International Journal of Computer Vision 59(1), 5–31, 2004

Structure from Motion Using Sequential Monte Carlo Methods

GANG QIAN* AND RAMA CHELLAPPA
Department of Electrical and Computer Engineering, Center for Automation Research, University of Maryland, College Park, MD 20742-3275, USA
gang.qian@asu.edu
rama@cfar.umd.edu

Received June 5, 2001; Revised August 11, 2003; Accepted August 12, 2003

Abstract. In this paper, the structure from motion (SfM) problem is addressed using sequential Monte Carlo methods. A new SfM algorithm based on random sampling is derived to estimate the posterior distributions of camera motion and scene structure for the perspective projection camera model. Experimental results show that challenging issues in solving the SfM problem, due to erroneous feature tracking, feature occlusion, motion/structure ambiguity, mixed-domain sequences, mismatched features, and independently moving objects, can be well modeled and effectively addressed using the proposed method.

Keywords: structure from motion, sequential Monte Carlo methods, video analysis

1. Introduction

In general, structure from motion (SfM) refers to the task of of recovering the 3D (stationary or dynamic) scene structure and sensor motion trajectory given a set of 2D (monocular or stereo) image frames obtained from a (calibrated or uncalibrated) optical camera. In this paper, we focus on SfM using image frames captured by a monocular calibrated camera—i.e., the principal point and the field of view (FOV) of the camera are assumed to be known.

The SfM problem has been a very active research area in computer vision since early eighties when Longuet-Higgins published the famous "eight point" algorithm (Longuet-Higgins, 1981). The past two decades have witnessed a multitude of methods for solving the SfM problem. These methods can be classified based on the camera projection models (perspective/affine projection), observation sources (image pair, image tuple or video sequence), and token types (sparse feature correspondence, dense optical flow field, lines or curves, or even direct SfM techniques that do not extract any tokens from the images). Reviews and comparison of different SfM methods can be found in Jerian and Jain (1991), Huang and Netravali (1994), Faugeras (1993), Tian et al. (1996), Jebara et al. (1999) and Oliensis (2000). Although many algorithms have been developed, few give satisfactory performance in real applications. To develop an SfM method that performs well in practice, one must consider the following issues: (1) observation noise (noise present in token correspondence or in computing optical flow), (2) feature occlusion, (3) motion/structure recovery ambiguities, (4) mixtures of image sequences having both small and large baselines and (5) mismatched tokens and/or independently moving objects in the observed image frames. Being able to handle these issues is critical for producing practical SfM algorithms. Although recently, elegant methods have been reported in Soatto and Brockett (1998) and Forsyth et al. (1999), more progress needs to be made in addressing the issues raised above.

*Dr. Qian is now affiliated with the Arts, Media and Engineering Program and the Department of Electrical Engineering at Arizona State University.

In this paper, we focus on developing a robust statistical SfM method using noisy sparse feature correspondences from calibrated video sequences under perspective projection. Since a moving camera can be viewed as a kinematic system with its motion characterized by state parameters, a state space model can be used to describe the camera motion. However, due to perspective projection, the observation equation of the kinematic system is nonlinear. Although nonlinear filtering techniques such as the extended Kalman filter and its variants (Broida et al., 1990; Azarbayejani and Pentland, 1995; Chiuso et al., 2002) have been applied to solve the SfM problem, the results are not satisfactory. Recently, sequential Monte Carlo (SMC) methods have received more attention for estimation, prediction, filtering and smoothing of nonlinear/non-Gaussian state space models. Several SMC methods have been proposed in various scenarios. Isard and Blake (1996) developed the CONDENSATION algorithm for shape and contour tracking. The Monte Carlo filter proposed by Kitagawa (1996) can be viewed as a generalized algorithm dealing with the state estimation problem for a nonlinear/non-Gaussian dynamic system. The particle filter (or bootstrap filter) (Gordon et al., 1993) is a variant of the SMC method. In nature, these methods are very similar: samples and weights are propagated from successive time instants to describe distributions of interest. Liu and Chen (1998) proposed a general SMC framework for dynamic systems, the sequential importance sampling and some of the above methods can be interpreted as special cases of the general sequential importance sampling framework. In this paper, we mainly follow the notations and approach of Liu and Chen (1998).

In the SfM problem, both camera motion and scene structure are estimated. In our approach, the camera motion is estimated first using an SMC method based on the epipolar constraint, and then the scene structure is recovered using the motion estimates. Recently, the SMC technique was used for the SfM problem (Forsyth et al., 1999) by Forsyth, Ioffe and Haddon. In Forsyth et al. (1999), the approach followed the spirit of the factorization method (Tomasi and Kanade, 1992). It was shown that the SMC method is capable of selecting valid feature points for object shape reconstruction and moving object segmentation. However, since the method is a batch algorithm and a high-dimensional parameter estimation problem is being solved, the procedure is time consuming. Although it is possible to develop a similar algorithm using the perspective projection model, which is more suitable for practical applications, convergence problems and the efficiency of the resulting algorithm may limit its usefulness.

In this paper, we develop a recursive algorithm for finding the posterior distribution of the sensor motion parameters using the perspective projection camera model and the sequential importance sampling (SIS) technique (Liu and Chen, 1998). The structure of the scene can be subsequently recovered. When the scene is dynamic, the algorithm can also detect points on the background. By finding the posterior distribution of the motion and scene structure parameters, a much clearer picture of the structure of the solution space can be obtained. Through this, not only can good estimates be found, but the uncertainty of the estimates can also be characterized.

2. Bayesian Motion Estimation

To solve the SfM problem means to find the optimal estimates for camera motion, scene geometry and segmentation which can provide the best interpretation of the observations using criteria such as the maximum *a posteriori* probability (MAP). In this paper, *Prob*(*parameters* | *observation*), the posterior distribution of the parameters, is approximated using random sampling methods. In this section, we first introduce the SIS technique. We then formulate the SfM problem using a state space model and then develop an SIS algorithm for finding the approximation to the posterior distribution of the state parameters. Before the presentation of algorithms, we would like to clarify some notations. Suppose we have a dynamic system.

$\mathbf{x}_t$, the state parameters of the dynamic system at time t
$\mathbf{y}_t$, the measurements observed from the system at time t
$\mathcal{X}_t = \{\mathbf{x}_i\}_{i=1}^t$, the state sequence up to time t
$\mathcal{Y}_t = \{\mathbf{y}_i\}_{i=1}^t$, the observation sequence up to time t

Let $\pi_t(\mathbf{x}_i) = p(\mathbf{x}_i \mid \mathcal{Y}_t)$ and $\pi_t(\mathcal{X}_i) = p(\mathcal{X}_i \mid \mathcal{Y}_t)$ be the posterior distributions of state at time i and that of the state sequence up to time i given observations up to time t.

2.1. Sequential Importance Sampling

The SIS method is a recently proposed technique for approximating the posterior distribution of the state

parameters of a dynamic system (Liu and Chen, 1998). Usually, the state space model of a dynamic system is described by observation and state equations. If the measurement is denoted by $\mathbf{y}_t$ and the state parameter by $\mathbf{x}_t$, the observation equation essentially provides the conditional distribution of the observation given the state, $f_t(\mathbf{y}_t \mid \mathbf{x}_t)$. Similarly, the state equation gives the Markov transition distribution from time t to time $t+1$, $q_t(\mathbf{x}_{t+1} \mid \mathbf{x}_t)$. The goal is to find the posterior distribution of the states $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)$ given all the available observations up to t, $\pi_t(\mathcal{X}_t)$, which can be further decomposed using conditional probability as

$$\begin{aligned}\pi_t(\mathcal{X}_t) &= P\big(\{\mathbf{x}_i\}_{i=1}^t \,\big|\, \{\mathbf{y}_i\}_{i=1}^t\big) \\ &= \prod_{i=1}^{t} f_i(\mathbf{y}_i \mid \mathbf{x}_i) q_i(\mathbf{x}_i \mid \mathbf{x}_{i-1})\end{aligned}$$

and $q_1(\mathbf{x}_1 \mid \mathbf{x}_0) = \pi(\mathbf{x}_0)$ is the prior distribution of the state parameters. One way to represent the approximation of the posterior distribution is by a set of samples and their corresponding weights.

Definition (Liu and Chen, 1998). A random variable X drawn from a distribution g is said to be **properly weighted** by a weighting function $w(X)$ with respect to the distribution π if for any integrable function h,

$$E_g h(X) w(X) = E_\pi h(X).$$

A set of random draws and weights $(x^{(j)}, w^{(j)})$, $j = 1, 2, \ldots,$ is said to be properly weighted with respect to (w.r.t.) π if

$$\lim_{m\to\infty} \frac{\sum_{j=1}^m h(x^{(j)}) w^{(j)}}{\sum_{j=1}^m w^{(j)}} = E_\pi h(X) \qquad (1)$$

for any integrable function h.

Suppose $\{\mathcal{X}_t^{(j)}\}_{j=1}^m$ is a set of random samples properly weighted by the set of weights $\{w_t^{(j)}\}_{j=1}^m$ w.r.t. π_t and let g_{t+1} be an arbitrary trial distribution. Then the recursive SIS procedure to obtain the random samples and weights properly weighting π_{t+1} is as follows.

SIS steps: (Liu and Chen, 1998) for $j = 1, \ldots, m$,

(A) Draw $X_{t+1} = \mathbf{x}_{t+1}^{(j)}$ from $g_{t+1}(\mathbf{x}_{t+1} \mid \mathcal{X}_t^{(j)})$. Attach $\mathbf{x}_{t+1}^{(j)}$ to form $\mathcal{X}_{t+1}^{(j)} = (\mathcal{X}_t^{(j)}, \mathbf{x}_{t+1}^{(j)})$.

(B) Compute the "incremental weight" u_{t+1} by

$$u_{t+1}^{(j)} = \frac{\pi_{t+1}\big(\mathcal{X}_{t+1}^{(j)}\big)}{\pi_t\big(\mathcal{X}_t^{(j)}\big) g_{t+1}\big(\mathbf{x}_{t+1} \,\big|\, \mathcal{X}_t^{(j)}\big)}$$

and let $w_{t+1}^{(j)} = u_{t+1}^{(j)} w_t^{(j)}$.

It can be shown (Liu and Chen, 1998) that $\{\mathcal{X}_{t+1}^{(j)}, w_{t+1}^{(j)}\}_{j=1}^m$ is properly weighted w.r.t. π_{t+1}. Hence, the above SIS steps can be applied recursively to obtain the properly weighted set for future time instants once the corresponding observations are available. It is not difficult to show that given the properly weighted samples $\{\mathcal{X}_t\}$ w.r.t. the joint posterior distribution $\pi_t(\mathcal{X}_t)$, the "marginal" samples formed by the components in $\{\mathcal{X}_t\}$ of $\mathbf{x}_i$ are properly weighted by the same set of weights respect to the marginal posterior distribution $\pi_t(\mathbf{x}_i)$. Once the properly weighted samples of the joint distribution are obtained, the marginal distributions are approximated by the "marginal" samples weighted by the same set of weights.

The choice of the trial distribution g_{t+1} is crucial in the SIS procedure since it directly affects the efficiency of the proposed SIS method. In our approach, we used

$$g_{t+1}(\mathbf{x}_{t+1} \mid \mathcal{X}_t) = q_{t+1}(\mathbf{x}_{t+1} \mid \mathbf{x}_t)$$

because of the convenience it provides during the computation and the satisfactory performance the resulting SIS method gives in estimating structure and motion. It can be shown that in this case $u_{t+1} \propto f(\mathbf{y}_{t+1} \mid \mathbf{x}_{t+1})$, the conditional probability density function of the observations at $t+1$ given the state sample $\mathbf{x}_{t+1}$. $f(\mathbf{y}_{t+1} \mid \mathbf{x}_{t+1})$ is also known as the likelihood function of $\mathbf{x}_{t+1}$ since the observations are fixed.

In SIS, an additional resampling step (Liu and Chen, 1998) often follows the sample weight evaluation after drawing new samples for current state. Assume that sample set $\mathcal{S}_t = \{\mathcal{X}_t^j\}_{j=1}^N$ is properly weighted by $\{w_t^j\}_{j=1}^N$. Resampling includes the following two steps.

Resampling: (Liu and Chen, 1998):

(A) Draw a new sample set $\mathcal{S}_t'$ from $\mathcal{S}_t$ according to the weights $w_t^{(j)}$.
(B) Assign equal weights to all samples in $\mathcal{S}_t'$.

A major benefit of resampling is to statistically reduce bad samples (with small weights) and encourage good samples so that good samples will produce

enough number of offspring to describe the distribution of future states. Since resampling will reduce the size of distinct samples, it might be harmful to do resampling when the variation of sample weights is small, i.e. when the samples are more or less equally weighted (important). Resampling is for a better empirical distribution of future states and it does not improve the estimation of current state since it introduces extra Monte Carlo variations in current samples. It is suggested to perform state estimation before resampling (Liu and Chen, 1998).

Sample Efficiency. SMC are importance sampling based methods. The efficiency of a SMC method can be measured by comparing it with the direct sampling from the target distribution. Quantitatively, it can be represented by the *effective sample size* (ESS), which is the size of the samples needed to be drawn from the target distribution to have the equivalent estimation accuracy using the SMC algorithm. Fixing the number of samples used in SMC, a large ESS indicates high efficiency. Although ESS depends on the statistics (functions of the states) to be estimated, it can be approximately computed (Kong et al., 1994) from sample weights by

$$ESS = \frac{m}{1 + m * var(w)} \tag{2}$$

where m is the number of samples used in SMC and w are the normalized weights of these samples. Over-resampling could greatly deteriorate the efficiency of a SMC algorithm and result in the *sample impoverishment* problem, in which case all samples collapse to only a few points in the state space. To avoid over-resampling and sample impoverishment, resampling is not performed in every recursion. Instead, only when the ESS is under certain threshold, resampling will be revoked.

2.2. Overview of the Algorithm

In the SfM problem, both camera motion and scene structure are estimated. There are two main strategies for doing this (Soatto and Perona, 1998). In one strategy, the camera motion is estimated first using geometric constraints on rigid body motion such as the epipolar constraint, and then the scene structure is recovered using the motion estimates. In the other strategy, structure and motion are estimated simultaneously. The second strategy results in a high-dimensional state space (the dimension increases linearly with the number of feature points), which is not favorable to the SIS procedure. In our approach, we use the first strategy: firstly the camera motion is solved and then the scene structure is computed based on motion estimates. The estimation of camera motion can be once more divided into two steps. The camera motion parameters without translation magnitude is firstly estimated using the well-known epipolar constraint. Then the translation magnitude is computed through triangulation. In the remaining part of the paper, the following notations are used.

- $\mathbf{x}_t$, partial camera motion (without translation magnitude) at time t
- γ_t, camera translation magnitude at time t
- $\mathbf{z}$, depth values of feature points
- $\mathbf{y}_t$, feature correspondences at time t
- $\mathcal{X}_t = \{\mathbf{x}_i\}_{i=1}^t$, partial camera motion sequence up to time t
- $\Gamma_t = \{\gamma_i\}_{i=1}^t$, sequence of camera translation magnitude up to time t
- $\mathcal{Y}_t = \{\mathbf{y}_i\}_{i=1}^t$, feature trajectories up to time t

The complete structure and motion estimation algorithm can be illustrated by Fig. 1. Three concatenated SIS procedures are involved in the complete algorithm, the first one for partial motion parameters, the second for translation magnitude and the last one for structure parameters. Each SIS procedure uses the results from both the preceding SIS procedure at the same time instant and the same SIS procedure at the previous time instant to update samples and weights. Hence, the complete camera motion and scene structure samples are built up progressively in dimension and recursively in time. Each of these SIS procedures will be discussed in details in the following sections.

2.3. State Space Model for Camera Motion

We first introduce the SIS procedure for partial camera motion parameters, without translation magnitude.

Parameterization of Camera Motion. Before discussing the parameterization of camera motion, we introduce two 3D Euclidean coordinate systems used in our research. One coordinate system is attached to the camera and uses the center of projection of the camera as its origin. It is denoted by C. The Z axis of C is along the optical axis of the camera, with the positive

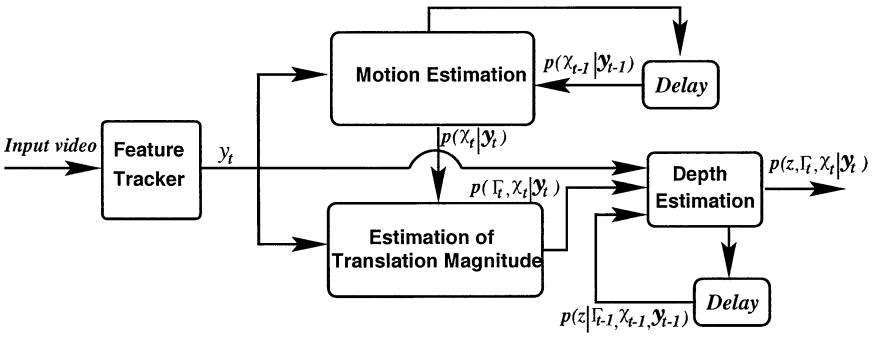

Figure 1. Overview of the algorithm.

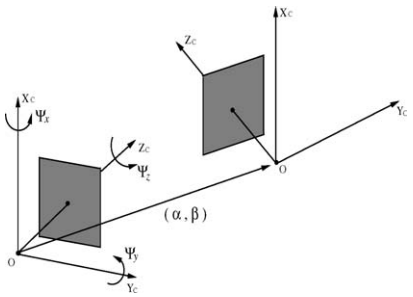

Figure 2. Motion parameters of a moving camera.

half-axis in the camera looking direction. The X-Y plane of C is perpendicular to the Z axis with the X and Y axes parallel to the borders of the image plane. Also, the X-Y-Z axes of C satisfy the right-hand rule. The other coordinate system is a world inertial frame, denoted by I. I is fixed on the ground. The coordinate axes of I are configured in such a way that initially, I and C coincide. When the camera moves, C travels with the camera and I stays at the initial position. It is worthy to mention that in our approach, the camera motion to be estimated is not between two successive time instants, instead, it is the globe overall motion of the camera in the world system I. By computing camera motion in I, the translation can be accumulated over time and the scene structure can be reliably estimated.

As illustrated by Fig. 2, the global camera motion can be described by five parameters.

$$\mathbf{m}_t = (\psi_x, \psi_y, \psi_z, \alpha, \beta)^T$$

(ψ_x, ψ_y, ψ_z) are the rotation angles of the camera about the coordinate axes of the inertial frame I and (α, β) are the elevation and azimuth angles of the camera translation direction, measured in the world system I. The unit vector in the translation direction is given by

$$\mathbf{T}(\alpha, \beta) = (\sin(\alpha)\cos(\beta), \sin(\alpha)\sin(\beta), \cos(\alpha))^T \quad (3)$$

In physical world, objects bear inertia during mechanical motion. Inertia resists sudden change of object motion and is proportional to the mass of moving objects. To capture the inertia of moving objects, we include $\dot{\mathbf{m}}_t$, the velocity of the above global motion parameter $\mathbf{m}_t$ in the motion state vector.

$$\dot{\mathbf{m}}_t = (\dot{\psi}_x, \dot{\psi}_y, \dot{\psi}_z, \dot{\alpha}, \dot{\beta})^T$$

Hence the state vector we use to represent camera motion is

$$\mathbf{x}_t = (\mathbf{m}_t, \dot{\mathbf{m}}_t)$$

State Space Model. Given the above motion parameterization, a state space model can be used to describe the behavior of a moving camera.

$$\mathbf{m}_{t+1} = \mathbf{m}_t + \dot{\mathbf{m}}_t + \mathbf{n}_m \quad (4)$$

$$\dot{\mathbf{m}}_{t+1} = \dot{\mathbf{m}}_t + \mathbf{n}_{\dot{m}} \quad (5)$$

$$\mathbf{y}_t = Proj(\mathbf{m}_t, \mathbf{P}) + \mathbf{n}_y \quad (6)$$

where $\mathbf{m}_t$ and $\dot{\mathbf{m}}_t$ are respectively the motion and moving velocity of the camera. $\mathbf{y}_t$ is the observation at time t, which contains 2D projections of feature points on the image plane at current time instant. $\mathbf{n}_m$ and $\mathbf{n}_{\dot{m}}$ denote the dynamic noise in the system, describing the time-varying property of camera motion parameters. If

no prior knowledge about motion is available, a random walk will be a suitable alternative for modeling the dynamic changes in camera motion.

Proj$(\cdot)$ denotes the perspective projection function, describing the projection of feature points with structure $\mathbf{P}$ in world system I onto the image plane after camera motion $\mathbf{m}_t$. It can be interpreted as follows. Suppose the 3D position of a point p in the world system I is $P = (X, Y, Z)^T$, and its 3D position in current camera-centered system C is $P_t = (X_t, Y_t, Z_t)^T$. Then the projection of p onto the image plane after camera motion $\mathbf{m}_t$ is

$$u = f\frac{X_t}{Z_t} \tag{7}$$

$$v = f\frac{Y_t}{Z_t} \tag{8}$$

where f is the focal length of the camera. If the camera motion parameter at time t is $\mathbf{x}_t = (\Psi, \alpha, \beta)$, then

$$P_t = \mathbf{R}(\Psi)(P - \gamma\mathbf{T}(\alpha, \beta)) \tag{9}$$

where camera translation vector $\mathbf{T}$ is given by (3) and γ is the translation magnitude. $\Psi = (\psi_x, \psi_y, \psi_z)$ denotes the camera rotational angles and the rotation matrix $\mathbf{R}(\Psi)$ can be computed by

$$\mathbf{R}(\Psi) = \begin{pmatrix} n_1^2 + (1 - n_1^2)\,\eta & n_1 n_2\,(1-\eta) + n_3\zeta & n_1 n_3\,(1-\eta) - n_2\zeta \\ n_1 n_2\,(1-\eta) - n_3\zeta & n_2^2 + (1 - n_2^2)\,\eta & n_2 n_3\,(1-\eta) + n_1\zeta \\ n_1 n_3\,(1-\eta) + n_2\zeta & n_2 n_3\,(1-\eta) - n_1\zeta & n_3^2 + (1 - n_3^2)\,\eta \end{pmatrix} \tag{10}$$

where $n = (n_1, n_2, n_3)^T = \frac{\Psi}{|\Psi|}$ is the direction cosine vector, $\zeta = \sin|\Psi|$, and $\eta = \cos|\Psi|$.

2.4. SIS for Bayesian Motion Estimation

Based on the above state space model, we designed an SIS method for finding an approximation to the posterior distribution of the motion parameters. As mentioned above, the trial distribution in the SIS procedure used in our approach is chosen as $g_{t+1}(\mathbf{x}_{t+1} \mid \mathcal{X}_t) = q_{t+1}(\mathbf{x}_{t+1} \mid \mathbf{x}_t)$. Therefore, during the SIS step (A), we will draw samples from the distribution of $\mathbf{x}_t + n_x$.

Computation of Likelihood Function. To derive the likelihood function, consider the case when only one point is observed. Assume that at the initial time instant, a point p is projected to (u_0, v_0) in the first image plane. At time t after camera motion $\mathbf{x}_t$, feature tracking results indicate that p is at (u_t, v_t) in the image plane at current camera position. Due to the feature tracking noise, (u_t, v_t) is a noisy measurement of the true projection of p. Suppose that the distribution of feature tracking noise is normal with zero mean and covariance matrix given by $\left[\begin{smallmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{smallmatrix}\right]$.

When the camera only rotates, the likelihood function can be computed directly as

$$f((u_t, v_t) \mid \mathbf{x}_t) = \frac{1}{2\pi\sigma^2}\exp-\left\{\frac{(u_t - u'(u_0, v_0, \Psi_t)^2}{2\sigma^2} + \frac{(v - v'(u_0, v_0, \Psi_t))^2}{2\sigma^2}\right\} \tag{11}$$

where (u', v') is the reprojected position of point p in current image plane after camera rotation. It can be computed using (u_0, v_0) and camera rotation angles Ψ_t according to (8). In this case, the scene structure $\mathbf{P}$ is not involved in the computation of (u', v').

When the camera motion includes non-zero translation, the computation of the likelihood $f((u_t, v_t) \mid \mathbf{x}_t)$ becomes more difficult. Equation (11) can not be used directly since $\mathbf{P}$ will be needed to compute (u', v') when $\mathbf{T}(\alpha, \beta)$ is non-zero. However $\mathbf{P}$ is not represented in the state vector. To overcome this difficulty, in our approach, we applied conditional expectation over the structure to derive the equation for the likelihood function:

$$f((u_t, v_t) \mid \mathbf{x}_t) = \int_{\mathbf{P}_l}^{\mathbf{P}_u} f(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{P})p(\mathbf{P})d\mathbf{P} \tag{12}$$

where $[\mathbf{P}_l, \mathbf{P}_u]$ denotes the range of the feature depths such that all the features can be observed at both camera positions with positive feature depth values. However, the computation of the integral in (12) requires knowledge of the prior distributions of the scene structure $\mathbf{P}$ and the translation magnitudes, which are unknown. The epipolar constraint (Faugeras, 1993) is used here to resolve this difficulty. Recall that the epipolar constraint says that the perspective projections of a 3D point on the two image planes taken from different viewpoints lie on their corresponding epipolar lines, which are the intersections of the two image planes with the epipolar plane containing the 3D point and the two centers of projection (COP) of the camera. Given the image position of a point in one view and camera

motion parameters between the two views, the epipolar line related to this point in the other view can be easily determined. Note that

$$E_{\mathbf{P}}\{f(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{P})\} = E_{\mathcal{P}_t}\{f(\mathbf{y}_t \mid \mathbf{x}_t, \mathcal{P}_t)\}$$

where $\mathcal{P}_t$ represents the image positions of the feature points at time t. The prior distribution of $\mathcal{P}_t$ is not available. We assume that $\mathcal{P}_t$ is uniformly distributed on the pixel sites on the corresponding epipolar line segments. Since no prior knowledge about the ratios between feature depth values and camera translation magnitude is given, the whole epipolar line is under consideration. However only the epipolar line segment that agrees with the positive-depth constraint is used to evaluate the above expectation.

In this paper, we assume that the feature point positions in the first image frame are exact. As illustrated in Fig. 3, let l denote the epipolar line segment of p at t and let (x_1, y_1), (x_2, y_2) be the two terminal points of l. The locations of the two points are easy to find given the camera motion $\mathbf{x}_t$ and the image position of p in the first frame using the positive-depth constraint. Let p_l be the project of p onto l and Δ be the distance from p to p_l. Let r_1 and r_2 are the distances from the two terminal points of l to p_l. The likelihood function of the motion parameter given a single point observation is

$$\begin{aligned} f((u, v) \mid \mathbf{x}_t) &= \frac{1}{2\pi(r_1 + r_2)\sigma^2} \int_{-r_1}^{r_2} \\ &\quad \times \exp -\frac{r^2 + \Delta^2}{2\sigma^2} dr \qquad (13) \\ &= \frac{\exp -\frac{\Delta^2}{2\sigma^2}}{2\sqrt{2\pi}\sigma(r_1 + r_2)} \\ &\quad \times \left(\operatorname{erf}\left(\frac{r_2}{\sqrt{2}\sigma} \right) + \operatorname{erf}\left(\frac{r_1}{\sqrt{2}\sigma} \right) \right) \qquad (14) \end{aligned}$$

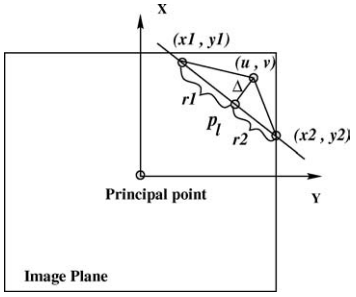

Figure 3. Epipolar line segment.

Given 2D trajectories of a set of feature points, multiple likelihood functions can be applied in different situations. If all the points are "good" features in the sense that they are not on moving objects nor mismatched, then $f(\mathbf{y}_t \mid \mathbf{x}_t)$ is obtained by multiplying the individual likelihood functions of the feature points.

$$f(\mathbf{y}_t \mid \mathbf{x}_t) = \prod_{i=1}^{M} f\left(\mathbf{y}_t^{(i)} \,\middle|\, \mathbf{x}_t\right) = \prod_{i=1}^{M} f\left(\left(u_t^{(i)}, v_t^{(i)}\right) \,\middle|\, \mathbf{x}_t\right) \qquad (15)$$

where M is the total number of the observed feature points. In some cases, some of the points are known to be not "good" beforehand. If the number of "bad" points is less than half of the total number of tracked feature points, the following equation can be used to compute the likelihood function for the observations:

$$f(\mathbf{y}_t \mid \mathbf{x}_t) = \prod_{i \in \mathcal{G}} f\left(\mathbf{y}_t^{(i)} \,\middle|\, \mathbf{x}_t\right) \qquad (16)$$

where $\mathcal{G} = \{i : f(\mathbf{y}_t^{(i)} \mid \mathbf{x}_t) \geq median(\{f(\mathbf{y}_t^{(i)} \mid \mathbf{x}_t)\}_{i=1}^{M}, i = 1, 2, \ldots, M\}$. The SIS method for Bayesian motion estimation is then as follows.

SIS Procedure for Camera Motion Distribution

1. *Initialization.* Suppose N samples are used to describe the motion distribution. Draw motion samples $\{\mathbf{x}_0^{(j)}\}_{j=1}^{N} = \{\mathbf{m}_0^{(j)}, \dot{\mathbf{m}}_0^{(j)}\}_{j=1}^{N}$ from $\pi(\mathbf{x}_0)$, which is the initial distribution of the motion parameters before the camera moves. Note that before camera begins moving, $\mathbf{x}_0 \neq 0$. Although the rotation angle Ψ and the translational vector are zero, the translational direction angles are uniformly distributed in their ranges. For each initial motion sample, the components of the rotation angles are all set to zero and the samples of α and β are drawn from two uniform distributions over $[0, \pi]$ and $[0, 2\pi]$, respectively.

   Since video sequences are used as image sources other than sets of image frames in arbitrary orders, the changes of the motion parameters usually are small. Assume that the rotational angular velocity obeys normal distribution and the translational direction angle velocity, a uniform distribution. For simplicity, dynamic noises in camera motion velocities are sampled from the following

distributions.

$$\begin{cases} n_{\dot{\psi}_\iota} \sim \mathcal{N}(0, \sigma_\iota), \iota \in \{x, y, z\} \\ n_{\dot{\kappa}} \sim U(-\delta_\kappa, \delta_\kappa), \kappa \in \{\alpha, \beta\} \end{cases} \tag{17}$$

where σ_ι,δ_α and δ_β are small positive real values. Since all the motion samples are drawn from a nearly exact distribution, equal weights are assigned to these samples.

To cover the case of pure rotation, some samples are treated as pure rotation samples according to probability P_{pr}, which can be set based on the prior knowledge of the camera motion. If P_{pr} is not available, 0.5 could be a good guess, when $t > 1$, the weights of pure rotation samples are computed using the likelihood function for pure rotation given by (11).

For $t = 1, \ldots, \tau$:

2. *Sample Generation.* Draw new samples for current time instant based on samples from previous time instant. Samples $\{\mathbf{x}_t^{(j)}\}_{j=1}^N$ are drawn based on the camera motion dynamics specified by (5) and (6). Hence for the jth sample $\mathbf{x}_t^{(j)} = (\mathbf{m}_t^{(j)}, \dot{\mathbf{m}}_t^{(j)})$,

$$\mathbf{m}_t^{(j)} = \mathbf{m}_{t-1}^{(j)} + \dot{\mathbf{m}}_{t-1}^{(j)} + \mathbf{n}_m \tag{18}$$

$$\dot{\mathbf{m}}_t^{(j)} = \dot{\mathbf{m}}_{t-1}^{(j)} + \mathbf{n}_{\dot{m}} \tag{19}$$

$\mathbf{n}_m$ and $\mathbf{n}_{\dot{m}}$ are sampled from similar distributions as in (17), but with different values of variances. Usually, the dynamic noise variances in $\mathbf{n}_m$ are smaller that those in $\mathbf{n}_{\dot{m}}$ since most of the changes in $\mathbf{m}_t$ have been taken care of by the prediction through $\dot{\mathbf{m}}_{t-1}$.

3. *Sample Transfer.* The status of camera motion in the sense of pure rotation can change from one time instant to the next time instant. Some samples in the two groups are exchanged to accommodate the shift in motion between pure rotation and general motion. If no information about the change of camera motion is given, half samples of pure rotation and general motion are exchanged. Let $\{w_{pr}^{(j)}\}$ be the weights corresponding to pure rotation samples computed using (11) and $\{w_g^{(j)}\}$ the weights of general motion samples. The posterior probability that the sequence is pure rotation can be inferred using these two weight sets by

$$P(\textit{pure rotation}) = \frac{\sum_j w_{pr}^{(j)}}{\sum_j w_{pr}^{(j)} + \sum_j w_g^{(j)}} \tag{20}$$

4. *Weight computation and re-sampling.*

   – Compute the sample weights $\{w_t^{(j)}\}$. For samples in the pure rotation group, the likelihood of a single feature point is computed using the function for pure rotation camera motion (11). Otherwise, the likelihood function for general camera motion (14) is used. If all feature points are good features without any mismatches, the product of the likelihood of all points is used as the weight for a motion sample as shown in (15). However, if some of the feature points are bad, (16) is used to compute the weight of a motion sample where only the likelihood of good feature points are used. The resulting samples and their corresponding weights $(\mathcal{X}_t^{(j)}, w_t^{(j)})$ are approximately properly weighted w.r.t. $\pi_t(\mathcal{X}_t)$.
   – Compute the current ESS using (2). When the ESS is lower than a threshold, (m/3 in our implementation), resample the above samples to amplify important motion samples so that good samples could be found for the motion parameters for the next time instant.

By using these properly weighted sample-weight sets for each time instant, the mean of the motion parameters can be computed directly. Also since the sample-weight sequences after re-sampling approximates $\mathcal{X}_t$ in distribution, the MAP estimates of $\mathcal{X}_t$ can also be obtained by locating the modes of $\pi_t(\mathcal{X}_t)$.

2.5. *Dynamic Scenes*

So far, we have discussed camera motion estimation for cases where the sets of features being tracked are good (static scene no mismatched feature points) and the number of bad feature points is less than the number of good ones. However, in real applications, the number of bad feature points is often larger than the number of points on the background. For example, multiple moving objects with different motion trajectories might be present in the scene and the features on these objects might be the majority of the tracked feature points. Therefore, the method proposed in the last section for dealing with mismatched feature points will not work well in this case. A new algorithm to simultaneously select good feature points and produce Bayesian camera motion estimates is required. This algorithm is now described.

Validity Vector. The state vector to be estimated is extended by adding a so-called validity vector $\mathbf{v}$:

$$\mathbf{x}_t = (\mathbf{v}_t, \mathbf{m}_t, \dot{\mathbf{m}}_t) \tag{21}$$

where $\mathbf{m}_t$ and $\dot{\mathbf{m}}_t$ represent the motion parameters. $\mathbf{v}_t$ is a $M \times 1$ vector. (Recall that M is the number of feature points). Each entry of $\mathbf{v}_t$ is associated with a feature point and indicates the possibility that the point is not mismatched and belongs to the background.

Trial Function. For the motion parameters $\mathbf{m}_t$, the trial function in the random sampling step in the SIS procedure remains the same. The trial function for the validity vector $\mathbf{v}$, $g(\mathbf{v}_{t+1} \mid \mathbf{v}_t)$, is similarly chosen as

$$g(\mathbf{v}_{t+1} \mid \mathbf{v}_t) = \pi_t(\mathbf{v}_{t+1} \mid \mathbf{v}_t) = Pr(\mathbf{v}_{t+1} \mid \mathbf{v}_t, \mathcal{Y}_t) \tag{22}$$

and the samples at $t+1$ are drawn using

$$\mathbf{v}_{t+1} = \gamma \mathbf{v}_t + \xi(\mathbf{m}_t, \mathbf{y}_t) + n_{\mathbf{v}} \tag{23}$$

where $n_{\mathbf{v}}$ is the noise in the validity vector and γ is an exponential forgetting factor, close to one. Equation (23) can be viewed as a definition for the conditional probability of $\mathbf{v}_{t+1}$ given $\mathbf{v}_t$. Both of them represent the possible time-varying nature of the validity vector: $\xi(\cdot)$ is a function used to update the current validity vector. Each element of ξ is given by

$$\xi_i(\mathbf{m}_t, \mathbf{y}_t) = \left(\frac{e_{th}}{e_i + 1}\right)^2 - sign(\lfloor e_i / e_{th} \rfloor) \frac{e_i + 1}{e_{th}} \tag{24}$$

where $e_i = e(\mathbf{m}_t, \mathbf{y}_t^{(i)})$ is the distance from the ith feature point to its associated epipolar line given the motion parameters $\mathbf{m}_t$. eth is a prechosen threshold for this distance. When e_i is larger than eth, $\xi_i(\cdot)$ is negative so that the corresponding value of $\mathbf{v}_t^{(i)}$ is decreased. On the other hand, if e_i is smaller than eth, $\xi_i(\cdot)$ produces a positive value and $\mathbf{v}_t^{(i)}$ is increased. Therefore, during the recursion, in the validity vector, the components related to the feature points whose 2D motion trajectories can be well explained by the motion samples (with low epipolar distances) will have high positive values. For mismatched features or features on moving objects, since their motion trajectories can not be interpreted by resulting motion samples, their related components in $\mathbf{v}$ will have low negative values. Hence, the validity vector samples can be used to segment out features on the background.

Likelihood Function. In this case, the likelihood function of the observation given the state parameter is obtained as

$$f(\mathbf{y}_t \mid \mathbf{x}_t) \propto I_{\left\{\sum \frac{x(i)}{|x(i)|} \leq 7\right\}}(\mathbf{v}_t^+) \sum_{i=1}^{M} \mathbf{v}_t^+(i) \exp\left\{\frac{-\varepsilon}{\sigma_u^2 + \sigma_v^2}\right\} \tag{25}$$

$I_{\{f_b(x)\}}(x)$ is an indicator function. $f_b(x)$ is a boolean function of x. If $f_b(x)$ is true, $I_{\{f_b(x)\}}(x)$ returns one, otherwise zero. $\mathbf{v}_t^+$ carries the positive components of ν_t, given by

$$\mathbf{v}_t^+(i) = \begin{cases} \mathbf{v}_t(i), & \text{if } \mathbf{v}_t(i) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

Since at least seven feature points are required to uniquely determined the camera motion, samples which have less then seven positive entries in the validity vectors are considered bad motion samples and eliminated from the SIS procedure. ε can be viewed as a weighted average of the squared epipolar distances of the features with positive entries in the validity vector. It is given by

$$\varepsilon = \frac{\sum_{i=1}^{M} e_t(i)^2 \mathbf{v}_t^+(i)}{\sum_{i=1}^{M} \mathbf{v}_t^+(i)} \tag{27}$$

Based on the above discussion, the SIS method in the last section can be modified here by retaining the sampling step for the motion parameters, adding the sampling step for the validity vector according to (23), and computing the weights using (25). During the recursion, $\mathbf{v}_t$ will evolve to a structure such that the entries corresponding to the background have similar large positive values while the other entries corresponding to the mismatched feature points or those on moving objects have small or negative values. Once the validity vector for the feature points is obtained, clustering methods such as the k-means algorithm can be applied to split the features into features on the background and features that belong to moving objects.

Relation to the EM Algorithm. The Expectation-Maximization algorithm (EM) (Hartley, 1958; Dempster et al., 1977) is a practical method for find the maxima of the likelihood of the state parameters of a system when some system parameters are unknown or only partial measurements are observed. The EM algorithm has been used to tackle the SfM problem

without correspondences (Dellaert et al., 2000). It is also very possible that it can be used for camera motion estimation in the presence of independently moving objects. Actually, the SIS procedure we have presented here is very close to the EM algorithm in spirit. To solve the camera motion with moving objects, there are two tasks to be accomplished. The first one is to find the feature points on the background; the second one is to find the camera motion relative to the background. If the EM algorithm is used, the basic estimation scheme could be as follows. Starting from an initial guess of the feature set on the background, at the M-step, estimate the motion parameters using current feature set assumed to be on the background. Then, at the E-step, using the newly obtained motion estimates, the probabilities of possible feature sets are evaluated. Large feature set with small residuals will have high probability. The expectation of the likelihood of the motion parameters is computed over all possible feature sets on the background. This expected value is then maximized over motion parameters based on current feature point assignment. The EM iteration is repeated. In our approach, the motion estimation using samples and weights computed from (25) can be viewed as the M-step but with many motion samples. and the update of the validity vector using (23) is similar to the E-step, with the assignment of the feature points are represented by the samples of the validity vectors. Our method using SIS can be viewed as a recursive version of the EM algorithm with only one step of iteration at each recursion.

3. Estimation of Translation Magnitude and Scene Structure

Recall that in the previous section, we selected the SfM computational scheme where the camera motion is estimated at first and then the structure parameters are recovered. Therefore, only rotation and translation direction angles are used to describe the camera motion and the translation size is not represented. In many practical applications, not only is one interested in the direction of translation, but also its magnitude. Remember that the translation is measured in the world coordinate system I. To be consistent, the structure parameters are also defined in system I, which is, in our configuration, identical to the camera-centered system before the movement of camera. In this paper only monocular image sequences are used as observations. It is well known that in this case the translation size and feature points' depths can only be recovered up to a global scale factor, i.e. only their relative sizes can be estimated. Usually, one of the length quantity such as the depth values is chosen as the length basis. This length quantity is called the *normalization basis* with normalized length 1. Translation magnitude and depth values are normalized with respect to the normalization basis. In our approach, we select the depth of the last feature point, z_M, as the normalization basis for translation magnitude and the depths of other feature points.

In this section, we show that the posterior distribution of translation sizes and the depths of the feature points can be approximated by a set of samples that are properly weighted by their weights w.r.t. the target distributions.

3.1. Estimation of Translation Magnitude

Denote the translation magnitude at time t by γ_t. Let the sequence of the translation magnitude up to current time t be $\Gamma_t = \{\gamma_\tau\}_{\tau=1}^t$. Recall that by using the SIS procedure for camera motion estimation described in the last section, $p(\mathcal{X}_t \mid \mathcal{Y}_t)$ can be described by properly weighted samples. In this section, we would like to find samples and weights describing the joint distribution $p(\Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)$. We have the following theorem.

Theorem 1. *Let r and s be two random variables with joint distribution $p(r, s)$. Assume that $\{s^{(i)}\}_{i=1}^m$ are properly weighted by normalized weights $\{w_s^{(i)}\}_{i=1}^m$ w.r.t. the marginal distribution $p(s)$ with $m \gg 1$.* ("*Normalized weights*" *means that the summation of the weights is* 1.) *For each sample of s, say, $s^{(i)}$, assume that $\{r_i^{(j)}\}_{j=1}^n$ are properly weighted by normalized weights $\{w_i^{(j)}\}_{j=1}^n$ w.r.t. the conditional distribution $p(r \mid s^{(i)})$ with $n \gg 1$. Then the combined samples $\{\{r_i^{(j)}, s^{(i)}\}_{j=1}^n\}_{i=1}^m$ are properly weighted by $\{\{w_i^{(j)} w_s^{(i)}\}_{j=1}^n\}_{i=1}^m$ w.r.t. the joint distribution $p(r, s)$.*

Proof: Let $h(r, s)$ be an integrable function of r and s.

$$Eh(r,s) = \int h(r,s)p(r,s)\,dr\,ds \tag{28}$$

$$= \int_s \int_r h(r,s)p(r \mid s)\,dr\,p(s)\,ds \tag{29}$$

Since $\{s^{(i)}\}_{i=1}^m$ are properly weighted samples w.r.t. $p(s)$, (29) can be written as

$$Eh(r,s) = \lim_{m\to\infty} \sum_{i=1}^{m} \int_r h\big(r, s^{(i)}\big)\, p\big(r \mid s^{(i)}\big)\, dr w_s^{(i)} \quad (30)$$

which can be approximated using sample-weight pairs properly weighted w.r.t. $p(r \mid s^{(i)})$ for $i = 1, \ldots, m$ as

$$Eh(r,s) = \lim_{m\to\infty} \sum_{i=1}^{m} w_s^{(i)} \lim_{n\to\infty} \sum_{j=1}^{n} h\big(r_i^{(j)}, s^{(i)}\big) w_i^{(j)} \quad (31)$$

$$= \lim_{m\to\infty} \sum_{i=1}^{m} \lim_{n\to\infty} \sum_{j=1}^{n} h\big(r_i^{(j)}, s^{(i)}\big) w_i^{(j)} w_s^{(i)} \quad (32)$$

Therefore, the combined samples $\{\{r_i^{(j)}, s^{(i)}\}_{j=1}^n\}_{i=1}^m$ are properly weighted by their respective weights $\{\{w_i^{(j)} w_s^{(i)}\}_{j=1}^n\}_{i=1}^m$ w.r.t. the joint distribution $p(r,s)$. □

The joint posterior distribution of Γ_t and $\mathcal{X}_t$ can be written as

$$p(\Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t) = p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) p(\mathcal{X}_t \mid \mathcal{Y}_t)$$

Samples and weights that describe $p(\mathcal{X}_t \mid \mathcal{Y}_t)$ can be computed easily using the SIS motion algorithm we proposed. If for each of the sample, say, $\mathcal{X}_t^{(j)}$, we can find samples of translation magnitude properly weighted w.r.t. the conditional distribution $p(\Gamma_t \mid \mathcal{X}_t^{(j)}, \mathcal{Y}_t)$, then the samples and weights w.r.t. their joint distribution can be formed using these two categories of sample-weight pairs based on Theorem 1. The samples of $p(\mathcal{X}_t \mid \mathcal{Y}_t)$ are obtained recursively during the SIS procedure. Basically, new sample stream $\mathcal{X}_t^{(j)}$ are formed by attaching the newly drawn sample $\mathbf{x}_t^{(j)}$ to the old stream $\mathcal{X}_{t-1}^{(j)}$, i.e. $\mathcal{X}_t^{(j)} = (\mathcal{X}_{t-1}^{(j)}, \mathbf{x}_t^{(j)})$. The weights are also updated accordingly. Therefore, we desire to have a recursive sampling scheme for the complete motion distribution $p(\Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)$. It is easy to show that

$$p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) \approx p(\Gamma_{t-1} \mid \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) p(\gamma_t \mid \gamma_{t-1}) \times \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)}{p(\mathbf{y}_t \mid \mathbf{x}_t)} \quad (33)$$

which indicates that we can recursively get samples of $p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t)$. The proof can be found in the Appendix. The following proposition shows that the above conditional distribution can be simplified.

Proposition 1. *Assume that M feature points are tracked through an image sequence. Without loss of generality, the depth of the feature point, z_M, is selected as the normalization basis for translation magnitude and structure parameters. Let $\mathbf{Y}_t$ be its 2D trajectories in the image plane up to t. Then the conditional distribution $p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t)$ is completely determined only by $\mathbf{Y}_t$, namely,*

$$p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) = p(\Gamma_t \mid \mathcal{X}_t, \mathbf{Y}_t) \quad (34)$$

Proof: Let $\mathbf{z}$ and $\tilde{\mathbf{z}}$ be feature depth vector and the normalized depth, respectively. Since z_M is the normalization basis, we have

$$\tilde{\mathbf{z}} = \frac{\mathbf{z}}{z_M}$$

Let γ_t be the normalized translation magnitude at time t. The representation of depths of the first $M-1$ points can be reparameterized using γ_t. Denote the relative depth by λ_t. Let's call λ_t the translation-normalized depth vector.

$$\lambda_t = \frac{\tilde{\mathbf{z}}^C}{\gamma_t} \quad (35)$$

where $\tilde{\mathbf{z}}^C$ is the truncated depth vector, containing the normalized depth values of the first $M-1$ points. Although $\tilde{\mathbf{z}}^C$ could be fixed when the scene is rigid, usually λ_t is time-varying since γ_t changes. Let Λ_t be the sequence of the translation-normalized depth vector up to t, i.e. $\Lambda_t = \{\lambda_\tau\}_{\tau=1}^t$. Denote the complementary set of $\mathbf{Y}_t$ by $\mathcal{Y}_t^C$ such that $(\mathcal{Y}_t^C, \mathbf{Y}_t) = \mathcal{Y}_t$. The original conditional distribution can be written as

$$p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) = \int_{\Lambda_t} p\Big(\Gamma_t, \Lambda_t \,\Big|\, \mathcal{X}_t, \mathcal{Y}_t^C, \mathbf{Y}_t\Big) d\Lambda_t$$

$$= \int_{\Lambda_t} \frac{p\Big(\mathcal{Y}_t^C \,\Big|\, \mathcal{X}_t, \Gamma_t, \Lambda_t\Big) p(\Gamma_t \mid \mathbf{Y}_t, \mathcal{X}_t) p(\Lambda_t \mid \mathcal{X}_t) p(\mathbf{Y}_t, \mathcal{X}_t)}{p(\mathcal{Y}_t^C \mid \mathbf{Y}_t, \mathcal{X}_t) p(\mathbf{Y}_t, \mathcal{X}_t)} d\Lambda_t$$

Since Λ_t is translation-normalized, we have $p(\mathcal{Y}_t^C \mid \mathcal{X}_t, \Gamma_t, \Lambda_t) = p(\mathcal{Y}_t^C \mid \mathcal{X}_t, \Lambda_t)$. Hence

$$p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) = \int_{\Lambda_t} \frac{p\big(\mathcal{Y}_t^C \,\big|\, \mathcal{X}_t, \Lambda_t\big) p(\Gamma_t \mid \mathbf{Y}_t, \mathcal{X}_t) p(\Lambda_t \mid \mathcal{X}_t)}{p\big(\mathcal{Y}_t^C \,\big|\, \mathcal{X}_t\big)} d\Lambda_t$$

$$= p(\Gamma_t \mid \mathbf{Y}_t, \mathcal{X}_t) \int_{\Lambda_t} \frac{p(\mathcal{Y}_t^C, \Lambda_t \mid \mathcal{X}_t)}{p(\mathcal{Y}_t^C \mid \mathcal{X}_t)} d\Lambda_t$$
$$= p(\Gamma_t \mid \mathbf{Y}_t, \mathcal{X}_t)$$

□

The related recursion formula for the simplified conditional distribution is

$$p(\Gamma_t \mid \mathcal{X}_t, \mathbf{Y}_t) \approx p(\Gamma_{t-1} \mid \mathcal{X}_{t-1}, \mathbf{Y}_{t-1}) p(\gamma_t \mid \gamma_{t-1}) \times \frac{p(y_t \mid \gamma_t, \mathbf{x}_t)}{p(y_t \mid \mathbf{x}_t)} \quad (36)$$

where y_t is the image plane position of the last feature point at time t. Based on the above discussion, we present the following SIS procedure for finding an approximation to the joint distribution $p(\Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)$.

SIS Procedure for Joint Motion Distribution

1. *Initialization.* At $t = 1$,
   - *Motion Sample Acquisition.* Compute the properly weighted samples $\{\mathbf{x}_1^{(j)}\}_{j=1}^m$ and their normalized weights $\{w_{x,1}^{(j)}\}_{j=1}^m$ w.r.t. $p(\mathbf{x}_1 \mid \mathcal{Y}_1)$ using the SIS procedure for camera motion distribution developed in the last section. Let $\mathcal{X}_1 = \{\mathbf{x}_1^{(j)}\}$ and $\mathcal{W}_{x,1} = \{w_{x,1}^{(j)}\}$.
   - *Samples of Translation Magnitude.*

     For each motion sample $\mathbf{x}^{(j)}$, $j = 1, \ldots, m$

     - Compute $\gamma^{*(j)}$ using $\mathbf{x}^{(j)}$ as

       $$\gamma^{*(j)} = \arg\max_{\gamma} p\big(y_1 \mid \gamma, \mathbf{x}^{(j)}\big) p(\gamma) \quad (37)$$

       Assume that the magnitude of the translation can be uniformly distributed in the domain of positive real numbers.
     - Draw a set of samples $\{\gamma_j^{(k)}\}$ around $\gamma^{*(j)}$ from a trial distribution

       $$g\big(\gamma \mid \gamma^{*(j)}\big) = \mathcal{N}\big(\gamma^{*(j)}, \Sigma_\gamma\big) \quad (38)$$

       This trial distribution can also be of other forms.
     - Evaluate weights for samples $\{\gamma_j^{(k)}\}$. Let $w_j^{(k)}$ be the weight of $\gamma_j^{(k)}$.

       $$w_j^{(k)} \propto \frac{p\big(y_1 \mid \gamma_j^{(k)}, \mathbf{x}^{(j)}\big) p\big(\gamma_j^{(k)}\big)}{g\big(\gamma_j^{(k)} \mid \gamma^{*(j)}\big) p\big(y_1 \mid \mathbf{x}^{(j)}\big)} \quad (39)$$

     - Normalize the weights by

       $$\tilde{w}_j^{(k)} = \frac{w_j^{(k)}}{\sum_k w_j^{(k)}} \quad (40)$$

     - Combine samples and weights. $\Upsilon_1^{(j)} = \{\gamma_j^{(k)}, \mathbf{x}^{(j)}\}_{k=1}^n$ and $\mathcal{W}_{\gamma,1}^{(j)} = \{\tilde{w}_j^{(k)} w_{x,1}^{(j)}\}_{k=1}^n$. Note that γ_j's in $\Upsilon_1^{(j)}$ and $\mathcal{W}_{\gamma,1}^{(j)}$ are approximately properly weighted w.r.t. the conditional distribution $p(\Gamma_1 \mid \mathcal{X}_1^{(j)}, \mathbf{Y}_1)$.
   - *Samples and Weights Collection.* According to Theorem 1, the collection of the above combined samples, $\{\Upsilon_1^{(j)}\}_{j=1}^m$ is approximately properly weighted by the collection of the related weights $\{\mathcal{W}_{\gamma,1}^{(j)}\}_{j=1}^m$ w.r.t. the joint distribution $p(\Gamma_1, \mathcal{X}_1 \mid \mathcal{Y}_1)$.

   For time instant $t > 1$.
2. *Motion Sample Acquisition.* Based on the motion sequence samples up to previous time, $\{\mathcal{X}_{t-1}^{(j)}\}_{j=1}^m$ and $\{w_{x,t-1}^{(j)}\}_{j=1}^m$, properly weighted samples $\{\mathcal{X}_t^{(j)}\}_{j=1}^m$ and their normalized weights $\{w_{x,t}^{(j)}\}_{j=1}^m$ w.r.t. $p(\mathcal{X}_t \mid \mathcal{Y}_t)$ can be obtained using the SIS procedure for motion distribution. A motion sequence sample up to current time, $\mathcal{X}_t^{(j)}$ is formed by adding motion sample of current time to the old motion sequence sample, $\mathcal{X}_t^{(j)} = (\mathcal{X}_{t-1}^{(j)}, \mathbf{x}_t^{(j)})$.

   For each motion sequence sample $\mathcal{X}^{(j)}$, $j = 1, \ldots, m$

3. *Translation Magnitude Sample Generation.*
   - Obtain samples of Γ_{t-1} by resampling $\Upsilon_{t-1}^{(j)}$ according to their weights $\mathcal{W}_{\gamma,1}^{(j)}$. Denote the newly drawn samples by $\{\Gamma_{t-1,j}^{(k)}\}$.
   - Predict translation magnitude of current time by sampling from the Markovian transition probability $p(\gamma_t \mid \gamma_{t-1})$. Let $\gamma_{t,j}^{(k)} = |\gamma_{t-1,j}^{(k)} + n_\gamma|$ with n_γ a Gaussian variable with zero mean and variance σ_γ. Attach it to $\Gamma_{t-1,j}^{(k)}$ to form $\Gamma_{t,j}^{(k)} = (\Gamma_{t-1,j}^{(k)}, \gamma_{t,j}^{(k)})$.
   - For each translation magnitude sample $\gamma_t^{(k)}$, compute its corresponding weight.

     $$w_j^{(k)} = \frac{p\big(y_t \mid \mathbf{x}_t^{(j)}, \gamma_t^{(k)}\big)}{p\big(y_t \mid \mathbf{x}_t^{(j)}\big)} \quad (41)$$

     where the $p(y_t \mid \mathbf{x}_t^{(j)}, \gamma_t^{(k)})$ and $p(y_t \mid \mathbf{x}_t^{(j)})$ can be easily computed using likelihood functions (11)

and (14). Note that the depth of the feature point in current measurements, (the trajectories of the last feature point) is 1.

– Normalize the weights by

$$\tilde{w}_j^{(k)} = \frac{w_j^{(k)}}{\sum_k w_j^{(k)}} \tag{42}$$

– Combine samples and weights. $\Upsilon_t^{(j)} = \{\Gamma_{t,j}^{(k)}, \mathcal{X}_t^{(j)}\}_{k=1}^n$ and $\mathcal{W}_{\gamma,t}^{(j)} = \{\tilde{w}_j^{(k)} \cdot w_{x,t}^{(j)}\}_{k=1}^n$. Note that $\Gamma_{t,j}$'s in $\Upsilon_t^{(j)}$ and $\mathcal{W}_{\gamma,t}^{(j)}$ are properly weighted w.r.t. the conditional distribution $p(\Gamma_t \mid \mathcal{X}_t^{(j)}, \mathbf{Y}_t)$. From Proposition 1, these samples and weights are also approximately properly weighted w.r.t. $p(\Gamma_t \mid \mathcal{X}_t^{(j)}, \mathcal{Y}_t)$.

– According to Theorem 1, the collection of the above combined samples, $\{\Upsilon_t^{(j)}\}_{j=1}^m$ is approximately properly weighted by the collection of the related weights $\{\mathcal{W}_{\gamma,t}^{(j)}\}_{j=1}^m$ w.r.t. the joint distribution $p(\Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)$. When data of next time instant is available, go to step 2.

By using the above SIS procedure, the joint camera motion distribution $p(\Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)$ can be approximated using weighted samples.

3.2. Depth Estimation

Assuming the scene is static, the 3D positions of the feature points in the world system I are constant. We denote $\mathbf{z}$ by the constant depth values of the feature points. Using the joint camera motion parameters, the conditional distribution of the feature depth values can be obtained. Then, Theorem 1 can be used to obtain joint structure and motion samples. The following theorem states the recursive relationship of the conditional depth distribution given joint motion parameters.

Theorem 2. *Given the joint camera motion distribution, the recursion equation for the conditional depth distribution of a set of static 3D points can be approximated as follows.*

$$p(\mathbf{z} \mid \Gamma_t, \mathcal{X}_t, \mathcal{Y}_t) \approx p(\mathbf{z} \mid \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) \times \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z})}{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)} \tag{43}$$

Proof:

$$\begin{aligned} & p(\mathbf{z} \mid \Gamma_t, \mathcal{X}_t, \mathcal{Y}_t) \\ &= \frac{p(\Gamma_t, \mathcal{X}_t, \mathcal{Y}_t, \mathbf{z})}{p(\Gamma_t, \mathcal{X}_t, \mathcal{Y}_t)} \\ &= \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}) p(\gamma_t \mid \gamma_{t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{z}, \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1})}{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t) p(\gamma_t \mid \gamma_{t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1})} \\ &= \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}) p(\mathbf{z}, \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1})}{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t) p(\Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1})} \\ &= p(\mathbf{z} \mid \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z})}{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)} \end{aligned}$$

□

Using the joint motion distribution, which can be obtained using the SIS procedure discussed in the previous section, both the joint structure/motion $p(\mathbf{z}, \Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)$ and marginal depth $p(\mathbf{z} \mid \mathcal{Y}_t)$ distributions are easy to compute.

$$p(\mathbf{z} \mid \mathcal{Y}_t) = E_{\Gamma_t, \mathcal{X}_t}\{p\{\mathbf{z}, \Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t\} \tag{44}$$

$$= \int_{\Gamma_t, \mathcal{X}_t} p(\mathbf{z}, \Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t)\, d\Gamma_t d\mathcal{X}_t \tag{45}$$

Moreover, an approximate recursive form of the marginal depth distribution can be easily derived.

$$\begin{aligned} p(\mathbf{z} \mid \mathcal{Y}_t) \approx & \frac{p(\mathcal{Y}_{t-1})}{p(\mathcal{Y}_t)} p(\mathbf{z} \mid \mathcal{Y}_{t-1}) \\ & \times \int_{\gamma_t, \mathbf{x}_t} p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}) p(\gamma_t \mid \gamma_{t-1}) \\ & \times p(\mathbf{x}_t \mid \mathbf{x}_{t-1})\, d\gamma_t d\mathbf{x}_t \end{aligned} \tag{46}$$

It is not difficult to derive Eq. (46) by using Theorem 2 and Eq. (33). The proof can be found in the Appendix. Although Eq. (46) can be used to find the weighted samples for marginalized depth distribution, we are more interested in obtaining weighted sample for the joint structure/motion distribution. To achieve this goal, we first use the recursive equation of depth distribution given by Theorem 2 and straightforwardly obtain samples and weights properly weighted w.r.t. $p(\mathbf{z} \mid \Gamma_t, \mathcal{X}_t, \mathcal{Y}_t)$, based on depth samples w.r.t. the previous conditional distribution $p(\mathbf{z} \mid \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1})$. Then, the samples w.r.t. the joint structure/motion distribution can be formed using Theorem 1. The depth distribution $p(\mathbf{z} \mid \mathcal{Y}_t)$ can be easily obtained by marginalization of the joint structure/motion distribution. Based on the above discussion, we have the following SIS procedure for recursively obtaining weighted samples for the joint motion and structure distribution.

SIS Procedure for Joint Motion/Structure Distribution

1. *Initialization.* At $t = 1$,

   – *Obtain Samples and Weights w.r.t.* $p(\gamma_1, \mathbf{x}_1 \mid \mathcal{Y}_1)$. Compute the properly weighted samples

$\{\gamma_1^{(j)}, \mathbf{x}_1^{(j)}\}$ and their normalized weights $\{w_m^{(j)}\}$ w.r.t. $p(\gamma_1, \mathbf{x}_1 \mid \mathcal{Y}_1)$ using the SIS procedure for joint camera motion distribution developed in the previous section.

– *Obtain Samples and Weights w.r.t.* $p(\mathbf{z} \mid \gamma_1, \mathbf{x}_1, \mathcal{Y}_1)$.

- For each motion samples $\{\gamma_1^{(j)}, \mathbf{x}_1^{(j)}\}$, compute $\mathbf{z}^{*(j)}$, the *maximum likelihood depth* (MLD), as follows.

$$\mathbf{z}^{*(j)} = \arg\max_{\mathbf{z}} p\big(\mathbf{y}_t \mid \mathbf{z}, \gamma^{(j)}, \mathbf{x}^{(j)}\big)\pi(\mathbf{z}) \quad (47)$$

- Draw a set of samples $Z_1^{(j)} = \{\mathbf{z}^{(j,k)}\}$ around $\mathbf{z}^{*(j)}$ from a trial distribution

$$g\big(\mathbf{z} \mid \mathbf{z}^{*(j)}\big) = \mathcal{N}\big(\mathbf{z}^{*(j)}, \Sigma\big) \quad (48)$$

This trial distribution can also be of other forms.

- Evaluate weights for samples in $Z_1^{(j)}$. Let $w^{(j,k)}$ be the weight of $\mathbf{z}^{(j,k)}$.

$$w^{(j,k)} \propto \frac{p\big(\mathbf{y}_t \mid \mathbf{z}^{(j,k)}, \gamma_1^{(j)}, \mathbf{x}_1^{(j)}\big)\pi\big(\mathbf{z}^{(j,k)}\big)}{g\big(\mathbf{z}^{(j,k)} \mid \mathbf{z}^{*(j)}\big)} \quad (49)$$

- Normalize the weights by

$$\tilde{w}^{(j,k)} = \frac{w^{(j,k)}}{\sum_k w^{(j,k)}} \quad (50)$$

Let $W_1^{(j)} = \{\tilde{w}^{(j,k)}\}_k$. According to Proposition 2 in Appendix, $(Z_1^{(j)}, W_1^{(j)})$ are properly weighted w.r.t to $p(\mathbf{z} \mid \gamma_1^{(j)}, \mathbf{x}_1^{(j)}, \mathcal{Y}_1)$

– Joint motion and depth samples can then be easily obtained by proper combination of joint motion and conditional depth samples, using Theorem 1. Therefore, the joint motion samples, conditional depth samples and joint motion/depth samples are all ready for the initial time instant.

For time instant $t > 1$

2. *Motion Sample Acquisition.* Compute the approximately properly weighted samples $\{\Gamma_t^{(j)}, \mathcal{X}_t^{(j)}\}_{j=1}^m$ and their normalized weights $\{w_m^{(j)}\}_{j=1}^m$ w.r.t. $p(\Gamma, \mathcal{X}_t \mid \mathcal{Y}_t)$ using the SIS procedure for joint motion distribution.
3. *Depth Sample and Weight Update*

– *Depth resampling.* For each motion samples $\{\Gamma_t^{(j)}, \mathcal{X}_t^{(j)}\}$, resample the previous depth sample set $Z_{t-1}^{(j)}$, w.r.t. $p(\mathbf{z} \mid \Gamma_{t-1}^{(j)}, \mathcal{X}_{t-1}^{(j)}, \mathcal{Y}_{t-1})$, according to their weights W_{t-1}. Denote the newly drawn samples by $Z_t^{(j)}$.

– *Weight update.* For each depth sample $z_t^{(k)}$ in $Z_t^{(j)}$, compute its corresponding weight using current motion samples.

$$w^{(j,k)} = p\big(\mathbf{y}_t \mid \gamma_t^{(j)}, \mathbf{x}_t^{(j)}, \mathbf{z}^{(k)}\big) \quad (51)$$

Let $W_t^{(j)}$ be $\{w^{(j,k)}\}_k$. Depth sample set $Z_t^{(j)}$ is properly weighted by $W_t^{(j)}$ w.r.t. $p(\mathbf{z} \mid \Gamma_t, \mathcal{X}_t, \mathcal{Y}_t)$.

Using Theorem 1, the joint motion and depth samples can then be easily obtained. When data of next time instant is available, go to step 2.

In the above SIS procedure, the depth distribution of individual feature point can be computed in parallel because the depth values of different feature points are statistically independent. The complete structure and motion estimation algorithm can be illustrated by Fig. 1. After the discussion of three SIS procedures for partial camera motion, translation magnitude and scene structure, we can see that the complete camera motion and scene structure estimates from a video sequence are obtained progressively in dimension and recursively in time.

4. Experimental Results and Performance Analysis

By using the above SIS methods, challenging issues in SfM problem such as occlusion, SfM ambiguity, mixed-domain sequence processing, and mismatched feature points can be elegantly handled.

4.1. A Case Study

We first show some results using synthetic sequences. In the first example, a sequence with 13 feature points was generated. The size of the images in pixels is 512×512 and the field of view is 0.9237 radian. The baseline of the camera motion is not large. The translation magnitude ranges from 1 to 35 percents of the closed feature depth. The feature trajectories were corrupted by additive white Gaussian noise (AWGN) with standard deviation (STD) 0.5 pixel. During the camera motion computation, 5000 samples were used in the SIS procedure. The resultant marginalized motion distributions are shown in Fig. 4. The ground truths of camera motion are indicated by thick solid lines. In

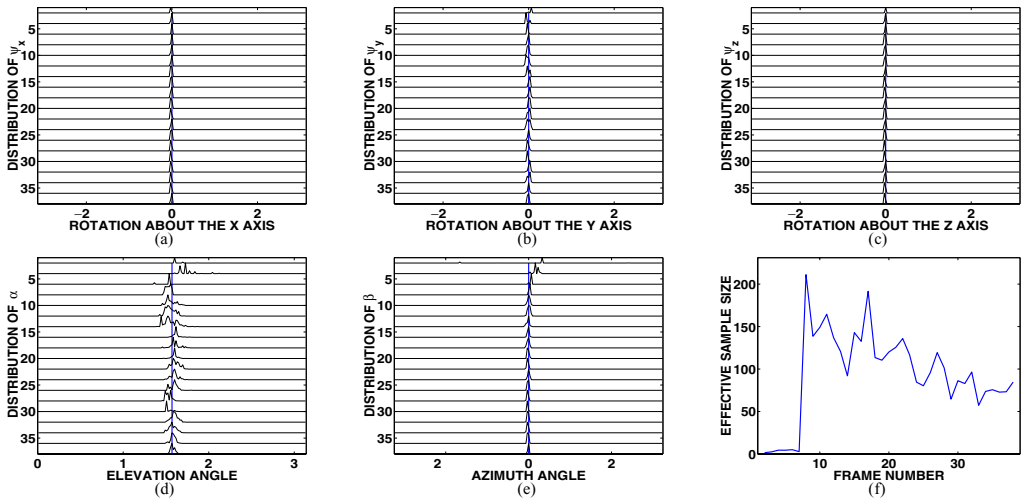

Figure 4. Camera motion estimation results in the case study. (a)–(e) are the posterior distributions of the camera motion, (f) shows the effective sample sizes of during the SIS procedure for camera motion.

Fig. 4 (a)–(c) in the first row are the respective distributions of rotational angles ψ_x, ψ_y and ψ_z. The first two figures in the second row show the distributions of translation direction angles α and β. In each figure, the distribution of the corresponding motion parameter at each time instant is shown from the top of the figure to the bottom with Y being the time instant and X being the value of each motion parameter. (ψ_x, ψ_y, ψ_z) are in the range $[-\pi, \pi]$. α is in $[0, \pi]$ and β in $[0, 2\pi]$. All the other motion distribution results in this paper can be interpreted in the same way. We see that the distributions of motion parameters have peaks very close to the ground truths. Figure 4(f) presents the effective sample sizes of the SIS procedure for motion distribution over the entire sequence. ESS at each time instant is computed using (2) from related sample weights.

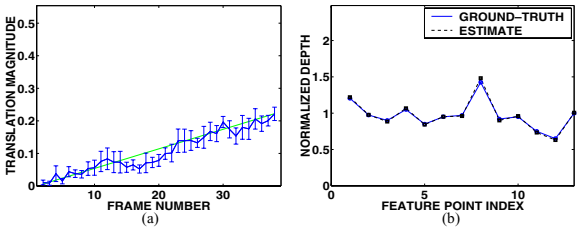

Figure 5. Camera translation magnitudes and feature point depth estimates in the case study. (a) shows the empirical means and variances of the translation magnitudes at different time instants, the thicker lines indicating the ground-truth, (b) shows the empirical mean of the depth samples from the two SIS procedures with (the diamonds) and without (the squares) scene rigidity constraint. The ground-truth is marked by circles.

Figure 5(a) shows the empirical mean and variance of translation magnitude at different time instant throughout the sequence. The ground-truth is shown by the dotted line. Figure 5(b) shows the empirical mean of the depth distribution. The ground-truth of feature depth is marked by circles. Since we developed two SIS procedures for depth estimation with and without using scene rigidity constraint, two set of samples were obtained and their empirical mean are shown here. Estimates with and without scene rigidity constraint are marked by diamonds and squares, respectively. Since both estimates are very close to the ground truth, it is difficult to distinguish one from the other. Figure 6 shows the means and variances of depth distribution using the SIS procedure for static scenes. We see that in the beginning the estimates are not accurate and the variances are large, but as the translation magnitude accumulates, the depth estimates converge to the ground-truth and the variances also reduce.

4.2. *Feature Occlusion*

During the tracking of feature points through a video sequence, some of the feature points may be occluded by objects over in some frames. The occluded feature points are unobservable during that period of time, but they may become visible in subsequent frames. In Tomasi and Kanade's factorization method (Tomasi and Kanade, 1992), the image positions of the occluded feature points are interpolated from their past and future positions. In our approach, feature occlusion does

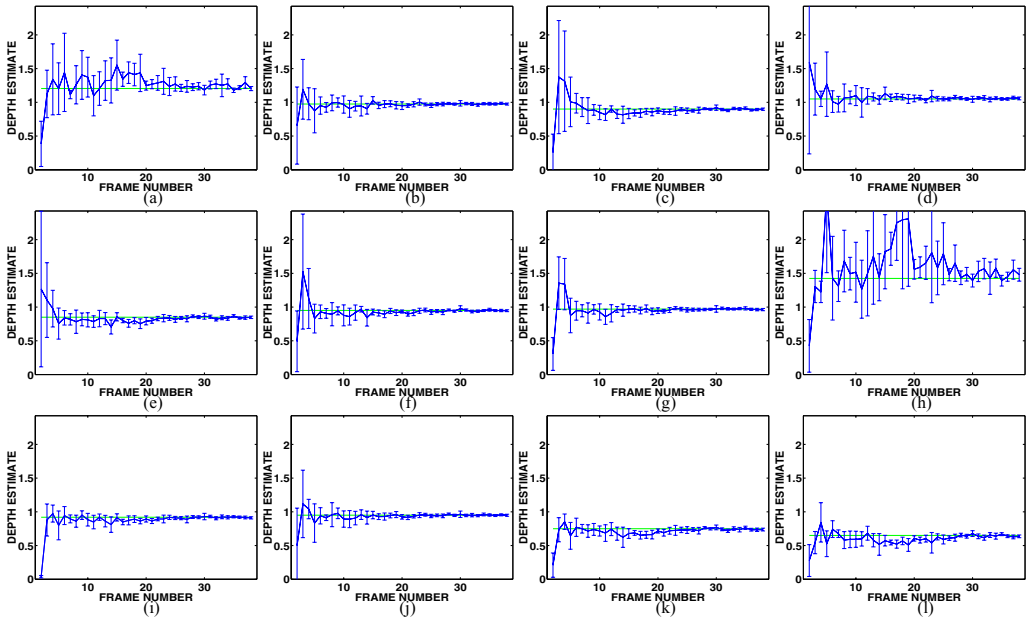

Figure 6. Posterior depth distributions in the case study.

not affect the basic theory of our algorithm. At each time instant, the occluded feature points are ignored when the weights are computed and only visible feature points are used as observations. The reduction in the number of observed feature points may make the motion estimation less accurate in the sense that the resulting posterior distribution has a larger variance than when more observations are used.

4.3. Motion and Structure Ambiguities

The ambiguities in motion and structure recovery have been noticed to be inherent (Adiv, 1989; Young and Chellappa, 1992). Although the ambiguities can be reduced by using rate data obtained from inertial sensors (Qian et al., 2001), inertial sensors are not widely applied yet and not many sequences were captured using

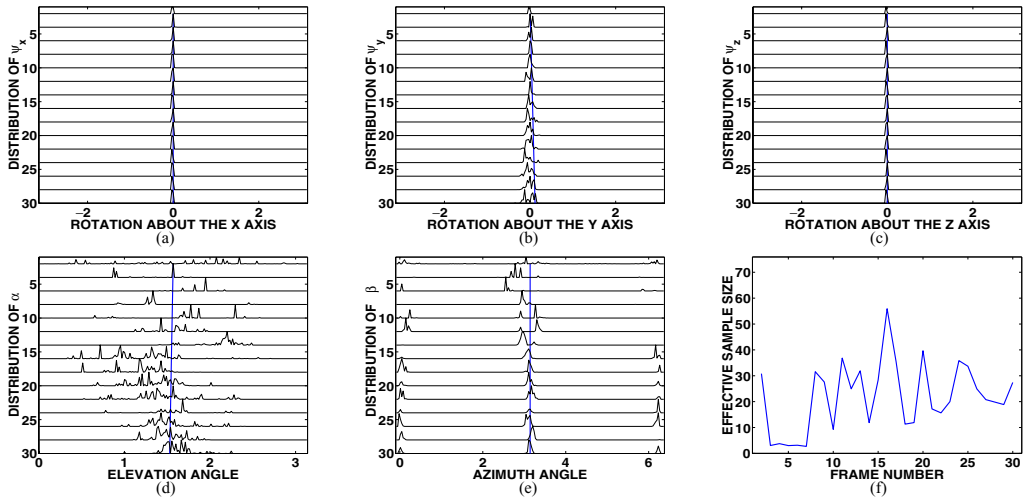

Figure 7. Posterior distributions of the motion parameters in the structure/motion ambiguity case. (f) shows the ESS in this case.

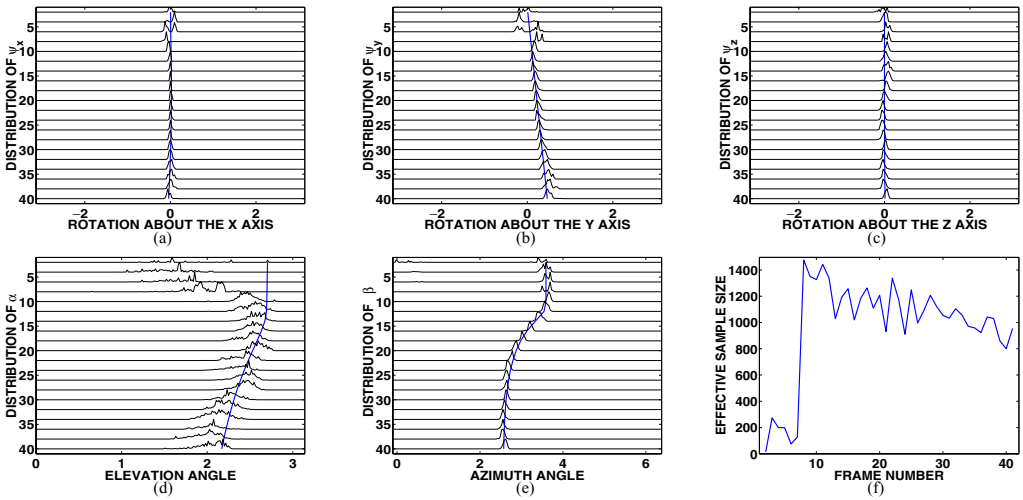

Figure 8. Results of processing a mixed-domain sequence. (a)–(e) are the posterior distributions of the camera motion, (f) shows the ESS in this case.

cameras with inertial sensors. One dominant ambiguity arises from the similarity between the feature trajectories generated by translation parallel to the image plane and out-plane rotation (Daniilidis and Nagel, 1993) when the size of the field of view is small. It is usually referred as *translation-rotation confusion*. Given feature correspondences tracked through an image sequence, two questions need to be answered. The first one is whether SfM ambiguities exist for this sequence. The second one is: if it exists, what are the admissible solutions? Using our SIS-based approach, both questions can be answered by looking at samples and weights describing the posterior distribution of the motion parameters. For feature trajectories obtained from an ambiguous image sequence, multiple modes can be found in distributions represented by the samples and weights. This answers the first question. Admissible solutions can be found by locating the modes of the distributions. in the solution space. The results shown in Fig. 7 are obtained using an ambiguous sequence. During the generation of the sequence, the camera moved downward along the vertical axis of the image plane and simultaneously rotated about the horizontal axis. The field of view of the camera is only 0.2 radian. This is a typical set-up for producing the ambiguity of translation-rotation confusion. We observe in Fig. 7(e) that at least two different peaks exist in the distribution of the translational direction angle β, one near 0 and the other near π. These two modes represent two possible translation direction along the image plane: the

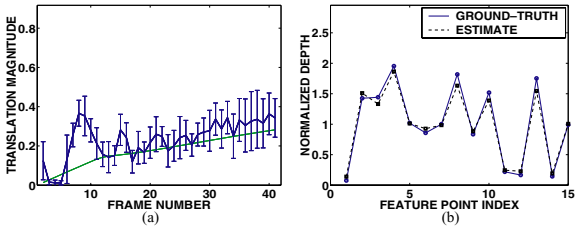

Figure 9. Camera translation magnitudes and feature point depth estimates for a mixed-domain sequence. (a) shows the empirical means and variances of the translation magnitudes at different time instants, the thicker lines indicating the ground-truth, (b) shows the empirical mean of the depth samples from the two SIS procedures with (the diamonds) and without (the squares) scene rigidity constraint. The ground-truth is marked by circles.

mode near π indicates downward motion and the other upward. The distributions of rotation about the horizontal axis of the image plane is shown in (b), which bear multi-mode patterns and large variances.

4.4. *Mixed-Domain Sequence Processing*

The processing of mixed-domain sequences was discussed by Oliensis (2000) as a challenge for SfM algorithms. Due to the constraints about camera translation (small or large) implicitly or explicitly assumed by researchers when SfM algorithms were designed, many existing SfM algorithms do not work well for mixed-domain sequences. In our approach, since no

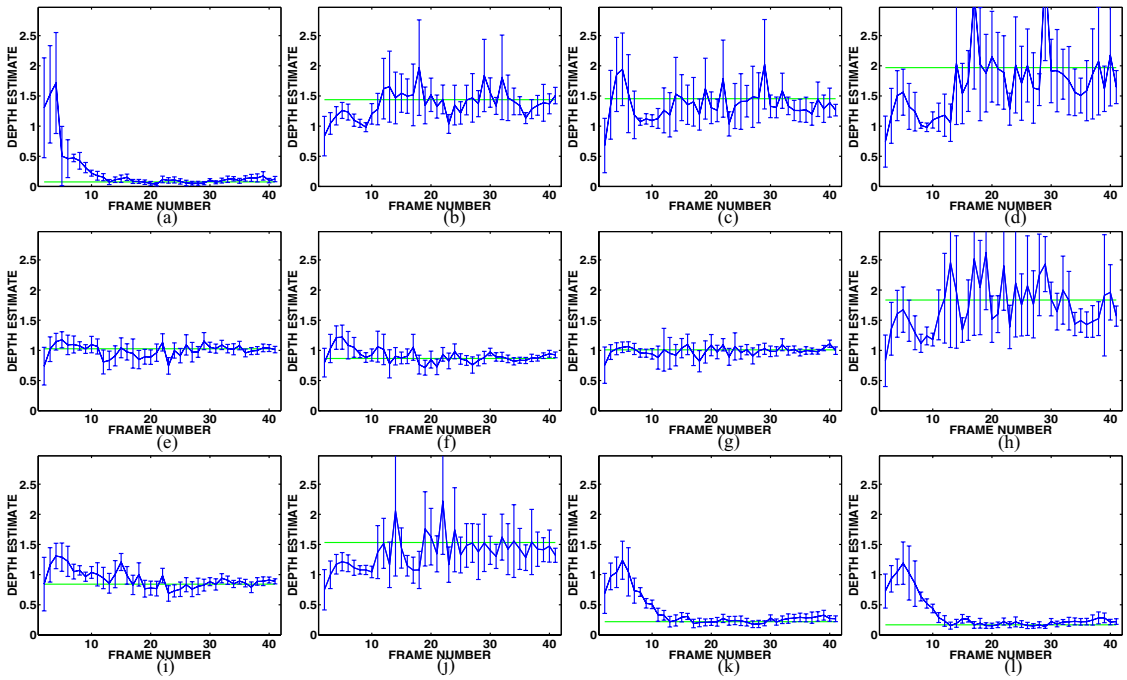

Figure 10. Posterior depth distributions in the mixed-domain sequence case.

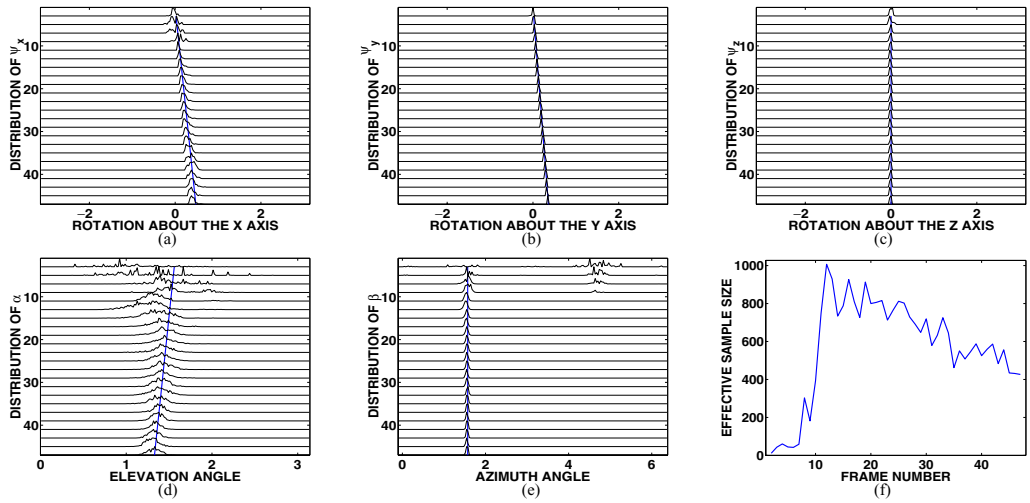

Figure 11. Posterior motion distributions when mismatched feature points present. (f) shows the ESS in this case.

assumptions about translation are made, mixed-domain sequences can be correctly processed. In this example, a sequence containing both small and large translation was used. The camera translation magnitude ranges from 0.2 to 3.5 times of the depth of the nearest feature point. Therefore, this sequence has both small and large camera translation. The pixel size of the image frame in this example is 256×256. The feature point correspondences were contaminated by AWGN with zero mean and two pixels of STD. In the beginning, the signal-to-noise-ratio (SNR) of the feature correspondences is low because of small camera motion, it is very easy for EKF-based algorithm to converge to false solutions. Since in these algorithms, only one

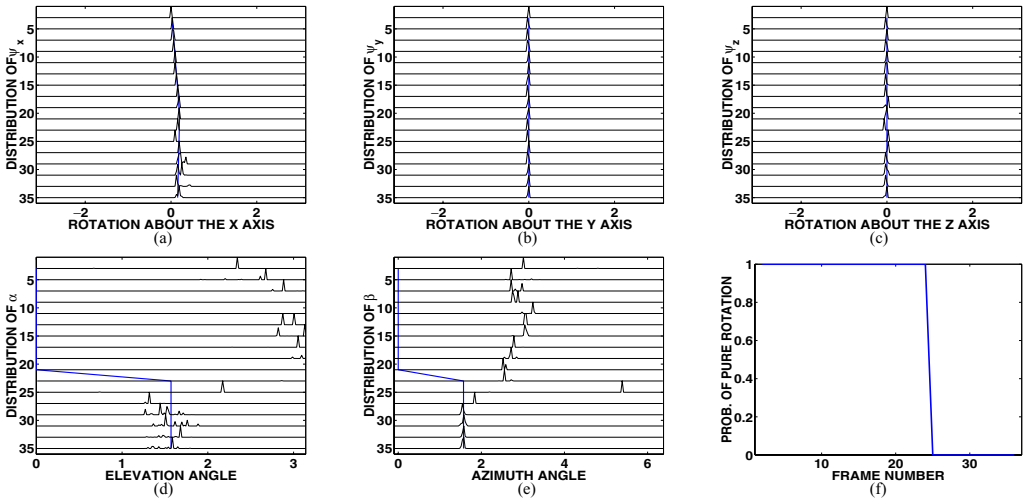

Figure 12. Posterior motion distributions when the camera first only rotates. (f) shows P(*pure rotation*) of this sequence.

"optimal" solution is kept in the recursion, they will be trapped in the false solution. However, in our approach, as the posterior distribution of camera motion is approximated by samples and weights, good samples producing small residual errors will be kept in the sample set. When the SNR is low, motion samples both close to and far from to true camera motion are kept. When the camera moves further, the SNR is increased and samples far from the true motion produce large residual errors hence less weights are assigned to them and gradually they vanish from the SIS procedure while good samples close to the true motion remain in the sample set and keep tracking the camera motion. That is the basic reason why our approach can handle a wide range of image sequences regardless the types of camera motion and scene structure. The results of processing this mixed-domain sequence are shown in Fig. 8. The camera translation and depth estimates are given in Fig. 9. The corresponding depth estimation results are shown in Fig. 10. The feature tracking noises in this example was higher than in the previous case since the sizes of the images were much smaller and the final results were slightly worse than that in the previous case.

4.5. Mismatched Feature Points

Mismatching of feature points is another source of measurement errors in the SfM problems. It happens when two or several features with similar appearances are located in image frames close to each other during the feature tracking process. This makes the feature tracker follow a wrong feature point instead of the correct one. The errors created by mismatched feature points cannot be statistically modeled using Gaussian random variables (Zhang, 1996). When mismatched feature points present, they affect the quality of the motion estimate if they are not dealt properly.

Recall that in Section 2, when we derived the equation of the joint likelihood of the whole observation of all feature points, we suggested to use (16) if some features are mismatched. The median value of the likelihood of all feature points is found at first. Then "good" features whose likelihood is larger than this median value are found and the product of their likelihood is used as the joint likelihood of the whole observation. By using (16), the effect of mismatched points can be removed and motion estimates can still be obtained. Simultaneously, the mismatched feature points can be detected by looking for feature points with large residual errors which are the resultant epipolar distances in our approach. In our experiment, 50 feature points were tracked through a synthetic sequence and among them, 20 were mismatched. Figure 11 shows the posterior distributions of the motion parameters, where the likelihoods were evaluated using (16).

4.6. Pure Rotation

In this experiment, the camera at first only rotated about the vertical axis of the image plane and after 20 frames, it started to move toward the right. In Fig. 12,

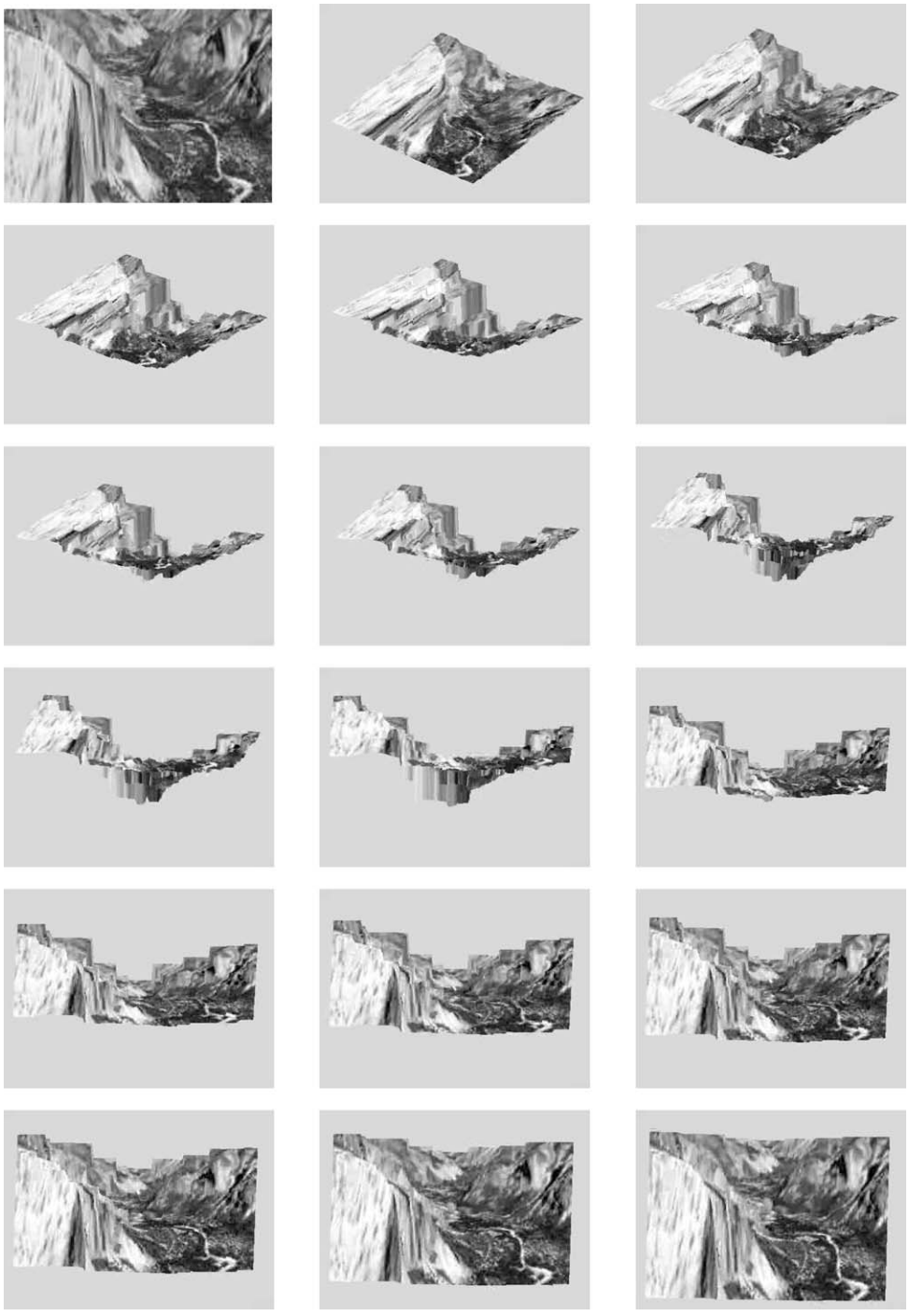

Figure 13. The texture map and reconstructed 3D model of the *Yosemite* sequence.

the posterior distributions of the motion parameters are shown. The posterior probability that the sequence is pure rotation is computed using (20) and is shown Fig. 12(f). We see that the probability is one in the beginning up to the 20th frame, which shows that the camera is only rotating. From around the 24th frame, it dramatically decreases to 0 because the camera started to translate.

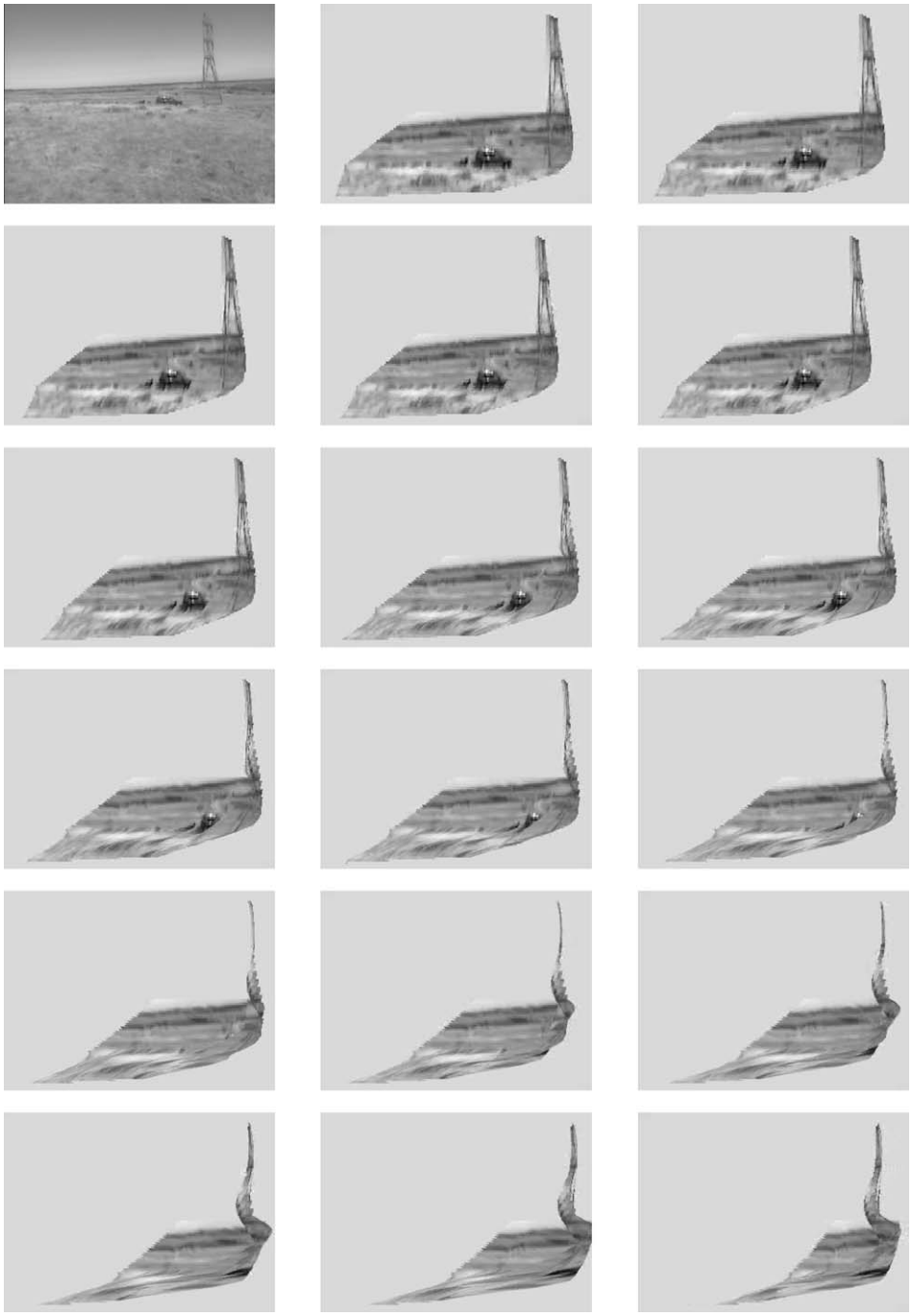

Figure 14. The texture map and reconstructed 3D model of the outdoor sequence.

4.7. Experiments Using Yosemite *Sequence*

We have applied the proposed approach to the well-known *Yosemite* sequence. Feature points were detected and tracked through the sequences by using the well known KLT feature tracker (Tomasi and Shi, 1994). The reconstructed 3D model is shown in Fig. 13. The depth variation in the

Figure 15. The intensity texture map and reconstructed 3D model of the face sequence.

valley can be clearly observed in the reconstructed model.

4.8. Experiments Using Real Images

Real image sequences have been used to test the proposed algorithm. The KLT tracker is used to provide feature correspondences.

Outdoor Sequence. In the first example, an outdoor sequence is used. In the scene, there is a moving vehicle and some of the feature points are located on the moving vehicle. The up-most-left figure in Fig. 14 shows one frame of the sequence and the remaining figures in Fig. 14 show the reconstructed 3D model. We can clearly see the 3D structure of the tower and the ground plane.

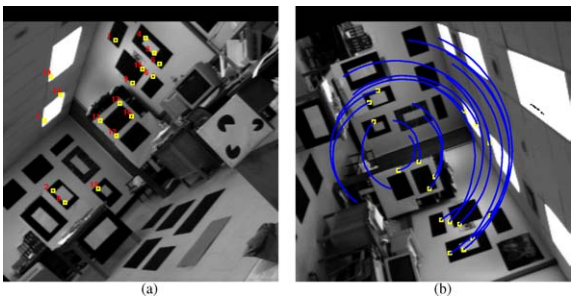

Figure 16. Feature locations and trajectories using the *PUMA2* sequence.

Face Sequence. In the second example, a 3D face model is reconstructed from a face sequence. Since the proposed method is feature-based, the depth values of a set of feature points are first computed using the proposed approach and then the depth field of the entire object is obtained by using the interpolation function *griddata* in MATLAB 6.1.0. Figure 15 show the intensity map of the face (the up-most-left one) and the reconstructed face model from different viewpoints.

UMASS PUMA2 Sequence. The 30-frame long *PUMA2* sequence was taken by mounting a camera at the end of a robot arm. The robot arm was rotated for 120 degrees (2.09 rad.) and 30 image frames were taken. The radius of the robot arm is approximately 1.8 feet. The *PUMA2* sequence is not a pure-rotation sequence, because when the camera moves along a circle, it translates slightly. However, since the radius of the robot arm is small compared to the depth of scene, in the initial part of the sequence, the motion of the sequence is very close to pure rotation about the optical axis. Figure 16 shows the location and trajectories of feature points used to compute the motion of the camera. The motion distributions and $P(\textit{pure rotation})$ of this sequence at different time instants are given in Fig. 17. Plot (c) shows the ground-truth (the thick line) and computed distribution of the rotation about the optical axis. Plot (f) shows the posterior probability of pure rotation of the sequence at different time instant. In the beginning, this probability is one and it means that the sequence was categorized as a pure-rotation sequence. As the translation size increased, this probability goes to zero which is true according to the generative model of the sequence. Although EKF based algorithm has been used to estimate the camera motion using *PUMA2* sequence and good motion estimates have been obtained (Wu et al., 1995), our algorithm is able to measure the probability of the pure rotation of the camera motion, without looking at the translation magnitude. It makes the measurement of this probability more direct and accurate and certainly is an advantage over other algorithms.

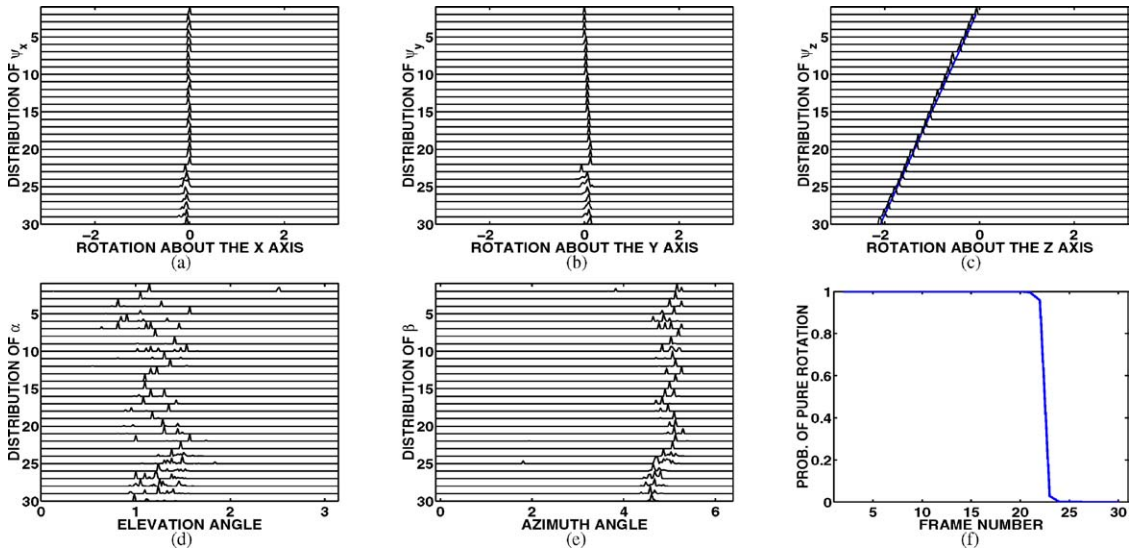

Figure 17. Posterior motion distributions using the *PUMA2* sequence. (f) shows $P(\textit{pure rotation})$ of this sequence.

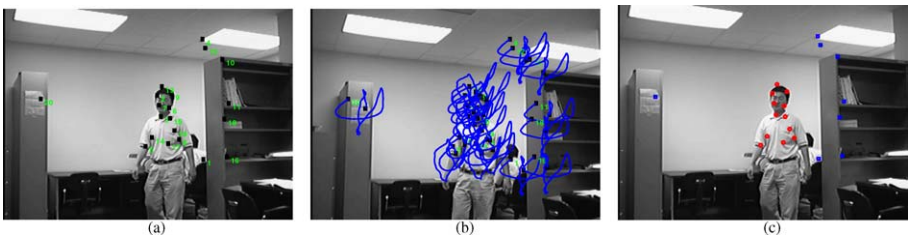

Figure 18. Feature locations and trajectories using the sequence with a walking person. (c) shows the feature points segmentation results. Points belonging to the walking person are marked by circles.

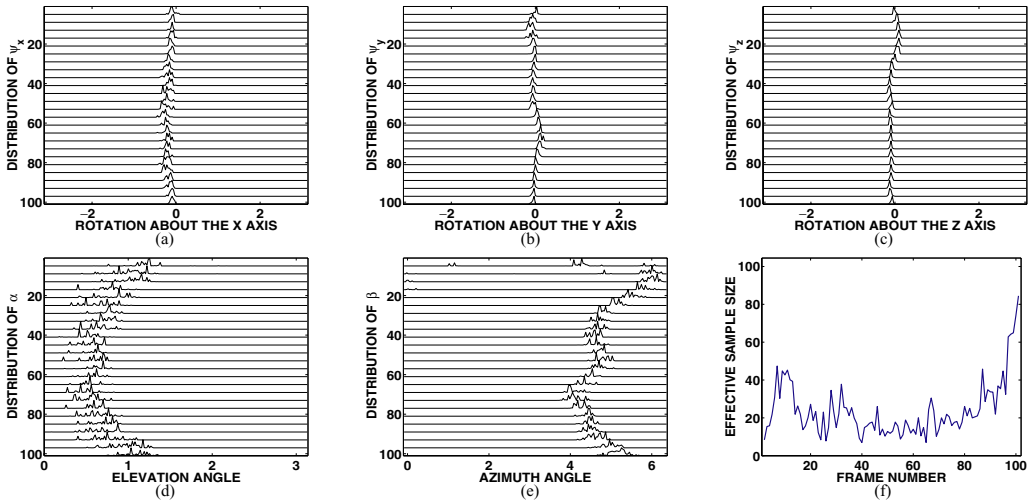

Figure 19. Posterior distributions of the motion parameters using the sequence with a walking person. (f) shows the ESS in this case.

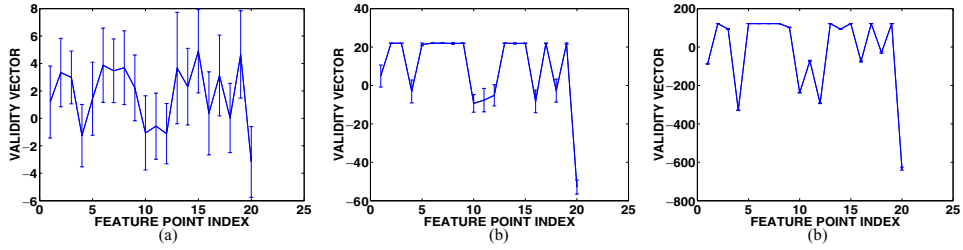

Figure 20. Empirical means and variances of the sample sets of the validity vector from using the sequence with a walking person. (a)–(c) shows the results at the initial, medium and later time instants. We can see that the validity vector samples gradually converge to the desired structure with low variations.

Dynamic Scenes. The SIS procedure for detecting moving objects in dynamic scenes has been tested using both synthetic and real images. One experimental results using real video is shown below. In this experiment, a sequence with a walking person was captured using a moving camera. Twenty feature points were detected and tracked throughout the sequence. Figure 18 shows their locations and trajectories in the image plane. Twelve feature points (No. 2, 3, 5, 6, 7, 8, 9, 13, 14, 15, 17, 19) are on the walking person and the remaining on the background. By using the modified SIS algorithm for dynamic scenes, the points on the walking person can be segmented out from the feature set. Since points on the walking person are relatively static and they outnumbered the background features, the camera motion computed here is relative

to the walking person. Figure 19 shows the posterior motion distributions. Figure 20(a)–(c) are the empirical means and variances using the samples and weights of the validity vectors at initial, medium and late time instants. We see that the structure of the validity vector evolved to the desired pattern: entries corresponding to the points on the relative background have high positive values and those related to the relative moving objects are negative. As more frames are processed, the difference between the entries of these two categories in the validity vectors becomes large and the variances of the validity vector samples decrease, which indicates the convergence of the computation of validity vector. The segmentation result is shown in Fig. 18(c). Feature points belonging to the walking person are marked by circles.

5. Conclusion

A sequential importance sampling based SfM algorithm has been presented. The posterior distribution of the camera motion and scene structure are approximated by sample and weight sets. Experimental results using both synthetic and real images show that the proposed method is capable of dealing with the difficulties met in the SfM problem due to errors in feature tracking, feature occlusion, motion/structure ambiguity, mixed-domain sequences, mismatched feature points, and independently moving objects. Although promising results have been obtained, the proposed SfM framework can still be improved in a number of ways. For example, one of the limitations is that current data flow in motion and structure estimation is only unidirectional, i.e., the translation magnitude and depth estimates are not used to improve the partial motion estimates. In future work, we will integrate the translation magnitude and depth estimates from previous time into the partial motion estimation stage at the current time so that better partial motion estimates will be obtained.

Appendix

A. A proposition

Proposition 2. *The distribution* $p(\mathbf{z}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{y}_t)$ *is proportional to* $p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}_t)\pi(\mathbf{z}_t)$.

Proof: The conditional distribution of $\mathbf{z}_t$ given the observation and the motion parameter $p(\mathbf{z}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{y}_t)$ can be written using the Bayes' rule as

$$p(\mathbf{z}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{y}_t) = \frac{p(\mathbf{z}_t, \gamma_t, \mathbf{x}_t, \mathbf{y}_t)}{p(\gamma_t, \mathbf{x}_t, \mathbf{y}_t)} = \frac{p(\mathbf{y}_t \mid \mathbf{z}_t, \gamma_t, \mathbf{x}_t)p(\gamma_t, \mathbf{x}_t, \mathbf{z}_t)}{p(\gamma_t, \mathbf{x}_t, \mathbf{y}_t)}$$

Since the motion parameter and the depth values of the feature points are independent, we have

$$p(\gamma_t, \mathbf{x}_t, \mathbf{z}_t) = p(\gamma_t, \mathbf{x}_t)\pi(\mathbf{z})$$

Because $\gamma_t, \mathbf{x}_t$ and $\mathbf{y}_t$ are all known and $p(\gamma_t, \mathbf{x}_t)$ and $p(\gamma_t, \mathbf{x}_t, \mathbf{y}_t)$ are constants, we have

$$p(\mathbf{z}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{y}_t) \propto p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}_t)\pi(\mathbf{z}_t)$$

Hence $p(\mathbf{z}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{y}_t)$ is proportional to $p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}_t)\pi(\mathbf{z}_t)$. □

B. Proof of Eq. (33)

$$p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) \approx p(\Gamma_{t-1} \mid \mathcal{X}_{t-1}, \mathcal{Y}_{t-1})p(\gamma_t \mid \gamma_{t-1}) \times \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)}{p(\mathbf{y}_t \mid \mathbf{x}_t)}$$

Proof:

$$\begin{aligned} & p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) \\ &= \frac{p(\Gamma_t, \mathcal{X}_t, \mathcal{Y}_t)}{p(\mathcal{X}_t, \mathcal{Y}_t)} \\ &\approx \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)p(\gamma_t \mid \gamma_{t-1})p(\mathbf{x}_t \mid \mathbf{x}_{t-1})p(\Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}))}{p(\mathbf{y}_t \mid \mathbf{x}_t)p(\mathbf{x}_t \mid \mathbf{x}_{t-1})p(\mathcal{X}_{t-1}, \mathcal{Y}_{t-1})} \\ &= p(\Gamma_{t-1} \mid \mathcal{X}_{t-1} \| X_{t-1} \| \mathcal{Y}_{t-1} \| \mathcal{Y}_{t-1})p(\gamma_t \mid \gamma_{t-1}) \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)}{p(\mathbf{y}_t \mid \mathbf{x}_t)} \end{aligned}$$

□

C. Proof of Eq. (46)

$$p(\mathbf{z} \mid \mathcal{Y}_t) \approx \frac{p(\mathcal{Y}_{t-1})}{p(\mathcal{Y}_t)} p(\mathbf{z} \mid \mathcal{Y}_{t-1}) \times \int_{\gamma_t, \mathbf{x}_t} p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z})p(\gamma_t \mid \gamma_{t-1}) \times p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) d\gamma_t d\mathbf{x}_t$$

Proof: Recall that

$$
\begin{aligned}
p(\Gamma_t \mid \mathcal{X}_t, \mathbf{Y}_t) &\approx p(\Gamma_{t-1} \mid \mathcal{X}_{t-1}, \mathbf{Y}_{t-1}) p(\gamma_t \mid \gamma_{t-1}) \\
&\quad \times \frac{p(y_t \mid \gamma_t, \mathbf{x}_t)}{p(y_t \mid \mathbf{x}_t)} \\
p(\mathbf{z} \mid \Gamma_t, \mathcal{X}_t, \mathcal{Y}_t) &\approx p(\mathbf{z} \mid \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) \\
&\quad \times \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z})}{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)} \\
p(\mathcal{X}_t \mid \mathcal{Y}_t) &\approx \frac{p(\mathcal{Y}_{t-1})}{p(\mathcal{Y}_t)} p(\mathbf{y}_t \mid \mathbf{x}_t) \\
&\quad \times p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathcal{X}_{t-1} \mid \mathcal{Y}_{t-1})
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
p(\mathbf{z} \mid \mathcal{Y}_t) &= \int_{\Gamma_t, \mathcal{X}_t} p(\mathbf{z}, \Gamma_t, \mathcal{X}_t \mid \mathcal{Y}_t) d\Gamma_t d\mathcal{X}_t \\
&= \int_{\Gamma_t, \mathcal{X}_t} p(\mathbf{z} \mid \Gamma_t, \mathcal{X}_t, \mathcal{Y}_t) p(\Gamma_t \mid \mathcal{X}_t, \mathcal{Y}_t) \\
&\quad \times p(\mathcal{X}_t \mid \mathcal{Y}_t) d\Gamma_t d\mathcal{X}_t \\
&\approx \int_{\Gamma_t, \mathcal{X}_t} p(\mathbf{z} \mid \Gamma_{t-1}, \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) \\
&\quad \times \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z})}{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)} \cdot \\
&\quad \times p(\Gamma_{t-1} \mid \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) p(\gamma_t \mid \gamma_{t-1}) \\
&\quad \times \frac{p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t)}{p(\mathbf{y}_t \mid \mathbf{x}_t)} \cdot \\
&\quad \times p(\mathbf{y}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathcal{X}_{t-1} \mid \mathcal{Y}_{t-1}) \\
&\quad \times \frac{p(\mathcal{Y}_{t-1})}{p(\mathcal{Y}_t)} d\Gamma_t d\mathcal{X}_t \\
&= \frac{p(\mathcal{Y}_{t-1})}{p(\mathcal{Y}_t)} \int_{\Gamma_{t-1}, \mathcal{X}_{t-1}} \\
&\quad \times p(\mathbf{z}, \Gamma_{t-1}, \mathcal{X}_{t-1} \mid \mathcal{Y}_{t-1}) d\Gamma_{t-1} d\mathcal{X}_{t-1} \cdot \\
&\quad \times \int_{\gamma_t, \mathbf{x}_t} p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}) p(\gamma_t \mid \gamma_{t-1}) \\
&\quad \times p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) d\gamma_t d\mathbf{x}_t \\
&= \frac{p(\mathcal{Y}_{t-1})}{p(\mathcal{Y}_t)} p(\mathbf{z} \mid \mathcal{Y}_{t-1}) \int_{\gamma_t, \mathbf{x}_t} p(\mathbf{y}_t \mid \gamma_t, \mathbf{x}_t, \mathbf{z}) \\
&\quad \times p(\gamma_t \mid \gamma_{t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) d\gamma_t d\mathbf{x}_t
\end{aligned}
$$

□

Acknowledgments

The work presented in this paper is partially supported by the Office of Naval Research under the grant N00014-011-0265. We are grateful to the anonymous reviewer for pointing out errors in an earlier version of the paper.

References

Adiv, G. 1989. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(5):477–489.

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

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

Chiuso, A., Favaro, P., Jin, H., and Soatto, S. 2002. Motion and structure causally integrated over time. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:523–535.

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

Dellaert, F., Seitz, S., Thorpe, C., and Thrun, S. 2000. Structure from motion without correspondence. In *IEEE Computer Vision and Pattern Recognition*, Hilton Head, SC.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

Doucet, A., Freitas, N., and Gordon, N. 2001. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag: New York.

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

Forsyth, D., Ioffe, S., and Haddon, J. 1999. Bayesian structure from motion. In *International Conference on Computer Vision*. Corfu, Greece, pp. 660–665.

Gordon, N., Salmon, D., and Smith, A. 1993. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings*, 140:107–113.

Hartley, H.O. 1958. Maximum likelihood from incomplete data. *Biometrics*, 14:174–194.

Hartley, R. and Zisserman, A. 2000. *Multiple View Geometry*. Cambridge, UK: Cambridge University Press.

Huang, T. and Netravali, A. 1994. Motion and structure from feature correspondences: A review. *Proceedings of the IEEE*, 82(2):252–268.

Isard, M. and Blake, A. 1996. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision*, Cambridge, UK, vol. I, pp. 343–356.

Jebara, T., Azarbayejani, A., and Pentland, A. 1999. 3-D structure from 2D motion. *IEEE Signal Processing Magazine*, 16:66–84.

Jerian, C. and Jain, R. 1991. Structure from motion: A critical analysis of methods. *IEEE Trans. Systems, Man and Cybernetics*, 21:572–588.

Kitagawa, G. 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25.

Kong, A., Liu, J.S., and Wong, W.H. 1994. Sequential imputation method and Bayesian missing data problems. *Journal of the American Statistical Association*, 89:278–288.

Liu, J.S. and Chen, R. 1998. Sequential monte carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, 93:1032–1044.

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

Oliensis, J. 2000. A critique of structure from motion algorithms. Technical Report www.neci.nj.com/∼homepages/oliensis/poleiccv.ps, NEC Research Institute, Princeton, NJ.

Qian, G., Chellappa, R., and Zheng, Q. 2001. Robust structure from motion estimation using inertial data. *Journal of the Optical Society of America A*, 18:2982–2997.

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

Soatto, S. and Perona, P. 1998. Reducing structure-from-motion: A general framework for dynamic vision Part 1: Modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(9):933–942.

Tian, T., Tomasi, C., and Heeger, D. 1996. Comparison of approaches to egomotion computation. In *IEEE Computer Vision and Pattern Recognition*, San Francisco, CA, pp. 315–320.

Tomasi, C. and Kanade, T. 1992. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154.

Tomasi, C. and Shi, J. 1994. Good features to track. In *IEEE Computer Vision and Pattern Recognition*, Seattle, WA, pp. 593–600.

Wu, T., Chellappa, R., and Zheng, Q. 1995. Experiments on estimating egomotion and structure parameters using long monocular image sequences. *International Journal of Computer Vision*, 15:77–103.

Young, G. and Chellappa, R. 1992, Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(10):995–1013.

Zhang, Z. 1996. Determining the epipolar geometry and its uncertainty: A review. Technical report, French National Institute for Research in Computer Science and Control (INRIA) No. 2927.