=1}^{L} \sigma_{V(i)}^2 + \sigma_x^2 \sum_{i=1}^{L} \prod_{\substack{j=1 \\ j\neq i}}^{L} \sigma_{V(j)}^2}{\prod_{i=1}^{L} \sigma_{V(i)}^2 + \sigma_x^2 \sum_{i=1}^{L-1} \prod_{\substack{j=1 \\ j\neq i}}^{L} \sigma_{V(j)}^2}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{\frac{1}{\sigma_{V(L)}^2}}{\frac{1}{\sigma_x^2} + \sum_{i=1}^{L-1} \frac{1}{\sigma_{V(i)}^2}}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{\frac{1}{Q_V(L)}}{\frac{1}{\sigma_x^2} + \sum_{i=1}^{L-1} \frac{1}{Q_V(i)}}\right). \end{aligned} \tag{25}$$

Hence, we are able to obtain a closed form expression for $\Delta I(L)$ in terms of the parameters of the input video sequence by starting from the basic equations of 3-D reconstruction from optical flow.

A. Estimation Theoretic Interpretation

Since we have considered the case of Gaussian noise, it is possible to give an alternative interpretation to the results in (25) from an estimation theoretic perspective. The mean squared distortion for M feature points is defined as

$$D(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{M}\sum_{k=1}^{M} E\left[(X_k - \hat{X}_k)^2\right]. \tag{26}$$

Let $p(X_k, H_k(1), \ldots, H_k(L))$ denote the joint density function of the parameter and observations. The mean square error estimator $\hat{X}_j$ of X_j, obtained from $\mathbf{H}^{(L)}$, is $\hat{X}_j(L) = E[X_j|H_j^{(L)}]$. Using the definition of CRLB we can write the following set of inequalities:

$$\begin{aligned} D &\geq \frac{1}{M}\sum_{k=1}^{M} \frac{1}{E\left[-\frac{\partial^2}{\partial X^2}\log p(X_k, H_k(1), \ldots, H_k(L))\right]} \\ &= \frac{1}{M}\sum_{k=1}^{M} \frac{1}{\frac{1}{\sigma_{x_j}^2} + \sum_{i=1}^{N} E\left[-\frac{\partial^2}{\partial X^2}\log p(H_k(i)|X)\right]} \\ &\geq \frac{1}{\frac{1}{M}\sum_{k=1}^{M}\left(\frac{1}{\sigma_{x_j}^2} + \sum_{k=1}^{N} \frac{1}{Q_{V_k(i)}}\right)} \\ &\triangleq \frac{1}{\frac{1}{M}\sum_{k=1}^{M} \frac{1}{D_k(L)}}. \end{aligned} \tag{27}$$

The last step is a result of the application of Jensen's inequality [37] and that $E[-(\partial^2/\partial X^2)\log p(H_k(i)|X)] = (1/Q_{V_k}(i))$. Recalling that (25) is for a particular feature point where the subscript has been suppressed for clarity of notation, let us denote $\Delta I_k \triangleq I(X_k; \mathbf{H}_k^{(L)}) - I(X_k; \mathbf{H}_k^{(L-1)})$. Then from (27) and the last expression of (25), we get

$$\Delta I_k = \frac{1}{2}\log\left(\frac{D_k(L-1)}{D_k(L)}\right). \tag{28}$$

Alternatively, using a Kalman filter with observation model as in (1), a constant parameter system model (i.e., $\mathrm{X(i+1)} = \mathrm{X(i)}$), and the initial condition that $\mathrm{Cov}[X(0)] = \sigma_x^2$, the innovations at the L^{th} stage, $\gamma_L = X_L - \hat{X}_L$. Then following the standard derivation for the Kalman filter (actually a recursive least squares problem because of the constant parameter) [37], it can be shown that variance of the innovations

$$P_{\gamma_L} = \sigma_{V(L)}^2\left(1 + \frac{\frac{1}{\sigma_{V(L)}^2}}{\frac{1}{\sigma_x^2} + \sum_{i=1}^{N-1} \frac{1}{\sigma_{V(i)}^2}}\right) \tag{29}$$

which shows that, for each feature point, the IMI is related to P_{γ_L} as

$$\Delta I = \frac{1}{2}\log\left(\frac{P_{\gamma_L}}{\sigma_{V(L)}^2}\right). \tag{30}$$

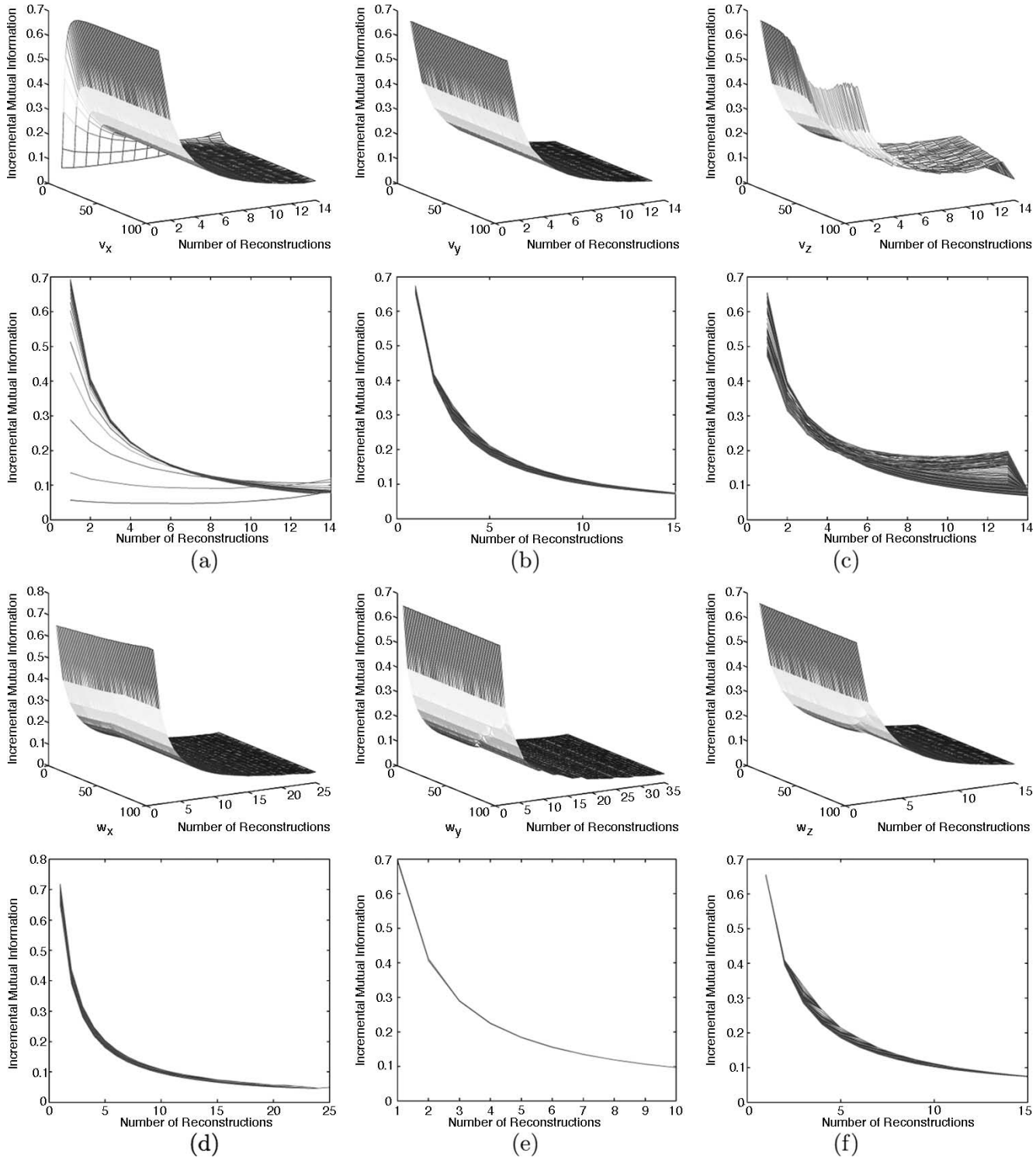


Fig. 3. Plots of the variation of IMI with different camera motion parameters. The 3-D plots represents the variation of the IMI with the number of intermediate reconstructions and the motion parameter, while the 2-D plots represents the variation of the IMI with the number of intermediate reconstructions for each value of the camera motion parameter. Each of the camera motion parameters are varied over a certain range, keeping the others fixed at their nominal value described in the text. The different plots are obtained for each of the six camera motion parameters whose range is as follows: (a) $v_X \in [0.001, 0.1]$; (b) $v_Y \in [0.001, 0.1]$; (c) $v_Z \in [0.001, 0.1]$; (d) $\omega_X \in (0, 3]$ degrees/frame; (e)$\omega_Y \in (0, 3]$ degrees/frame; and (f) $\omega_Z \in (0, 3]$ degrees/frame.

VII. Experimental Results

In this section, we present the results of experiments carried out in order to analyze the criterion based on IMI using both simulated and real data. With the simulation data, we analyze the dependencies of the IMI on different camera motion parameters. We also show how the IMI can be used to identify a few troublesome points which affect the quality of the entire reconstruction. Next, we consider real 3-D range data and analyze the reconstruction using the IMI criterion

for quality adjustment. Finally, a video sequence as captured by an ordinary video camera is considered and the effect of our criterion on the 3-D reconstruction algorithm is studied. In all the experiments, we consider the situation where the MI $I(X, \mathbf{H}^{(L)})$ is nondecreasing, which is the interesting case as explained in Section V. The fusion algorithm is described in [19] and uses a least median of squares estimator, which is solved by using the Robbins-Monro stochastic approximation algorithm.

A. Experiments With Simulated Data

The aim of the following experiments with simulation data is to analyze the effect of different camera motion parameters on the IMI criterion, given that the true depth values are known. Also, we will analyze the effect of different levels of noise in the feature positions. For this purpose, a set of fifty 3-D points were generated so that their true positions are known. The initial positions of these points were set randomly. Different kinds of motion were applied to these points so as to obtain various motion tracks for each of them. The perspective projections of these points were generated on a 512×512 pixel grid and Gaussian noise with zero mean and known variance was added to these 2-D locations. This resulted in creating a set of tracked features. The median value of the true motion between two consecutive frames (median computed over all frames and features) was around 15 pixels in both the horizontal and vertical directions. Pairs of such motion tracks, corresponding to pairs of frames of an image sequence, are given as the input to the 3-D reconstruction algorithm. In order to take advantage of a longer baseline between the images, the pairs of frames are chosen by skipping two intermediate frames between them, e.g. frames 1 and 4, 2, and 5, etc. to form pairs. We solve for the depth using the true value of the camera motion. This is a simple least-squares solution of (12), which ensures that the errors in the depth computation are solely due to noise in the feature positions (i.e., without noise in the feature positions, the reconstruction will be perfect). This allows us to separate and study the effects of noise in feature positions on the depth estimate. In practice, errors in camera motion estimates and algorithmic imperfections will also affect the solution.

1) Variation With Motion Parameters: In this set of experiments, we vary the different camera motion parameters one at a time and analyze the change in the IMI. For each kind of motion, the image projections are obtained, which are the input to the 3-D reconstruction algorithm. Since the 3-D reconstruction algorithm assumes that the motion between pairs of frames is small, we have to be careful in choosing the range over which the camera motion parameters can be allowed to vary. The range for each motion parameter was sampled uniformly at 100 points. For this experiment, a small Gaussian noise with a variance of 5 pixels was added to each of the feature positions independently. The IMI was computed using (25). The number of frames considered for each reconstruction varied a little, depending upon the length of time that the feature positions remained within the field of view of the camera, for that kind of motion. In Fig. 3, we plot the IMI averaged over all the feature positions, as each of the six camera motion parameters are varied. The 3-D plots represent the variation of the IMI with the number of intermediate reconstructions and for each value of the particular camera motion parameter that is changed. The change with the camera motion parameter can be seen along the y-axis. In the 2-D plot, separate curves are obtained for each value of the motion parameter of interest. Thus the 2-D plots are the different y-z cross-sections of the 3-D plots in the first column. The nominal values of the different parameters are set at $v_x = v_y = v_z = 0.01$ and $\omega_X = \omega_Y = \omega_Z = 1$ degree/frame. The exact units of the translational motion do not matter because the equations involve the FOE $x_f = (v_x/v_z)$ and $y_f = (v_Y/v_Z)$. We vary each parameter at a time, keeping all the other fixed at the nominal value.

The interesting fact to note from the curves in Fig. 3 is that the IMI does not change much with the rotational motion. This is to be expected because, under perspective projection and the set-up of (12), the rotational motion does not affect the depth computation. Therefore as long as the rotational motion is not so large as to introduce errors in the image-plane motion computation, the quality of the reconstruction should not change too much with the rotation. The IMI should, however, change with the translational motion, and this is shown in the respective plots.

2) Variation With Noise in Feature Positions: In the above experiments, we assumed that the noise in the feature positions was small and studied the variation of the IMI with different camera motion parameters. However, in practice, the noise will vary depending upon a number of external parameters, like lighting conditions, imaging system, etc. Hence, we now study how noise in feature positions will affect the IMI estimates of the quality of the reconstruction. We use the same set of 3-D points used in the above experiment and generate a set of motion tracks for all the points using the nominal values of the camera motion parameters. For this experiment, the motion parameters are kept fixed, while the noise level in the feature positions is varied. Fig. 4(a), (b) plots the reconstructed depth and the IMI, when Gaussian noise with standard deviations of 2 and 5 pixels, respectively, is added to the positions of the feature points. A similar plot with noise of standard deviation of 5 pixels is shown in Fig. 4(c), but the range of depth values is larger, as is clear from the vertical axis. Since the motion between consecutive frames is smaller due to the larger depth, we considered the motion between every fifth frame in the sequence. The IMI is plotted for every single feature in the lower figure of each of the pairs of plots. For the case of low noise (standard deviation of two pixels), the IMI decreases continuously for every feature point. However, for higher noise cases, the IMI behaves in a much more complicated manner. It decreases monotonically for some of the features, while for the others this is not the case. The monotonic decrease corresponds to those points which are reconstructed accurately. The points for which the IMI behaves erratically are the ones at which the reconstruction is not very accurate.

3) IMI Estimated From Monte Carlo Simulations: In this section, we show how to estimate the IMI with an unknown (non-Gaussian) noise distribution, using Monte Carlo sampling to approximate the distribution, as explained in Section V-A. As in the previous experiments, a set of features were tracked over a number of frames, with the camera motion set at the nominal value. However, noise was added to the feature positions according to a uniform distribution. For each feature point, the

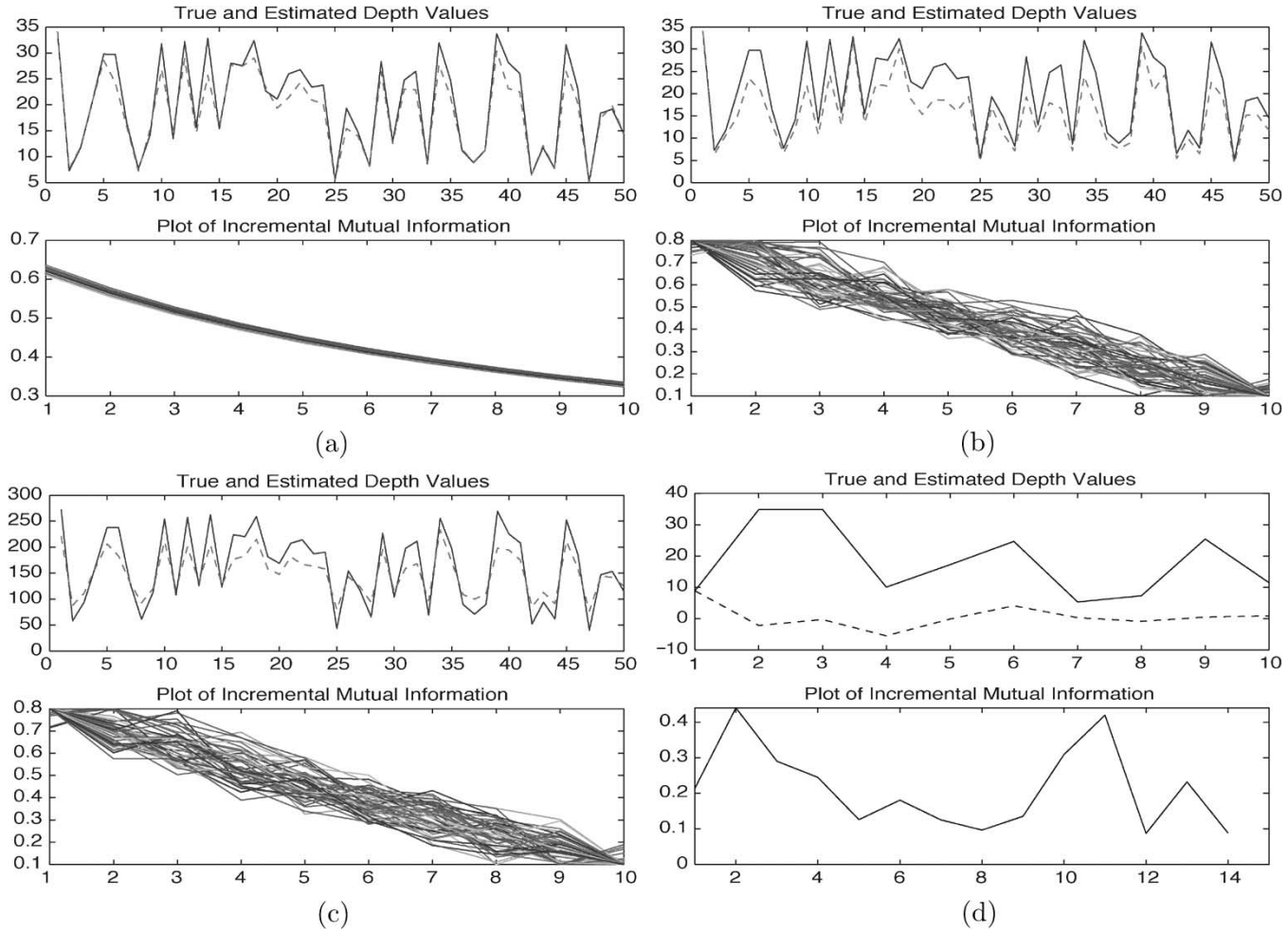


Fig. 4. (a)–(b) Plots of the depth reconstruction for the 50 feature points along with the IMI for each feature, for different levels of noise in the feature positions. The upper figure in each pair of plots represents the depth reconstruction for the feature points, while the lower figure represents the IMI as a function of the number of intermediate reconstructions, for each feature point separately. For the depth reconstruction, the solid line represents the true depth, and the dotted line represents the reconstructed value. The standard deviation of the noise in the three plots is as follows: (a) Two pixels and (b) five pixels. (c) Same plot as above, except that the range of the depth values is larger. The standard deviation of the noise is five pixels. (d) Uniform noise example: the upper plot shows the true depth values of the 3-D points (the solid line) and the fused estimate from the intermediate reconstructions from all the ten frames (the dotted lines). The lower plot shows the change in the IMI, obtained using Monte Carlo simulations, with increasing number of intermediate reconstructions.

noise was added uniformly in a 5×5 block around that point. In this case, we cannot use the closed form expressions that we could derive for the Gaussian noise case. Hence, the technique for estimating the IMI, as explained in Section V-A, was followed. The noise in the feature points was assumed to be independent. The points were tracked over 50 frames, out of which the first thirty were used for estimating the distribution of the two-frame depth estimates using (8). The IMI was computed by following the steps described in the above-mentioned section. The 3-D positions of the points were estimated from the SfM equations in (12), assuming that the camera motion is known and using the least-squares method described before. For some of the points, the results were erroneous as is clear from the first plot of Fig. 4(d). The lower plot of the same figure depicts this case where the IMI remains large and does not follow the steadily decreasing trend as in some of the previous examples. This is an example of a situation where the reconstruction quality is not of the desired fidelity. However, the reprojection error, computed from (15), decreases monotonically, with increasing number of intermediate reconstructions.

B. Experiments With Range and Video Data

We now show how our IMI criterion can be used in real-life 3-D reconstruction problems. For this purpose, we consider two experiments. The first involves the situation where 3-D range data of a face is available. A video sequence is generated from this range data. We estimate the structure of the face using the video sequence of it, use the IMI criterion for assessing the quality of the reconstruction, and finally compare this depth estimate with the true value. In the second experiment, we use a video sequence of a face captured with a ordinary video camera and plot the IMI criterion and the 3-D reconstruction of it.

1) Experiments With Face Range Data: We experimented with a publicly available database of 3-D models obtained from a Minolta 700 range scanner.[4] . We will report numerical results from our algorithm on some of the data available here, though we will not publish the images or 3-D models of the subjects. In order to perform an accurate analysis of our methods, we require

[4]The data is available [Online] at: http://sampl.eng.ohio-state.edu/sampl/data/3-DDB/RID/minolta/faces-hands.1299/index.html

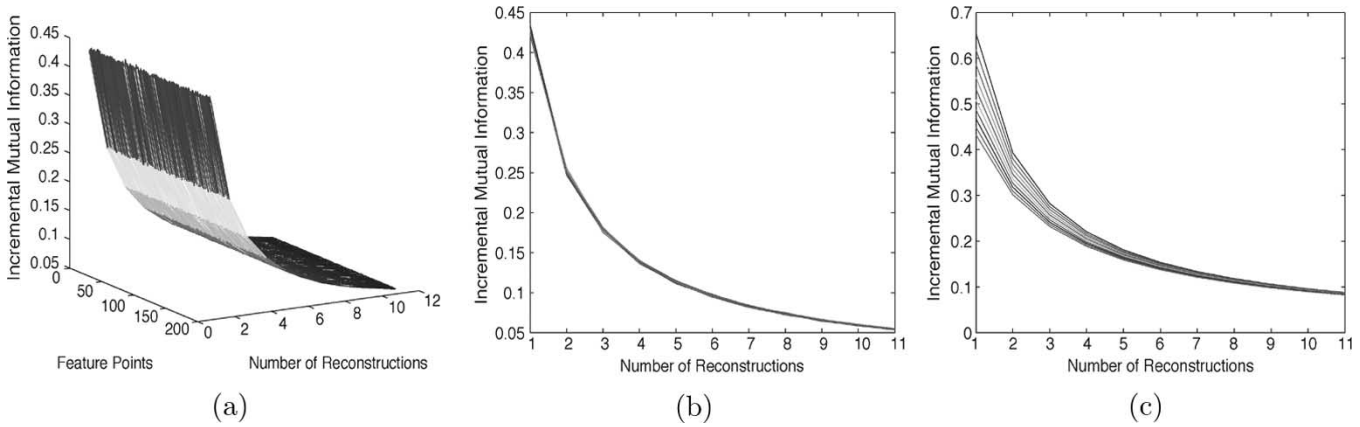


Fig. 5. (a) Plot of the IMI versus number of intermediate reconstructions and feature points. (b) Plot of IMI versus number of intermediate reconstructions for each feature point. (c) Plot of IMI versus number of intermediate reconstructions for each feature point, where Gaussian noise with standard deviation of five pixels is added to the feature positions.

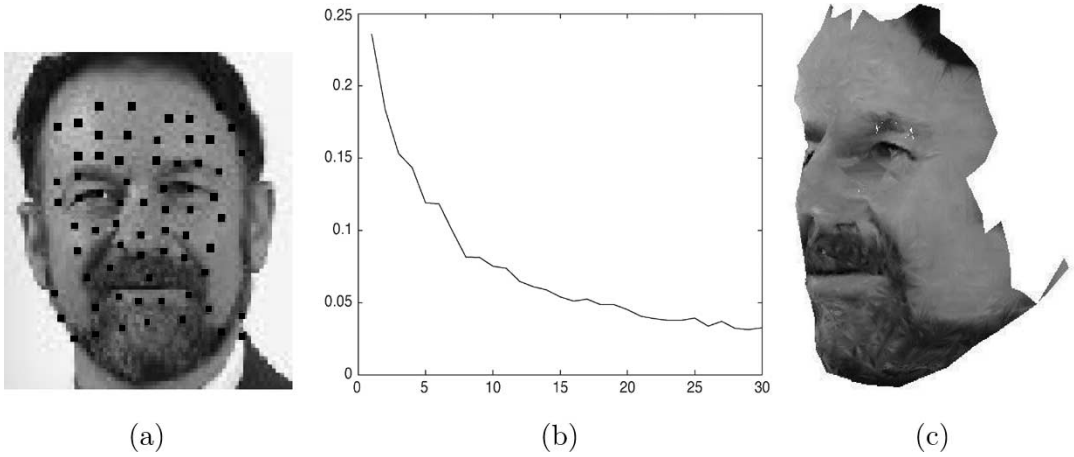


Fig. 6. Three-dimensional reconstruction from video using the method of measuring the IMI to judge the quality of the result. (a) One of the images from the video along with the set of tracked features used for the reconstruction. (b) The change in the IMI with the number of images. (c) One view from the reconstructed model.

a video sequence of the person and the 3-D depth values. This, however, is not available on this particular database or in any other that we know of. Thus we had to generate a sequence of images in order to apply our algorithm.[5] This was done using the 3-D model and the texture map provided on the web-site.

The 3-D reconstruction of the face was generated using our algorithm described in [44] and [45]. For this paper, we use the example referred to as "frame001" on this website. The average error in the reconstruction was about 3%. The IMI was computed using Theorem 1 and (25). The plot of the IMI for this example is shown in Fig. 5. In Fig. 5(a), the variation of the IMI with the number of i, the variation of the IMI with the number of intermediate reconstructions and the feature points is shown in a 3-D plot. In Fig. 5(b), a set of 2-D plots of the IMI versus the number of reconstructions, for each feature point is shown. In Fig. 5(c), we plot the IMI (for each feature point) when Gaussian noise with a standard deviation of 5 pixels is added to the feature positions. In this case, the average reconstruction error was around 4% and the IMI varied significantly depending on the feature point. Such variations can be used to identify points which can potentially lead to erroneous reconstructions.

2) Experiments With Video Data: Finally, we present a result on a video sequence captured with an ordinary video camera. The video consists of a person moving his head in front of a static camera. The aim was to reconstruct the model of the head of the person from this video. The focal length of the camera was known. Fig. 6(a) represents an image from the video along with some of the feature points which were tracked. Fig. 6(b) represents the change in the IMI between the unknown 3-D structure and the intermediate reconstructions from every pair of frames. The covariance of the error in the intermediate reconstructions were estimated using the result of Theorem 1. This was used to estimate the IMI using (25). Based on this measure, the 3-D model was reconstructed using 25 frames and Fig. 6(c) shows one particular view of this model. Details of the 3-D reconstruction algorithm can be found in [45].

[5]The optical flow computed with the generated sequences may be more accurate than in a normal setting. However, we will also analyze after adding noise to the feature positions. Hence the comparison of the 3-D reconstruction accuracy should still be useful.

VIII. Conclusion

In this paper, we have introduced a method to evaluate the quality of 3-D reconstruction from a video sequence in information theoretic terms. We showed that the 3-D reconstruction problem can be represented using a channel model, where the channel characteristic can be estimated from the input parameters of the video. Such a conceptual representation allows us to derive a criterion for evaluating the reconstruction by computing the change in the MI between the unknown scene structure and the 3-D estimates obtained from increasing numbers of images from the video sequence. This information theoretic criterion is useful because it can deal with biased estimators and can capture the effects of higher-order statistics. We showed how the IMI can be estimated in terms of the parameters of the feature points tracked across the video sequence. Finally, we presented results using both simulated and real data.

Appendix
Computing the Error Covariances

We will derive an expression for the determinant of the matrix $\mathbf{Q}_{V^{(N)}} + \mathbf{1}_N Q_X \mathbf{1}_N^T$ in (23). Represent by $\mathbf{A}_N$ the following matrix:

$$\mathbf{A}_N = \begin{bmatrix} \sigma_x^2+\sigma_{v_1}^2 & \sigma_x^2 & \cdots & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2+\sigma_{v_2}^2 & \cdots & \sigma_x^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2+\sigma_{v_N}^2 \end{bmatrix}. \quad (31)$$

The aim is to compute the determinant of $\mathbf{A}_N$. We will do so by using the method of induction. Consider $N = 2$. Then the determinant of $\mathbf{A}_2$ denoted by $|\mathbf{A}_2| = \sigma_x^2(\sigma_{v_1}^2+\sigma_{v_2}^2)+\sigma_{v_1^2}\sigma_{v_2}^2$. For $N = 3$, $|\mathbf{A}_3| = \sigma_x^2(\sigma_{v_1}^2\sigma_{v_2}^2+\sigma_{v_2}^2\sigma_{v_3}^2+\sigma_{v_3}^2\sigma_{v_1}^2)+\sigma_{v_1}^2\sigma_{v_2}^2\sigma_{v_3}^2$. Assume that

$$|\mathbf{A}_N| = \prod_{i=1}^{N}\sigma_{v_i}^2 + \sum_{i=1}^{N}\sigma_x^2\prod_{\substack{j=1\\ j\neq i}}^{N}\sigma_{v_j}^2. \quad (32)$$

Now consider the matrix

$$\mathbf{A}_{N+1} = \begin{bmatrix} \sigma_x^2+\sigma_{v_1}^2 & \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2+\sigma_{v_2}^2 & \cdots & \sigma_x^2 & \sigma_x^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2+\sigma_{v_N}^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2+\sigma_{v_{N+1}}^2 \end{bmatrix}. \quad (33)$$

Subtracting the second to last row from the last one, we get the following matrix

$$\begin{bmatrix} \sigma_x^2+\sigma_{v_1}^2 & \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2+\sigma_{v_2}^2 & \cdots & \sigma_x^2 & \sigma_x^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2+\sigma_{v_N}^2 & \sigma_x^2 \\ 0 & 0 & \cdots & -\sigma_{v_N}^2 & \sigma_{v_{N+1}}^2 \end{bmatrix}. \quad (34)$$

Then

$$|\mathbf{A}_{N+1}| = \sigma_{v_{N+1}}^2|\mathbf{A}_N| + \sigma_{v_N}^2|\mathbf{B}| \quad (35)$$

where

$$\mathbf{B} = \begin{bmatrix} \sigma_x^2+\sigma_{v_1}^2 & \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2+\sigma_{v_2}^2 & \cdots & \sigma_x^2 & \sigma_x^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_x^2 & \sigma_x^2 & \cdots & \sigma_x^2+\sigma_{v_{N-1}}^2 & \sigma_x^2 \\ \sigma_x^2 & \cdots & \sigma_x^2 & \sigma_x^2 & \sigma_x^2 \end{bmatrix}. \quad (36)$$

Now, using the fact that the determinant of a matrix remains unchanged by elementary row and column operations, we get

$$\begin{aligned} |\mathbf{B}| &= \begin{vmatrix} \sigma_x^2+\sigma_{v_1}^2 & -\sigma_{v_1}^2 & -\sigma_{v_1}^2 & \cdots & -\sigma_{v_1}^2 & -\sigma_{v_1}^2 \\ \sigma_x^2 & \sigma_{v_2}^2 & 0 & \cdots & 0 & 0 \\ \vdots & 0 & \sigma_{v_3}^2 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_x^2 & 0 & 0 & \cdots & \sigma_{v_{N-1}}^2 & 0 \\ \sigma_x^2 & 0 & 0 & \cdots & 0 & 0 \end{vmatrix} \\ &= (-1)^{N+1}\sigma_x^2 \begin{vmatrix} -\sigma_{v_1}^2 & -\sigma_{v_1}^2 & \cdots & -\sigma_{v_1}^2 & -\sigma_{v_1}^2 \\ \sigma_{v_2}^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_{v_3}^2 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{v_{N-1}}^2 & 0 \end{vmatrix} \\ &= \sigma_x^2\sigma_{v_1}^2\sigma_{v_2}^2\cdots\sigma_{v_{N-1}}^2. \end{aligned} \quad (37)$$

Then substituting (32) and (37) into (35), we get

$$\begin{aligned} \mathbf{A}_{N+1} &= \prod_{i=1}^{N+1}\sigma_{v_i}^2 + \sigma_{v_{N+1}}^2\left(\sum_{i=1}^{N}\sigma_x^2\prod_{\substack{j=1\\ j\neq i}}^{N}\sigma_{v_j}^2\right) + \sigma_x^2\prod_{i=1}^{N}\sigma_{v_i}^2 \\ &= \prod_{i=1}^{N+1}\sigma_{v_i}^2 + \sum_{i=1}^{N}\sigma_x^2\prod_{\substack{j=1\\ j\neq i}}^{N+1}\sigma_{v_j}^2 + \sigma_x^2\prod_{i=1}^{N}\sigma_{v_i}^2, \\ &= \prod_{i=1}^{N+1}\sigma_{v_i}^2 + \sum_{i=1}^{N}\sigma_x^2\prod_{\substack{j=1\\ j\neq i}}^{N+1}\sigma_{v_j}^2 \end{aligned} \quad (38)$$

which proves the hypothesis for $|\mathbf{A}_N|$ in (32).

ACKNOWLEDGMENT

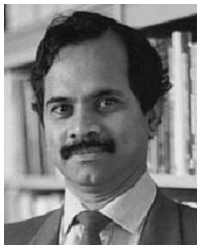
The authors would like to thank Prof. A. Papamarcou for many interesting and helpful suggestions.

REFERENCES

[1] V. Nalwa, *A Guided Tour of Computer Vision*. Reading, MA: Addison Wesley, 1993.
[2] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[3] O. D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA: MIT Press, 1993.
[4] J. Weng, T. S. Huang, and N. Ahuja, "3-D motion estimation, understanding, and prediction from noisy image sequences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 370–389, 1987.
[5] G. S. Young and R. Chellappa, "Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 995–1013, Oct. 1992.
[6] Z. Y. Zhang, "Determining the epipolar geometry and its uncertainty: a review," *Int. J. Comput. Vis.*, vol. 27, pp. 161–195, Mar. 1998.
[7] J. Oliensis, "A critique of structure-from-motion algorithms," *Comput. Vis. Image Understanding*, vol. 84, no. 3, pp. 407–408, Dec. 2001.
[8] Z. Zhang and O. Faugeras, *3-D Dynamic Scene Analysis*. New York: Springer-Verlag, 1992.
[9] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *AI*, vol. 17, pp. 185–203, 1981.
[10] J. Weng, N. Ahuja, and T. S. Huang, "Optimal motion and structure estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 864–884, Sept. 1993.
[11] T. J. Broida and R. Chellappa, "Performance bounds for estimating three-dimensional motion parameters from a sequence of noisy images," *J. Opt. Soc. Amer. A*, vol. 6, pp. 879–889, 1989.
[12] K. Daniilidis and H. H. Nagel, "The coupling of rotation and translation in motion estimation of planar surfaces," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1993, pp. 188–193.
[13] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from motion causally integrated over time," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 523–535, Apr. 2002.
[14] S. Soatto and R. Brockett, "Optimal structure from motion: local ambiguities and global estimates," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1998, pp. 282–288.
[15] Y. Ma, J. Kosecka, and S. Sastry, "Linear differential algorithm for motion recovery: a geometric approach," *Int. J. Comput. Vis.*, vol. 36, pp. 71–89, Jan. 2000.
[16] Z. Sun, V. Ramesh, and A. M. Tekalp, "Error characterization of the factorization method," *Comput. Vis. Image Understanding*, vol. 82, pp. 110–137, May 2001.
[17] D. D. Morris, K. Kanatani, and T. Kanade, "Gauge fixing for accurate 3-D estimation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2001, pp. II:343–II:350.
[18] A. K. Roy-Chowdhury, "Statistical Analysis of 3-D Modeling From Monocular Video Streams," Ph.D. thesis, University of Maryland, College Park, 2002.
[19] A. K. Roy-Chowdhury and R. Chellappa, "Stochastic approximation and rate-distortion analysis for robust structure and motion estimation," *Int. J. Comput. Vis.*, vol. 55, no. 1, pp. 27–53, Oct. 2003.
[20] R. Walter, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.
[21] C. Fermuller and Y. Aloimonos, "Statistics explains geometrical optical illusions," in *Foundations of Image Understanding*, ch. 14, 2001.
[22] A. K. Roy-Chowdhury and R. Chellappa, "Effect of stattistical bias on 3-D reconstruction from video," Univ. of Maryland, College Park, Tech. Rep., CAR-TR-972, 2001.
[23] P. A. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, Sept. 1997.
[24] B. Schiele and J. L. Crowley, "Transinformation for active object recognition," in *Proc. Int. Conf. Computer Vision*, 1998, pp. 249–254.
[25] J. Denzler and C. M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 145–157, Feb. 2002.
[26] J. W. Fisher III and J. C. Principe, "A nonparametric methodology for information theoretic feature extraction," in *Proc. DARPA'97*, 1997, pp. 1077–1084.
[27] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 158–171, Feb. 2002.
[28] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[29] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Appearance-based active object recognition," *Image Vis. Comput.*, vol. 18, no. 9, pp. 715–727, June 2000.
[30] L. Paletta, M. Prantl, and A. Pinz, "Learning temporal context in active object recognition using bayesian analysis," in *Proc. ICPR00*, vol. I, 2000, pp. 695–699.
[31] S. Watanabe, *Pattern Recognition: Human and Mechanical*. New York: Wiley, 1985.
[32] ——, "Pattern recognition as a quest for minimum entropy," *Pattern Recognit.*, vol. 14, pp. 381–387, 1981.
[33] T. Hofmann and J. M. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 1–14, Jan. 1997.
[34] A. Renyi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math., Statistics and Probability*, 1960, pp. 547–561.
[35] K. Kanatani, "Geometric information criterion for model selection," *Int. J. Comput. Vis.*, vol. 26, no. 3, pp. 171–189, Feb. 1998.
[36] M. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Statist. Assoc.*, to be published.
[37] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1988.
[38] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1315–1321, May 1999.
[39] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *Proc. Int. Conf. Computer Vision*, 1999, pp. 781–788.
[40] T. J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 90–99, Jan. 1986.
[41] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[42] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1991.
[43] S. Srinivasan, "Extracting structure from optical flow using fast error search technique," *Int. J. Comput. Vis.*, vol. 37, pp. 203–230, 2000.
[44] A. K. Roy-Chowdhury, S. Krishnamurthy, T. Vo, and R. Chellappa, "3-D face reconstruction from video using a generic model," in *Proc. Int. Conf. Multimedia and Expo*, Lausanne, Switzerland, 2002.
[45] A. K. Roy-Chowdhury and R. Chellappa, "3-D face reconstruction from monocular using using uncertainty analysis and a generic model," *Comput. Vis. Image Understanding*, vol. 91, no. 1–2, pp. 188–213, July–Aug. 2003.

Amit K. Roy-Chowdhury received the B.S. degree in electrical Eengineering from Jadavpur University, Calcutta, India, in 1985, the M.S. degree in engineering in systems science and automation from the Indian Institute of Science, Bangalore, in 1997, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Maryland, College Park, in 2002, where he worked on statistical error characterization of 3-D modeling from monocular video sequences.

He is an Assistant Professor in the Electrical Engineering Department, University of California, Riverside. He was previously with the Center for Automation Research, University of Maryland, as a Research Associate. He was involved in projects related to face, gait, and activity modeling and recognition. His research interests are in signal, image and video processing, computer vision, and pattern recognition.



Rama Chellappa received the B.E. (Hons.) degree from the University of Madras, Madras, India, in 1975 and the M.E. (Distinction) degree from the Indian Institute of Science, Bangalore, in 1977. He received the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981, respectively.

Since 1991, he has been a Professor of electrical engineering and an affiliate Professor of Computer Science at the University of Maryland, College Park. He is also affiliated with the Center for Automation Research (Director) and the Institute for Advanced Computer Studies (permanent member). Prior to joining the University of Maryland, he was an Assistant Professor (1981 to 1986) and an Associate Professor (1986 to 1991) and Director of the Signal and Image Processing Institute (1988 to 1990) with the University of Southern California (USC), Los Angeles. Over the last 22 years, he has published numerous book chapters and peer-reviewed journal and conference papers. He has edited a collection of Papers on Digital Image Processing (Los Alamitos, CA: IEEE Computer Society Press, 1992), coauthored a research monograph on *Artificial Neural Networks for Computer Vision* (with Y.T. Zhou) (New York: Springer-Verlag, 1990), and co-edited a book on *Markov Random Fields: Theory and Applications* (with A.K. Jain) (New York: Academic, 1993). His current research interests are face and gait analysis, 3-D modeling from video, automatic target recognition from stationary and moving platforms, surveillance and monitoring, hyperspectral processing, image understanding, and commercial applications of image processing and understanding.

Dr. Chellappa has served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON NEURAL NETWORKS. He was Co-Editor-in-Chief of *Graphical models and Image Processing*. He is now serving as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He served as a member of the IEEE Signal Processing Society Board of Governors from 1996 to 1999. Currently, he is serving as the Vice President of Awards and Membership for the IEEE Signal Processing Society. He has served as a General the Technical Program Chair for several IEEE international and national conferences and workshops. He received several awards, including the National Science Foundation (NSF) Presidential Young Investigator Award, an IBM Faculty Development Award, the 1990 Excellence in Teaching Award from School of Engineering at USC, the 1992 Best Industry Related Paper Award from the International Association of Pattern Recognition (with Q. Zheng), and the 2000 Technical Achievement Award from the IEEE Signal Processing Society. He was elected as a Distinguished Faculty Research Fellow (1996 to 1998) at the University of Maryland, he is a Fellow of the International Association for Pattern Recognition, and he received a Distinguished Scholar-Teacher award from the University of Maryland in 2003.