Deterministic and Statistical Properties of Multi-Resolution 3D Modeling

Amit RoyChowdhury	Haiying Liu	Rama Chellappa
amitrc@cfar.umd.edu	hyliu@cfar.umd.edu	rama@cfar.umd.edu

The authors are with Center for Automation Research, University of Maryland, College Park, MD 20742. This work was partially supported by DARPA/ONR grant N00014-00-1-0908.

Abstract

Multi-resolution schemes for 3D modeling from an input video sequence are becoming very popular. However, multi-resolution techniques may not be the best solution strategy in many scenarios and it is important to understand the characteristics of such algorithms. In this paper, we present a multi-resolution structure from motion algorithm, using monocular video as input, that exploits the bilinear relationship between depth and translation parameters to propagate the estimates through a coarse-to-fine reconstruction framework. We present a detailed analysis of the deterministic and statistical properties of such an algorithm. We show that our optimization procedure is guaranteed to converge to the local minimum at each resolution. We also derive analytical expressions for the error covariances of depth and motion estimates, obtained using a multi-resolution structure from motion reconstruction algorithm, as a function of the error covariance of the feature tracks in the the input video. The method is based on using the implicit function theorem for deriving the error covariance at each resolution, and propagating the statistics from coarse to fine resolution, again taking advantage of the bilinear nature of the problem. The statistical calculations do not require the assumptions of Gaussianity of the error distributions and are valid for dense depth reconstruction estimates. Simulations have been carried out using real-life video sequences. It is shown how multi-resolution techniques can either succeed or fail in reconstructing the same scene depending on the quality of the input video of that scene, thus justifying the need for a theoretical analysis of error propagation in the 3D modeling from video.

Index Terms

Multi-resolution, structure and motion estimation, error analysis, bilinear parameterization, face modeling.

I. Introduction

Extraction of 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 [1], [2] and [3], among others. Recent research on SfM issues has concentrated on sensitivity, robustness and error characterization of existing techniques [4], [5], [6], [7], etc. The errors that 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 [8] or [2]). They can usually be handled by imposing additional constraints on the solution space. The statistical errors are a result of the poor quality of 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.

Multi-resolution techniques (e.g. [9], [10], [11]) have recently become popular in 3D structure recovery from a video sequence. In this paper, we concentrate on multi-resolution SfM techniques from monocular video with a small baseline. We focus on the continuous (differential) version of the SfM equations [12]. We show that the locally optimal solution can be obtained at each resolution of the multi-resolution reconstruction scheme by taking advantage of the bilinear parameterization of the structure and motion equations. Even though multi-resolution techniques have many advantages, they may not always be effective or useful. We will show later with an example (Figures 3, 4), that multi-resolution techniques may or may not succeed in reconstructing the same scene, depending on the quality of the input video of the scene (which will vary due to illumination conditions, imaging sensors, etc.). Thus, it is important to understand the effect of the quality of the input video on the final 3D reconstruction. We show how to estimate the quality of the 3D reconstruction as a function of statistics of the input video. We start with reconstruction at a single resolution and derive a closed-form expression for the error covariance of the reconstruction as a function of the error covariance of the feature tracks in the video sequence. We then show that it is possible to extend the mathematical methods used here to the multi-resolution reconstruction case and derive similar closed form expressions. The derivation is based on the implicit function theorem of real analysis [13], and does not require the assumption of Gaussianity of the error statistics. It has been used previously in vision for the derivation of the uncertainty in the fundamental matrix [1] and for establishing partial results on the uniqueness of the structure and motion parameters when a long sequence is used [14]. [1] We show the effect of such an analysis on real-life 3D reconstruction problems.

II. Related Work

Pioneered by the seminal work of Longuet-Higgins [16] and the eight-point algorithm developed independently by Tsai and Huang [17], 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

[1] We have recently come to know that a somewhat similar method was applied for error calculations in medical imaging applications [15].

the hope that the extra information would help correct the flaws that are inevitably present in two-frame reconstructions. The problem of tracking an object across multiple frames was addressed in [18] where a known object and its past position and velocity were used to predict its new location. Broida and Chellappa investigated the use of the extended Kalman filter [19] for estimating motion and structure from a sequence of monocular images [20]. Azarbayejani and Pentland extended this work to include the estimation of the focal length of the camera, along with motion and structure [21]. Tomasi and Kanade developed an algorithm for shape and motion estimation under orthographic projection using the factorization theorem [22]. Szeliski and Kang proposed a non-linear least squares optimization scheme using the Levinburg-Marquardt method for solving the problem [23]. Oliensis developed a multi-frame algorithm under perspective projection in [24], which was extended recently in [25]. Most of these multi-frame methods can be characterized as batch processing (but not necessarily recursive), which means that the problem of estimating the 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 proposed a fusion algorithm that computes the final reconstruction from intermediate reconstructions by analyzing the uncertainties in them, rather than from image data directly [26].

In spite of the existence of numerous algorithms for SfM [2], [3], [8], constructing accurate 3D models reliably from images is still a challenging problem. Many researchers have analyzed the sensitivity and robustness of many of the existing algorithms. The work of Weng et al. [27] 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 [28]. 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 [29]. Similar results were derived by Daniilidis and Nagel in [30] and the coupling of the translation and rotation for a small field of view was studied. They also proved that many algorithms for three-dimensional 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 and Soatto [31] and Soatto and Brockett [32] 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 [3]. Ma, Kosecka and Sastry [6] also addressed the issues of sensitivity and robustness in their motion recovery algorithm. Sun, Ramesh and Tekalp [7] proposed an error characterization of the factorization method for 3D shape and motion recovery from image sequences using matrix perturbation theory. Morris, Kanatani and Kanade [33] analyzed the non-trivial effects of unknown scale factor, referred to in the literature as *gauge* freedom, on the covariance calculations in SfM. In [34], Roy Chowdhury and Chellappa showed that it is possible to analytically compute the error covariance of 3D reconstruction as a function of the error covariance of the optical flow estimates, using the implicit function theorem [13]. Recently, Fermuller *et. al.* [35] have shown that the bias in optical flow estimates can be used to explain certain geometrical optical illusions. We have extended their work to prove that the 3D estimate from SfM using optical flow is also significantly statistically biased [36], [37].

III. Problem Formulation

Consider a coordinate frame O-XYZ attached rigidly to a camera with the origin at the center of perspective projection and the Z-axis perpendicular to the image plane o-xy. Assume that the camera is in motion with respect to a single rigid body imaged scene with translational velocity $\mathbf{V} = [v_X, v_Y, v_Z]^T$ and rotational velocity $\mathbf{\Omega} = [\omega_X, \omega_Y, \omega_Z]^T$. We assume that the camera motion between two consecutive frames in a video sequence is small, and use optical flow for motion field analysis. If $p(x,y)$ and $q(x,y)$ are the horizontal and vertical velocity fields of a point (x,y) in the image plane, they are related to the 3D object motion and scene depth by [12]

$$p = (x - fx_f)h + \frac{1}{f}xy\omega_X - (f + \frac{1}{f}x^2)\omega_Y + y\omega_Z \qquad (1)$$

$$q = (y - fy_f)h + (f + \frac{1}{f}y^2)\omega_X - \frac{1}{f}xy\omega_Y - x\omega_Z, \qquad (2)$$

where $h(x,y) = v_Z/z(x,y)$ is the scaled inverse scene depth, f is the focal length of the camera, and $(x_f, y_f) = (\frac{v_X}{v_Z}, \frac{v_Y}{v_Z})$ is known as the *focus of expansion* (FOE). For N such points, normalizing linear distances with respect to the focal length and defining [2],

$$\begin{aligned}
\mathbf{h} &= [h_1, h_2, ..., h_N]^T_{N\times 1} && (3)\\
\mathbf{u} &= [p_1, q_1, p_2, q_2, ..., p_N, q_N]^T_{2N\times 1} \triangleq [u_i]^T_{i=1,2,\cdot,2N} && (4)\\
\mathbf{r}_i &= [x_i y_i, -(1+x_i^2), y_i]^T_{3\times 1} && (5)\\
\mathbf{s}_i &= [1+y_i^2, -x_i y_i, -x_i]^T_{3\times 1} && (6)\\
\mathbf{\Omega} &= (\omega_X, \omega_Y, \omega_Z)^T_{3\times 1}, && (7)\\
\mathbf{Q} &= [\mathbf{r}_1,\ \mathbf{s}_1,\ \mathbf{r}_2,\ \mathbf{s}_2,\ ...,\ \mathbf{r}_N,\ \mathbf{s}_N]^T_{2N\times 3}\,, && (8)\\
\mathbf{P} &= \text{diag}\,[x_i - x_f \quad y_i - y_f]^T_{2N\times N, i=1,...,N}\,, && (9)\\
\mathbf{B} &= [\mathbf{P}\ \mathbf{Q}]_{2N\times(N+3)} && (10)\\
\mathbf{z} &= [\mathbf{h} \quad \mathbf{\Omega}]^T_{(N+3)\times 1}\,, && (11)
\end{aligned}$$

it can be shown that

$$\mathbf{u} = \mathbf{P}\mathbf{h} + \mathbf{Q}\mathbf{\Omega} = \begin{bmatrix} \mathbf{P} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{\Omega} \end{bmatrix} \triangleq \mathbf{B}\mathbf{z}. \tag{12}$$

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

Consider the cost function which minimizes the re-projection error (i.e. bundle adjustment)

$$\begin{aligned}
C &= \frac{1}{2}\sum_{i=1}^{2N}(u_i - \hat{u}_i)^2 = \frac{1}{2}\sum_{i=1}^{2N} C_i^2 && (13)\\
C_i &= u_i - \hat{u}_i && (14)
\end{aligned}$$

where $(\hat{p}_i, \hat{q}_i)$ are the projections of the depth and motion estimates, $\mathbf{z}$, onto the image plane and are obtained from the right hand side of the equations (2).

IV. Multi-resolution Algorithm

We now show how to solve the above problem optimally in a multi-resolution framework. *For each pixel* at resolution level l (represented by a superscript), equation (2) can be re-written in a hierarchical way [10].

$$\mathbf{u}^l = h^l \mathbf{G}^l \mathbf{V} + \mathbf{K}^l \mathbf{\Omega} \tag{15}$$

[2] The i^{th} point is represented by the subscript i.

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

where

$$\mathbf{u}^l = [\frac{p}{2^l}, \frac{q}{2^l}]^T, \quad h^l = \frac{1}{\left(\frac{Z}{2^l}\right)}, \quad x^l = \frac{x}{2^l}, \quad y^l = \frac{y}{2^l}, \tag{16}$$

$$\mathbf{G}^l \;=\; \begin{bmatrix} -\frac{f}{2^{2l}} & 0 & \frac{x^l}{2^l} \\ 0 & -\frac{f}{2^{2l}} & \frac{y^l}{2^l} \end{bmatrix}, \tag{17}$$

$$\mathbf{K}^l \;=\; \begin{bmatrix} \frac{2^l x^l y^l}{f} & -\left(\frac{f}{2^l} + \frac{2^l {x^l}^2}{f}\right) & y^l \\ \left(\frac{f}{2^l} + \frac{2^l {y^l}^2}{f}\right) & -\frac{2^l x^l y^l}{f} & -x^l \end{bmatrix}. \tag{18}$$

It can be observed that (15) is a bilinear system. That is, the unknown parameters $\mathbf{\Theta}^l = [\mathbf{m}^T \;\; h^l]^T$ can be naturally split up into motion vector $\mathbf{m} = [\mathbf{V}^T \;\; \mathbf{\Omega}^T]^T$ and structure parameter h^l such that the system is linear in $\mathbf{m}$ for fixed h^l and linear in h^l for fixed $\mathbf{m}$:

- Motion subsystem:

$$\mathbf{S}^l\mathbf{m} = \mathbf{u}^l \tag{19}$$

  where $\mathbf{S}^l = \left[h^l\mathbf{G}^l \;\; \mathbf{K}^l\right]$ and $\mathbf{u}^l$ are known, and the motion $\mathbf{m} = [\mathbf{V}^T \quad \mathbf{\Omega}^T]^T$ is unknown.

- Structure subsystem:

$$\mathbf{v}^l h^l = \mathbf{w}^l \tag{20}$$

  where $\mathbf{v}^l = \mathbf{G}^l\mathbf{V}$ and $\mathbf{w}^l = \mathbf{u}^l - \mathbf{K}^l\mathbf{\Omega}$ are known, and the inverse depth h^l is unknown.

Denote estimated structure by $\hat{h}^l$, estimated motion by $\hat{\mathbf{m}}$, and the corresponding optical flow computed from (15) by $\hat{\mathbf{u}}^l$. The cost function (13) can be rewritten as:

$$C(\mathbf{\Theta}^l) = \frac{1}{N^l}\sum_{i=1}^{N^l} ||\mathbf{u}_i^l - \hat{\mathbf{u}}_i^l||_2^2 \tag{21}$$

where N^l is the total number of pixels at level l. The SfM problem is then formulated as finding both structure and motion so that (21) is minimized. A natural way of solving such a bilinear system is to treat it as a sequence of least-squares problems. First, the structure propagated from the previous coarser level is fixed and the motion is solved using the motion subsystem. Then, the computed motion is fixed and the structure is solved using the structure subsystem. Mathematically,

$$\begin{aligned} \mathbf{m}(j) &= \arg\min_{\mathbf{m}} C\left(\mathbf{m}, h^l(j-1)\right) && (22) \\ h^l(j) &= \arg\min_{h^l} C\left(h^l, \mathbf{m}(j)\right), && (23) \end{aligned}$$

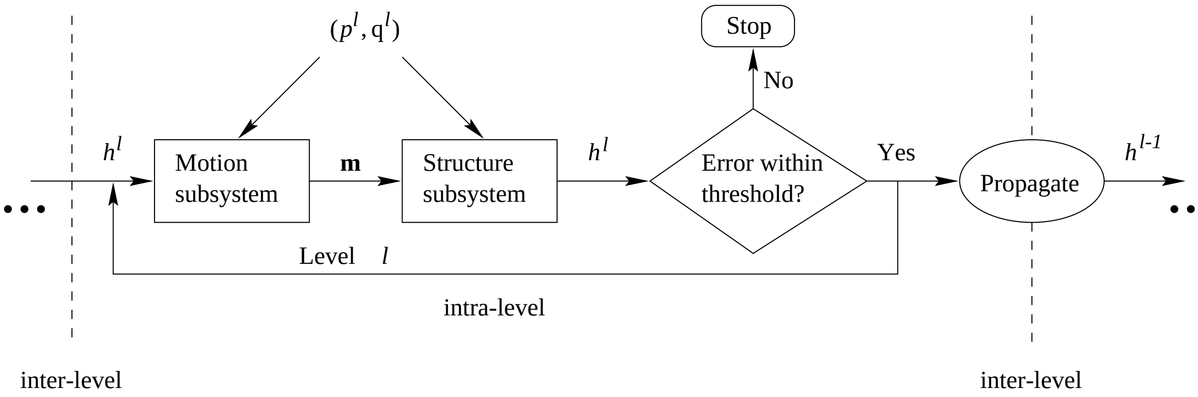

Fig. 1. Hierarchical iterative algorithm.

where j is the iteration index. When the motion at the current level converges, the structure is propagated to the next level. It is illustrated in the appendix that this is indeed a descent method that will converge to a local minimum (see also [38], Chapter 10).

The motivation for formulating the structure from motion problem in a hierarchical way is that (x^l, y^l, Z^l) and (p^l, q^l) define a hierarchical "scene", in which the sizes of all objects and the optical flow are reduced by half when the level l increases by one. Thus in the coarsest level, the structure tends to be flat. This can be exploited to generate an initial guess at each resolution that, hopefully, leads to the global minimum. Our experiments show that compared to single-resolution, the multi-resolution method reduces the error variance in structure estimation.

It is well known that (2) alone only yields structure and translation up to a scale transformation [12]. Therefore, it is convenient to initialize inverse depth h^l at the coarsest level to an arbitrary constant. Then the motion $(\mathbf{V}, \mathbf{\Omega})$ is estimated from (19). The resulting $\mathbf{m} = \begin{bmatrix}\mathbf{V}^T & \mathbf{\Omega}^T\end{bmatrix}^T$ is then passed to (20) and h^l is re-computed. This estimation procedure is repeated at each resolution level (intra-level iteration) until the changes in the motion $(\mathbf{V}, \mathbf{\Omega})$ is below a threshold. h^l is then propagated to the next finer level and refined (inter-level iteration), until h^0 is solved. The algorithm is listed as follows. Its flow chart is shown in Figure 1.

1) Set l = coarsest level.
2) Assume that $h^l(0)$ is flat, i.e. $h^l(0) \leftarrow c$, where c is a positive constant.
3) Set $i \leftarrow 0$.
4) Set $i \leftarrow i + 1$.

5) When $h^l(i-1)$ is known from previous level $l-1$, we have the motion subsystem:

$$\mathbf{S}^l(i-1)\mathbf{m}(i) = \mathbf{u}^l \tag{24}$$

where $\mathbf{S}^l(i-1) = \begin{bmatrix} h^l(i-1) & \mathbf{G}^l & \mathbf{K}^l \end{bmatrix}$ and $\mathbf{m}(i) = \begin{bmatrix} \mathbf{V}(i)^T & \mathbf{\Omega}(i)^T \end{bmatrix}^T$. Solve the linear subsystem using the method of least squares, and obtain the estimate of $\mathbf{m}(i) = (\mathbf{V}(i), \mathbf{\Omega}(i))$ at level l.

6) When $\mathbf{V}(i)$ and $\mathbf{\Omega}(i)$ are computed from the previous step, we have the structure subsystem:

$$\mathbf{v}^l(i)h^l(i) = \mathbf{w}^l(i) \tag{25}$$

where $\mathbf{v}^l(i) = \mathbf{G}^l\mathbf{V}(i)$ and $\mathbf{w}^l = \mathbf{u}^l - \mathbf{K}^l\mathbf{\Omega}(i)$. Solve the linear subsystem using the method of least squares for each pixel, and refine the estimation of $h^l(i)$ at the level l.

7) Go to 4, until $\|\mathbf{m}(i) - \mathbf{m}(i-1)\| < \epsilon$, where ϵ is a small positive number; or a maximum number of iterations is reached if not convergent yet.

8) Set $l \leftarrow l-1$.

9) Propagate the structure $Z^l = \frac{1}{h^l}$ from level $l+1$ by

$$x^l \leftarrow 2x^{l-1}, \qquad y^l \leftarrow 2y^{l-1}, \qquad Z^l \leftarrow 2Z^{l-1} \tag{26}$$

and interpolate Z^l at $(2x^{l-1}+1, 2y^{l-1}+1)$.

10) Go to step 3, until $l = 0$, i.e. the original image level.

11) Set $Z \leftarrow \frac{1}{h^0}$.

V. Error Analysis

We now analyze the statistical characteristics of the above algorithm. Our aim is to derive an analytical expression relating the statistics of depth and motion estimates to a function of the statistics of the feature tracks in the input video sequence. For ease of understanding, we first derive the result for the single resolution case, and these show how it can be extended to the multi-resolution scenario.

A. Single Resolution Error Analysis

We 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 $\mathbf{R_z}$ (see equation (12)), for the single resolution case. The cost function in (13) requires a non-linear optimization, which rarely gives a good solution unless a very good initial condition is available. Different methods have been proposed to deal with this. These involve estimating the camera motion first followed by the depth, recursively updating the camera motion and depth one at a time using the previously available estimate of the other, etc. [2]. For the case of reconstruction from a monocular video that we deal with, it is often possible to estimate the FOE from the first two/three frames and assume it to be constant over the next few which are used to reconstruct the structure. Knowledge of the FOE makes the system of equations in (12) linear, because of the bilinear parameterization in (2). We will derive the error covariance expression for this special case, because it is simple and enough to prove the main result on multi-resolution error analysis. For the corresponding result in the more general scenario, refer to [34].

Theorem 1: Define

$$\begin{aligned}\mathbf{A}_{\bar{i}p} &= [-(x_{\bar{i}} - x_f)\mathbf{I}_{\bar{i}}(N)| - \mathbf{r}_{\bar{i}}] = [\mathbf{A}_{\bar{i}ph}|\mathbf{A}_{\bar{i}pm}] \\ \mathbf{A}_{\bar{i}q} &= [-(y_{\bar{i}} - y_f)\mathbf{I}_{\bar{i}}(N)| - \mathbf{s}_{\bar{i}}] = [\mathbf{A}_{\bar{i}qh}|\mathbf{A}_{\bar{i}qm}] \qquad (27)\end{aligned}$$

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 $i = 1, ..., n = 2N$) 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 $\mathbf{A}_{\bar{i}p}$ and q in $\mathbf{A}_{\bar{i}q}$ denotes that the elements of the respective vectors are derived from the p^{th} and q^{th} components of the motion in (2). Then

$$\begin{aligned}\mathbf{R_z} &= \mathbf{H}^{-1}\left(\sum_{ij} \frac{\partial C_i^T}{\partial \mathbf{z}}\frac{\partial C_i}{\partial \mathbf{u}}\mathbf{R_u}\frac{\partial C_j^T}{\partial \mathbf{u}}\frac{\partial C_j}{\partial \mathbf{z}}\right)\mathbf{H}^{-T} \qquad (28) \\ &= \mathbf{H}^{-1}\left(\sum_{\bar{i}=1}^{N}\left(\mathbf{A}_{\bar{i}p}{}^T\mathbf{A}_{\bar{i}p}R_{\bar{i}p} + \mathbf{A}_{\bar{i}q}{}^T\mathbf{A}_{\bar{i}q}R_{\bar{i}q}\right)\right)\mathbf{H}^{-T}, \qquad (29) \\ \text{and} \quad \mathbf{H} &= \sum_{\bar{i}=1}^{N}\left(\mathbf{A}_{\bar{i}p}{}^T\mathbf{A}_{\bar{i}p} + \mathbf{A}_{\bar{i}q}{}^T\mathbf{A}_{\bar{i}q}\right), \qquad (30) \\ \text{where} \quad \mathbf{R_u} &= \text{diag}[R_{\bar{i}p}, R_{\bar{i}q}]_{\bar{i}=1,...N}. \qquad (31)\end{aligned}$$

B. *Proof of Error Covariance Result*

We use the implicit function theorem [13] to prove the above result. The detailed proof of the above result can be found in [34].

C. Multi-Resolution Error Analysis

The method described above can now be applied in order to derive an exact expression for the error covariance of the 3D model obtained using the multi-resolution reconstruction scheme in Section IV. It relies upon the bilinear parameterization explained before. Since the algorithm guarantees convergence to a local minimum (the best that bundle adjustment can do), the error calculations are relevant and can be useful in practical situations. Moreover, the multi-resolution scheme provides a good initial condition of each resolution, and thus we can reasonably expect to reach the global optimum, or close to it. The error is first computed for the motion subsystem in Section IV, using the method described above for the single resolution case, and passed on to the structure subsystem of equations. The error computed for this subsystem is propagated to the next higher level of resolution. For multiple iterations within a level (intra-level), the procedure can be repeated before propagating the error covariance to the next higher level. However, for simplicity, we will derive the result for a single iteration within each level. Note that the assumption that the FOE is known is not required for the multi-resolution error calculations, and hence this error analysis is valid for the most general scenario.

At resolution $l+1$, we know the inverse depth and camera motion, $\mathbf{h}^{l+1}$ and $\mathbf{m}^{l+1}$ respectively, and are interested in computing these quantities at the next higher resolution l. We also know the error covariance of these quantities, $\mathbf{R}_{\mathbf{m}}^{l+1}$ and $\mathbf{R}_{\mathbf{h}}^{l+1}$ and our aim is to estimate them for the next higher resolution. At the first stage of the algorithm, we use the depth values from the lower resolution to update the camera motion parameters at this higher resolution. Thus, in equation (2), we know p^l, q^l and $\mathbf{h}^{l+1}$. The cost function that minimizes the square of the re-projection error is (see Section III for notational details)

$$C^l = \sum_{\bar{i}=1}^{N^l} \left[\left((p_{\bar{i}}^l - x_{\bar{i}} h_{\bar{i}}^{l+1}) - (\hat{p}_{\bar{i}}^l - x_{\bar{i}} \hat{h}_{\bar{i}}^{l+1}) \right)^2 + \left((q_{\bar{i}}^l - y_{\bar{i}} h_{\bar{i}}^{l+1}) - (\hat{q}_{\bar{i}}^l - y_{\bar{i}} \hat{h}_{\bar{i}}^{l+1}) \right)^2 \right]. \quad (32)$$

Thus the vector of unknowns $\mathbf{z} = [x_f^l, y_f^l, \omega_X^l, \omega_Y^l, \omega_Z^l]$. Then,

$$\frac{\partial C_i^l}{\partial \mathbf{z}} = \begin{cases} \mathbf{B}_{\bar{i}p}{}^l, & i \text{ odd} \\ \mathbf{B}_{\bar{i}q}{}^l, & i \text{ even} \end{cases}, \quad (33)$$

where $\mathbf{B}_{\bar{i}p}{}^l = [h_{\bar{i}}^{l+1}, 0, -\mathbf{r}_{\bar{i}}^l]$ and $\mathbf{B}_{\bar{i}q}{}^l = [0, h_{\bar{i}}^{l+1}, -\mathbf{s}_{\bar{i}}^l]$. The Hessian matrix of the camera motion parameters can be obtained similar to (30) as

$$\mathbf{H}^l_{\mathbf{m}} = \sum_{\bar{i}=1}^{N^l} \left({\mathbf{B}_{\bar{i}p}}^{lT} {\mathbf{B}_{\bar{i}p}}^{l} + {\mathbf{B}_{\bar{i}q}}^{lT} {\mathbf{B}_{\bar{i}q}}^{l} \right). \tag{34}$$

From (2), we see that since $\mathbf{h}^{l+1}$ is used to update the motion component, the input at this stage of the reconstruction is $\mathbf{u}^l_{\bar{i}} = [\tilde{p}^l_{\bar{i}}, \tilde{q}^l_{\bar{i}}] = [p^l_{\bar{i}} - x_{\bar{i}} h^{l+1}_{\bar{i}}, q^l_{\bar{i}} - y_{\bar{i}} h^{l+1}_{\bar{i}}]$. Thus the input covariance

$$\mathbf{R}^l_{u\bar{i}} = [R^l_{u\bar{i}p}, R^l_{u\bar{i}q}] = [R^l_{\bar{i}p} + x^2_{\bar{i}} R^{l+1}_{h_{\bar{i}}}, R^l_{\bar{i}q} + y^2_{\bar{i}} R^{l+1}_{h_{\bar{i}}}], \tag{35}$$

where R_{ni} is the ith component of R_n. The partial with respect to the input, at any resolution level l, is

$$\frac{\partial C^l_i}{\partial \mathbf{u}^l} = \left[\frac{\partial C^l_i}{\partial \tilde{p}^l_1} \ \frac{\partial C^l_i}{\partial \tilde{q}^l_1} \ \cdots \ \frac{\partial C^l_i}{\partial \tilde{p}^l_N} \ \frac{\partial C^l_i}{\partial \tilde{q}^l_N} \right] = \mathbf{I}_i(2N^l), \tag{36}$$

which is a 1 × 2N dimensional array. Therefore, the error covariance of the estimates of the camera motion parameters at resolution l is

$$\mathbf{R}^l_{\mathbf{m}} = \mathbf{H}^{l^{-1}}_{\mathbf{m}} \left(\sum_{\bar{i}=1}^{N^l} \left({\mathbf{B}_{\bar{i}p}}^{lT} {\mathbf{B}_{\bar{i}p}}^{l} R^l_{u\bar{i}p} + {\mathbf{B}_{\bar{i}q}}^{lT} {\mathbf{B}_{\bar{i}q}}^{l} R^l_{u\bar{i}q} \right) \right) \mathbf{H}^{l^{-T}}_{\mathbf{m}}. \tag{37}$$

The next step in the algorithm is to update the depth values depending on the camera motion parameters estimated in the previous step. Given $\mathbf{m}^l$ and $\mathbf{R}^l_{\mathbf{m}}$, we want to estimate $\mathbf{h}^l$ and $\mathbf{R}^l_{\mathbf{h}}$. The unknown parameter vector is $\mathbf{z} = [\mathbf{h}^l_1, ..., \mathbf{h}^l_{N^l}]$ and the input is $\mathbf{u}^l_{\bar{i}} = [p^l_{\bar{i}} - \mathbf{r}^{l^T}_{\bar{i}} \mathbf{\Omega}^l, q^l_{\bar{i}} - \mathbf{s}^{l^T}_{\bar{i}} \mathbf{\Omega}^l]$. The cost function, representing the re-projection errors that we want to minimize, is

$$C^l = \sum_{\bar{i}=1}^{N^l} \left[\left((p^l_{\bar{i}} - \mathbf{r}^{l^T}_{\bar{i}} \mathbf{\Omega}^l) - (\hat{p}^l_{\bar{i}} - \mathbf{r}^{l^T}_{\bar{i}} \hat{\mathbf{\Omega}}^l) \right)^2 + \left((q^l_{\bar{i}} - \mathbf{s}^{l^T}_{\bar{i}} \mathbf{\Omega}^l) - (\hat{q}^l_{\bar{i}} - \mathbf{s}^{l^T}_{\bar{i}} \hat{\mathbf{\Omega}}^l) \right)^2 \right]. \tag{38}$$

Then, from (2),

$$\frac{\partial C^l_i}{\partial \mathbf{z}} = \begin{cases} {\mathbf{D}_{\bar{i}p}}^{l}, & i \text{ odd} \\ {\mathbf{D}_{\bar{i}q}}^{l}, & i \text{ even} \end{cases}, \tag{39}$$

where ${\mathbf{D}_{\bar{i}p}}^{l} = [-(x^l_{\bar{i}} - x^l_f)\mathbf{I}_{\bar{i}}(N^l)]$ and ${\mathbf{D}_{\bar{i}q}}^{l} = [-(y^l_{\bar{i}} - y^l_f)\mathbf{I}_{\bar{i}}(N^l)]$. The partial with respect to the input is the same as in (36). The input covariance

$$\begin{aligned} \mathbf{R}^l_{u\bar{i}} &= [R^l_{u\bar{i}p}, \quad R^l_{u\bar{i}q}] \\ &= [R^l_{\bar{i}p} + x^{l2}_{\bar{i}} y^{l2}_{\bar{i}} R^l_{\omega_x} + (1 + x^{l2}_{\bar{i}})^2 R^l_{\omega_y} + y^{l2}_{\bar{i}} R^l_{\omega_z}, \\ &\quad R^l_{\bar{i}q} + (1 + y^{l2}_{\bar{i}})^2 R^l_{\omega_x} + x^{l2}_{\bar{i}} y^{l2}_{\bar{i}} R^l_{\omega_y} + x^{l2}_{\bar{i}} R^l_{\omega_z}], \end{aligned} \tag{40}$$

where the components of the motion are assumed uncorrelated and $R^l_{\omega_x}, R^l_{\omega_y}, R^l_{\omega_z}$ are the variances for each of the components. The Hessian matrix of the inverse depth can be obtained as in (30) as

$$\mathbf{H}^l_{\mathbf{h}} = \sum_{\bar{i}=1}^{N^l} \left(\mathbf{D}_{\bar{i}p}{}^{lT} \mathbf{D}_{\bar{i}p}{}^{l} + \mathbf{D}_{\bar{i}q}{}^{lT} \mathbf{D}_{\bar{i}q}{}^{l} \right). \tag{41}$$

Then, from Theorem 1, the error covariance of the estimates of the inverse depth at resolution l is

$$\mathbf{R}^l_{\mathbf{h}} = \mathbf{H}^{l^{-1}}_{\mathbf{h}} \left(\sum_{\bar{i}=1}^{N^l} \left(\mathbf{D}_{\bar{i}p}{}^{lT} \mathbf{D}_{\bar{i}p}{}^{l} R^l_{u\bar{i}p} + \mathbf{D}_{\bar{i}q}{}^{lT} \mathbf{D}_{\bar{i}q}{}^{l} R^l_{u\bar{i}q} \right) \right) \mathbf{H}^{l^{-T}}_{\mathbf{h}}. \tag{42}$$

Summarizing the above result, we have shown that starting with a solution for motion and structure, $\mathbf{m}^{l+1}$ and $\mathbf{h}^{l+1}$ with error covariances $\mathbf{R}^{l+1}_{\mathbf{m}}$ and $\mathbf{R}^{l+1}_{\mathbf{h}}$, we can obtain the error covariance for the solution at the next higher level of resolution, $\mathbf{R}^l_{\mathbf{m}}$ and $\mathbf{R}^l_{\mathbf{h}}$, by studying the propagation of the errors through the multi-resolution 3D modeling strategy. Thus, starting with the coarsest resolution, it is possible to analyze the quality of the reconstruction at any higher level of resolution.

Equation (42) is derived under the condition that $\mathbf{m}^l$ is known. If this condition is imposed in the derivation of Theorem 1 and $l = 1$, $(\mathbf{A}_{\bar{i}p}, \mathbf{A}_{\bar{i}q})$ would be redefined as $(\mathbf{A}_{\bar{i}ph}, \mathbf{A}_{\bar{i}qh})$ and this would be exactly the same as $(\mathbf{D}_{\bar{i}p}{}^{1}, \mathbf{D}_{\bar{i}q}{}^{1})$. Hence for $l = 1$, (42) is equal to the single resolution case (29).

VI. Experimental Evaluations

In this section, both computer-rendered image sequences and real video sequences are used to present results of the algorithm and numerically simulate the statistical result derived above.

A. Computer rendered image sequences

The source data were downloaded from http://sampl.eng.ohio-state.edu/ ˜sampl/ data/ 3DDB/ RID/minolta/ faces-hands.1299/ index.html. The data include face/hand texture images, range images, and corresponding masks for the valid data. Two frames of each image sequence were generated from a virtual camera with the ground truth focal length and motion parameters shown in Table I, and the optical flow is computed accordingly. All frames are 200×200 pixels. This

TABLE I

MOTION PARAMETERS OF FACES. (V_X, V_Y, V_Z) IS IN PIXELS AND CAN ONLY BE COMPUTED UP TO SCALE. $(\Omega_X, \Omega_Y, \Omega_Z)$ IS IN RADIANS. $f = 1000$ PIXELS.

Motion	V_X	V_Y	V_Z	Ω_X	Ω_Y	Ω_Z
Face 1	85.022	88.363	-2.631	-0.0853	0.0819	0.0010
Face 2	87.329	88.003	-1.836	-0.0847	0.0840	0.0008
Actual	120	120	0	-0.084	0.084	0

database was used in order to analyze the accuracy of the 3D reconstruction against the ground truth, i.e. true depth and motion values.

Figure 3(a) shows one frame of the ”Face 1” sequence. The invalid range data is filtered out when generating the sequence. An artificial checkerboard background is generated for each frame as a reference plane. Arbitrarily guessed flat depth is used at the coarsest level $l = 3$. Figures 3(c)-(d) show the face structure recovered at the end of the iteration process at level $l = 2$ and 0. Note that the axis scales are different in the figures. Figures 3(e) shows the recovered ”Face 1” structure with skin texture from one viewing angle. It can be observed that the face pops up from the flat surface at the coarsest level, and the structure is well propagated and refined through the hierarchical “scene” from coarse to fine.

Figure 4(a) shows the structure error statistics when adding noise of zero mean and variance of one pixel to the optical flow. It can be observed that the error variance decreases with the intra- and inter- levels of iterations. As a comparison, the result from a single level iteration is also shown in the same plot. Although the total number of iterations is twice as many as the hierarchical one, it still ends up with higher error variance for the structure estimates. When error variance of optical flow is too high due to the image noise (Figure 3(b) shows the image with noise and Figure 3(f) shows the noise distribution), the multi-resolution algorithm does not improve the structure estimate. Figure 4(b) shows the failure of a multi-resolution scheme in reconstructing the same sequence as the quality of the input data gets worse. The error variance decreases from level 3 to level 2. After that, it begins fluctuating and increases during inter-level propagation.

B. Real video sequences

We also used video sequences captured using a hand-held video camera to qualitatively test the robustness of our method. Figure 5(a) shows one frame of face from a real video sequence. The subject was moving his head left and right. The optical flow was calculated using the algorithm proposed in [39]. The frame are 320×240 pixels. Figure 5(b) shows that the structure of the face is recovered reasonably well. The structure and motion parameters were recovered using the algorithm designed in Section IV of this paper. The number of levels of resolution used for this experiment are three. The background is removed for clarity.

VII. Conclusion

In this paper, we have presented a multi-resolution SfM algorithm, analyzed its convergence properties, and derived a result on the statistical error characteristics of our algorithm. The algorithm exploits the bilinear parameterization of the depth and motion parameters in order to obtain a solution which is locally optimal at each level of the reconstruction. The accuracy of a multi-resolution modeling scheme depends upon the quality of the input video; therefore it is important to understand when to apply such a method. One of the main results of this paper is an analytical expression for the error covariance of the structure and motion estimates as a function of the error covariance of the feature tracks in the multi-resolution framework. The derivation is based on the implicit function theorem and does not require assumptions such as Gaussianity of the error statistics. Experiments with real video data demonstrate the accuracy of the reconstruction. The significance of the statistical calculations is demonstrated through numerical simulations.

Appendix

We now illustrate that the above algorithm converges to the local minimum of depth and motion estimates at each resolution level l.

A nonlinear system $\mathbf{y}(\mathbf{\Theta})$ is called *bilinear system* when its parameter vector $\mathbf{\Theta}_{d\times 1}$ can be split up into two parts $\mathbf{\Theta} = [\rho_{m\times 1} \quad \eta_{n\times 1}]^T, d = m + n$, such that $\mathbf{y}(\mathbf{\Theta})$ is linear in ρ for fixed η and linear in η for fixed ρ.

In our case, $\rho = \mathbf{m}$, $\eta = h^l$ in Section IV.

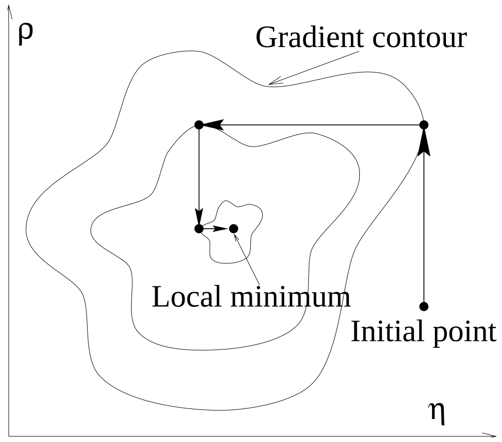

Fig. 2. Illustration of convergence.

Figure 2 illustrates the convergence process in a two dimensional case. Essentially, the algorithm splits the searching routine in space $\mathbb{R}^d$ into two searches in space $\mathbb{R}^m$ and $\mathbb{R}^n$ corresponding to ρ and η respectively. Since in $\mathbb{R}^m$ and $\mathbb{R}^n$, $\mathbf{y}(\rho, \eta(i-1))$ and $\mathbf{y}(\rho(i), \eta)$ are linear, the two sub-searches converge to the local minimum.

REFERENCES

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

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

[3] John Oliensis, "A critique of structure-from-motion algorithms," *Computer Vision and Image Understanding*, vol. 80, pp. 172–214, 2000.

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

[5] Z.Y. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *International Journal of Computer Vision*, vol. 27, pp. 161–195, March 1998.

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

[7] Z. Sun, V. Ramesh, and A.M. Tekalp, "Error characterization of the factorization method," *Computer Vision and Image Understanding*, vol. 82, pp. 110–137, May 2001.

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

[9] K.J. Hanna, "Direct multi-resolution estimation of ego-motion and structure from motion," in *Proc. of IEEE Workshop on Visual Motion*, 1991, pp. 156–162.

[10] H. Liu, R. Chellappa, and A. Rosenfeld, "A hierarchical approach for obtaining structure from two-frame optical flow," in *Proc. of IEEE Workshop on Motion and Video Computing*, 2002, pp. 214–219.

[11] R. Mandelbaum, G. Salgian, and H. S. Sawhney, "Correlation-based estimation of ego-motion and structure from motion and stereo," in *Proc. of International Conference on Computer Vision*, 1999, pp. 544–550.

[12] Vishvijit Nalwa, *A Guided Tour of Computer Vision*, Addison Wesley, 1993.

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

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

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

[16] H.C. Longuet-Higgins, “A computer algorithm for reconstructing a scene from two projections,” *Nature*, vol. 293, pp. 133–135, September 1981.

[17] R.Y. Tsai and T.S. Huang, “Estimating 3-d motion parameters of a rigid planar patch: I,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, pp. 1147–1152, December 1981.

[18] D.B. Gennery, “Tracking known three-dimensional objects,” in *AAAI-82*, 1982, pp. 13–17.

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

[20] T.J. Broida and R. Chellappa, “Estimating the kinematics and structure of a rigid object from a sequence of monocular images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 497–513, 1991.

[21] A. Azarbayejani and A. Pentland, “Recursive estimation of motion, structure, and focal length,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 562–575, 1995.

[22] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: A factorization method,” *International Journal of Computer Vision*, vol. 9, pp. 137–154, November 1992.

[23] R. Szeliski and S.B. Kang, “Recovering 3d shape and motion from image streams using non-linear least squares,” *Journal of Visual Computation and Image Representation*, vol. 5, pp. 10–28, 1994.

[24] J. Oliensis, “A multi-frame structure-from-motion algorithm under perspective projection,” *International Journal of Computer Vision*, vol. 34, pp. 1–30, August 1999.

[25] J. Oliensis and Yacup Genc, “Fast and accurate algorithms for projective multi-image structure from motion,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 546–559, June 2001.

[26] J. Inigo Thomas and J. Oliensis, “Dealing with noise in multiframe structure from motion,” *Computer Vision and Image Understanding*, vol. 76, pp. 109–124, 1999.

[27] J. Weng, N. Ahuja, and T.S. Huang, “Optimal motion and structure estimation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 864–884, September 1993.

[28] T.J. Broida and R. Chellappa, “Performance bounds for estimating three-dimensional motion parameters from a sequence of noisy images,” *Journal of the Optical Society of America A*, vol. 6, pp. 879–889, 1989.

[29] G.S. Young and R. Chellappa, “Statistical analysis of inherent ambiguities in recovering 3-d motion from a noisy flow field,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 995–1013, October 1992.

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

[31] A. Chiuso and S. Soatto, “3d motion and structure causally integrated over time: Theory and practice,” Tech. Rep., ESSRL 99-003, Washington University, Saint Louis, 1999.

[32] S. Soatto and R. Brockett, “Optimal structure from motion: Local ambiguities and global estimates,” in *Proc. of Conference on Computer Vision and Pattern Recognition*, 1998, pp. 282–288.

[33] D.D. Morris, K. Kanatani, and T. Kanade, “Gauge fixing for accurate 3D estimation,” in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2001, pp. II:343–350.

[34] Amit K. Roy Chowdhury and R. Chellappa, “Stochastic approximation and rate distortion analysis for robust structure and motion estimation,” *International Journal of Computer Vision*, pp. 27–53, October 2003.

[35] C. Fermller, D Shulman, and Y. Aloimonos, “The statistics of optical flow,” *Computer Vision and Image Understanding*, vol. 82, no. 1, pp. 1–32, April 2001.

[36] A. Roy Chowdhury, *Statistical Analysis of 3D Modeling From Monocular Video Streams*, PhD Thesis, 2002.

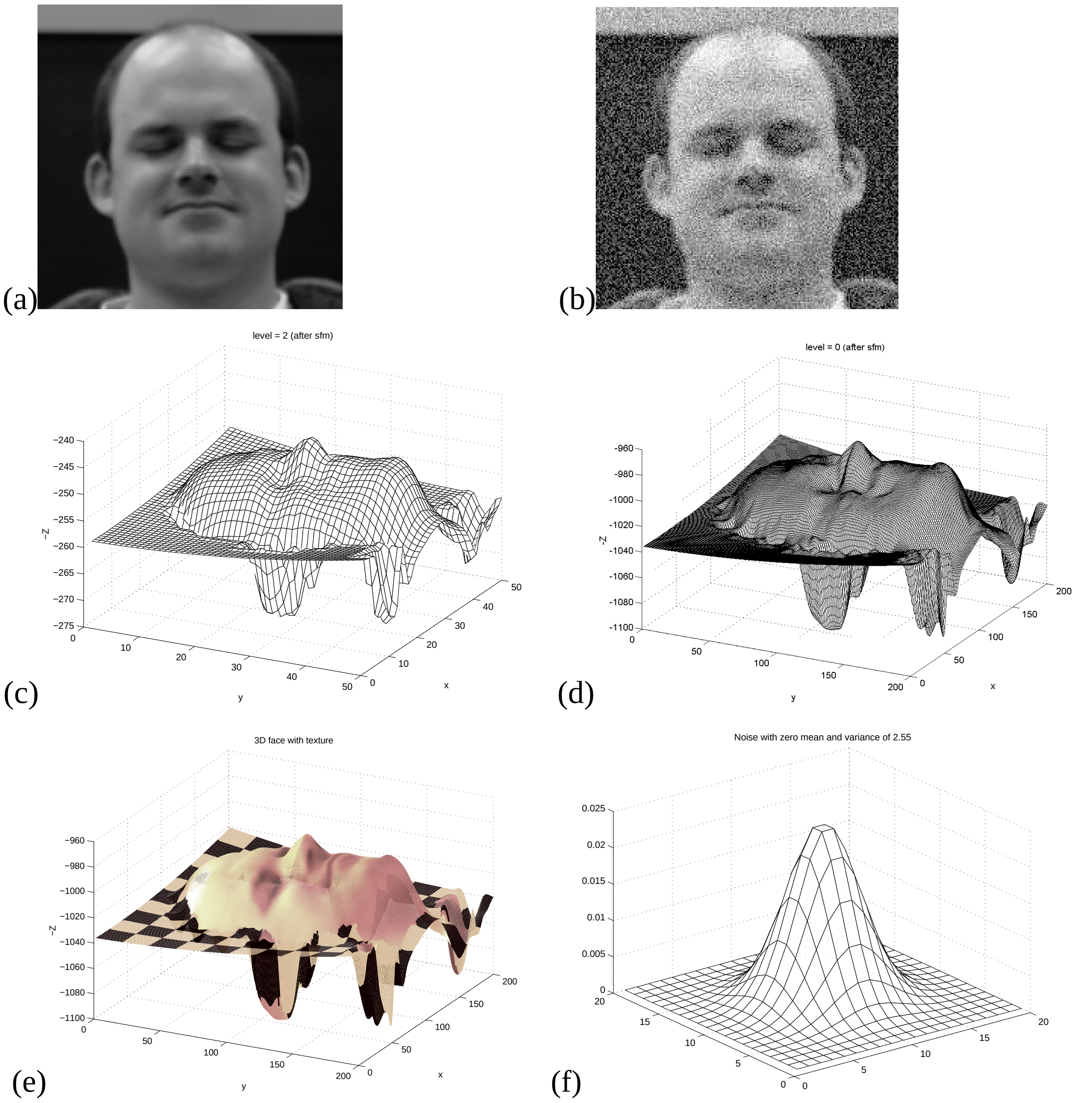

Fig. 3. Face sequences (200×200) (a) One frame of Face 1. (b) The frame with the noise of zero mean and variance 2.55. (c)(d) Depth recovered at levels $l = 2, 0$ (e) Recovered Face 1 structure with skin texture from one angle (f) Distribution of noise.

[37] A. Roy Chowdhury and R. Chellappa, "Statistical error propagation in 3d modeling from monocular video," in *CVPR Workshop on Statistical Analysis in Computer Vision*, 2003.

[38] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Inc., 1987.

[39] H. Liu, R. Chellappa, and A. Rosenfeld, "Accurate dense optical flow estimation and segmentation using adaptive structure tensors and a parametric model," *IEEE Trans. on Image Processing*, p. (to appear), 2003.

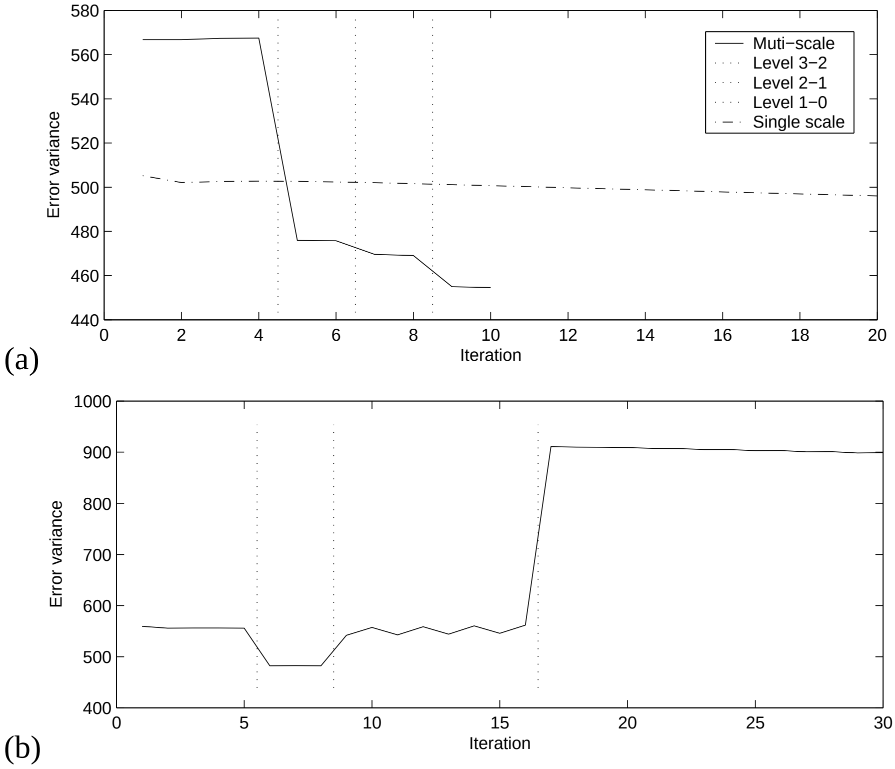

Fig. 4. Error statistics. Optical flow contains noise of zero mean and variance of 1. Solid line: Multi-scale (four levels with totally 10 iterations). Dotted lines: Level transitions. Dash dot line: Single scale (one level with totally 20 iterations). (a) Face 1. (b) Failed example of Face 1 when noise variance is high.

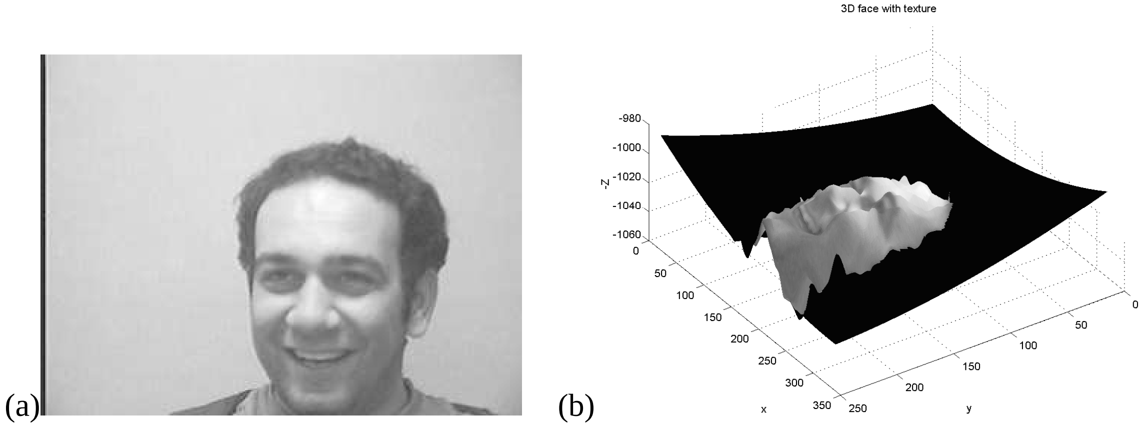

Fig. 5. Real sequence (320×240, with noisy optical flow) (a) One frame of face image from real video sequences 1 and 2. (b) Recovered face structures 1 and 2 with skin texture.