Fusing Depth and Video using Rao-Blackwellized Particle Filter

Amit Agrawal and Rama Chellappa

University of Maryland
College Park, MD 20742 USA
aagrawal@cfar.umd.edu

Abstract. We address the problem of fusing sparse and noisy depth data obtained from a range finder with features obtained from intensity images to estimate ego-motion and refine 3D structure of a scene using a Rao-Blackwellized particle filter. For scenes with low depth variability, the algorithm shows an alternate way of performing Structure from Motion (SfM) starting with a flat depth map. Instead of using 3D depths, we formulate the problem using 2D image domain parallax and show that conditioned on non-linear motion parameters, the parallax magnitude with respect to the projection of the vanishing point forms a linear subsystem independent of camera motion and their distributions can be analytically integrated. Thus, the structure is obtained by estimating parallax with respect to the given depths using a Kalman filter and only the ego-motion is estimated using a particle filter. Hence, the required number of particles becomes independent of the number of feature points which is an improvement over previous algorithms. Experimental results on both synthetic and real data show the effectiveness of our approach.

1 Introduction

SfM refers to the estimation of 3D scene structure and sensor motion given monocular or stereo images. With increasing availability of range sensors in 3D modeling, the need for fusing the depth information from these sensors with information from intensity images naturally arises. However, the available depth data from range sensors is often noisy and coarse. We focus on using such coarse and noisy depth information along with sparse noisy features detected from intensity images to refine the depths and estimate the ego-motion under perspective projection. A moving camera can be viewed as a dynamical system where a state space model can be used to describe the motion and/or depths with the noisy 2D feature points denoting the observations. This approach was used in [1] where a Kalman filter was used to fuse information from a sequence of images. In [2], the problem was reformulated to reduce the size of the state space and an extended Kalman filter was used to handle the non-linearities. The use of prior rate data for SfM was demonstrated in [3]. Recently, Sequential Monte Carlo (SMC) methods have emerged as a powerful tool for estimation, prediction and filtering of non-linear dynamical systems. Kitagawa [4] proposed a Monte Carlo filter for non-linear/non-gaussian dynamical systems. The bootstrap filter proposed by Gordon [5] is a similar variant of the SMC method. Several SMC

methods have been proposed to handle various problems in computer vision such as shape and contour tracking [6], SfM [7], tracking [8], and self-calibration [9]. Most of the existing methods for SfM can be divided into two broad categories [10]

- Methods which *reduce* SfM such as [7]. These methods eliminate depths from the state space (e.g. utilizing the epipolar constraints) and estimate the camera motion. The structure can be estimated after motion estimation in a variety of ways such as using another Sequential Importance Sampling (SIS) procedure as in [7].
- Methods which attempt to estimate both structure and motion simultaneously. Here the state space consists of both structure and motion parameters and the size of state space increases linearly with the number of feature points.

With respect to the problem at hand, both approaches have limitations. The first approach eliminates depths in motion estimation. Thus, even if we have some prior depth information, such information can not be used to improve the motion estimates. The second approach has a considerable disadvantage that the size of state space increases linearly with the number of feature points as structure is also a part of the state space. Thus the number of particles in the particle filtering scheme needs to increase [11][12] which makes it computationally inefficient and unstable. To overcome the above limitations, we propose a very simple analytical formulation which has the following advantages

- Structure and motion are estimated *simultaneously* with a method for incorporating prior depth information. Thus structure is also a part of the state space. Thus our approach can deal with general 3D scenes and is *not* restricted.
- Although structure is a part of state space, the number of particles required is independent of the number of feature points.

The approach proposed in this paper is based on the Rao-Blackwellisation [13] and marginalized particle filter schemes [11]. If one can find a linear subsystem in the state space model conditioned on the rest of the states, the distributions corresponding to linear states can analytically integrated. We show that by working in a 2D domain using parallax, the parallax magnitude with respect to the projection of the vanishing point form a linear subsystem conditioned on the non-linear motion parameters. Thus, in our formulation, the non-linear part of the state space for which a particle filter is used consists of only the motion parameters and the distributions of the linear part (consisting of parallax magnitudes) is estimated using a Kalman filter. Prior information on depths can be transferred as prior information on parallax magnitudes and hence an efficient way of incorporating prior depth information can be obtained. In addition, for scenes with low depth variability, the approach can be viewed as an alternate way of performing SfM starting from a flat depth map.

2 Algorithm

Let the Z axis of the camera point in the direction of the principal axis. At time instant 0, the camera coordinate system is aligned with the world coordinate sys-

tem. We parameterize the motion at time t as $\mathbf{m}^t = (\omega_x, \omega_y, \omega_z, \alpha, \gamma, s)$ where $\Psi^t = (\omega_x, \omega_y, \omega_z)$ are the *total* rotational angles along the X, Y and Z axis upto current time t, (α, γ) denotes the elevation and azimuth angles and s denotes the scale. The translation direction is then $[\sin(\alpha)\cos(\gamma), \sin(\alpha)\sin(\gamma), \cos(\alpha)]^T$ and the *total* translation upto current time is $T(\alpha, \gamma, s) = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} = s\times \begin{bmatrix} \sin(\alpha)\cos(\gamma) \\ \sin(\alpha)\sin(\gamma) \\ \cos(\alpha) \end{bmatrix}$. Thus the overall camera motion in the world coordinate system is estimated. The rotation matrix R^t is then given by (1)

$$R^t = \begin{bmatrix} \eta_1^2 + (1-\eta_1^2)\tau & \eta_1\eta_2(1-\tau) + \eta_3\zeta & \eta_1\eta_3(1-\tau) - \eta_2\zeta \\ \eta_1\eta_2(1-\tau) - \eta_3\zeta & \eta_2^2 + (1-\eta_2^2)\tau & \eta_2\eta_3(1-\tau) + \eta_1\zeta \\ \eta_1\eta_3(1-\tau) + \eta_2\zeta & \eta_2\eta_3(1-\tau) - \eta_1\zeta & \eta_3^2 + (1-\eta_3^2)\tau \end{bmatrix} \tag{1}$$

where $\eta = (\eta_1, \eta_2, \eta_3)^T = \Psi^t/|\Psi^t|$ is the direction cosine vector, $\zeta = \sin|\Psi^t|$ and $\tau = \cos|\Psi^t|$. Let $P = [X, Y, Z]^T$ denote a 3D point on the rigid scene in the world coordinate system. The projection of P on to the image plane at time 0 is given by

$$\mathbf{p} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} X/Z \\ Y/Z \end{bmatrix} \tag{2}$$

where we assume that the focal length of the camera equals one (or the image pixels have been normalized w.r.t. focal length). Thus given the projection $\mathbf{p}$ at time 0, we can parameterize the 3D coordinates[1] as $X = uZ$, $Y = vZ$. At each time instant t, the 3D point P^t is given by the following motion model

$$P^t = R^t P + T^t \tag{3}$$

Let $R^t = \begin{bmatrix} r_{11}\ r_{12}\ r_{13} \\ r_{21}\ r_{22}\ r_{23} \\ r_{31}\ r_{32}\ r_{33} \end{bmatrix}$. Using (2) and (3), the projection of P^t, $\mathbf{p}^t = \begin{bmatrix} u^t \\ v^t \end{bmatrix}$ is

$$u^t = \frac{Za + T_x}{Zc + T_z}, v^t = \frac{Zb + T_y}{Zc + T_z} \tag{4}$$

where for simplicity $a = r_{11}u + r_{12}v + r_{13}$, $b = r_{21}u + r_{22}v + r_{23}$ and $c = r_{31}u + r_{32}v + r_{33}$. The prior depth information (also referred to as *reference depths*) gives us some estimate $\widehat{Z}$ of the 3D point P which essentially correspond to a different point Q along the 3D ray. Let $\mathbf{q}^t = h(\mathbf{m}^t, \widehat{Z})$ denotes the projection of Q^t at time t. Thus, from (4) $\mathbf{q}^t = \begin{bmatrix} u_q^t \\ v_q^t \end{bmatrix} = \begin{bmatrix} (\widehat{Z}a + T_x)/(\widehat{Z}c + T_z) \\ (\widehat{Z}b + T_y)/(\widehat{Z}c + T_z) \end{bmatrix}$. It is well known that $\mathbf{q}^t - \mathbf{p}^t$ lies along the epipolar direction $\mathbf{q}^t - \mathbf{e}^t$ where $\mathbf{e}^t$ denotes the epipole [14] [15]. It is also the parallax due to $\widehat{Z}$. The parallax can be parameterized as a scalar (parallax magnitude β^t) times the vector along the epipolar direction, i.e. $(\mathbf{q}^t - \mathbf{p}^t) = \beta^t(\mathbf{q}^t - \mathbf{e}^t)$ The exact form of β^t is then given by

$$\beta^t = \frac{\widehat{Z} - Z}{\widehat{Z}} \frac{T_z}{Zc + T_z} \tag{5}$$

[1] We assume the bias in feature points for the first image to be zero for simplicity.

Thus we have $\mathbf{p}^t = \mathbf{q}^t - (\mathbf{q}^t - \mathbf{p}^t) = \mathbf{q}^t - \beta^t(\mathbf{q}^t - \mathbf{e}^t)$

The above equation gives a *linear* relationship between the observed projection $\mathbf{p}^t$ and the parallax magnitude β^t. However, from (5), we observe that β^t depends on T_z or current motion $\mathbf{m}^t$. Thus expressing β^t in terms of β^{t-1} and $\mathbf{m}^{t-1}$ becomes cumbersome due to the dependence of β^t on $\mathbf{m}^t$. We can get a much simpler formulation if we use a different parametrization based on the projection of the vanishing point of the 3D ray corresponding to point P. Although the vanishing point is generally used in context of parallel set of lines, here by vanishing point (**vp**) we mean the intersection of the 3D ray corresponding to feature point $\mathbf{p}$ (in first frame) with the plane at infinity. Let $\mathbf{pvp}^t$ denote the projection of the vanishing point at time t. As $Z \to \infty$, using (4), we have $\mathbf{pvp}^t = \begin{bmatrix} a/c \\ b/c \end{bmatrix}$. Thus, if we write $\mathbf{p}^t = \mathbf{q}^t - (\mathbf{q}^t - \mathbf{p}^t) = \mathbf{q}^t - \beta^t(\mathbf{q}^t - \mathbf{pvp}^t)$ one can solve for β^t as $\beta^t = -\frac{\widehat{Z}-Z}{Z}\frac{c}{c+T_z/Z}$ Thus when $T_z \ll Z$, i.e. motion in Z direction is small compared to the depths, we have

$$\beta^t \approx -(\widehat{Z} - Z)/Z \tag{6}$$

which is constant and **independent** of camera motion across all the frames. Thus by formulating the parallax with respect to the projection of the vanishing point, the parallax magnitude becomes independent of camera motion and is linear with the observations $\mathbf{p}$, given (or conditioned on) the motion and reference depths.

2.1 State Space Model

Suppose we track N feature points $i = 1 \dots N$ across K time instants $t = 1 \dots K$. To capture the motion dynamics, we use a 1-step predictive model for motion. Let $\dot{\mathbf{m}}^t = (\dot{\omega_x}, \dot{\omega_y}, \dot{\omega_z}, \dot{\alpha}, \dot{\gamma}, \dot{s})$. The state vector at time t is $\mathbf{x}^t = (\mathbf{x}^t_{nl}, \mathbf{x}^t_l) = (\mathbf{m}^t, \dot{\mathbf{m}}^t, \beta^t_1, \beta^t_2, \dots, \beta^t_N)$ consisting of two parts: the *non-linear* states $\mathbf{x}^t_{nl} = (\mathbf{m}^t, \dot{\mathbf{m}}^t)$ and the *linear* states $\mathbf{x}^t_l = (\beta^t_1, \beta^t_2, \dots, \beta^t_N)$ consisting of the parallax magnitudes for all the feature points. The state equations can then be written as

$$\begin{aligned} \mathbf{m}^{t+1} &= \mathbf{m}^t + \dot{\mathbf{m}}^t + \mathbf{n}^m \qquad \dot{\mathbf{m}}^{t+1} = \dot{\mathbf{m}}^t + \mathbf{n}^{\dot{m}} \\ \beta_i^{t+1} &= \beta_i^t + w_i \quad \text{for i=1...N} \end{aligned} \tag{7}$$

where the state noise, $\mathbf{n}^m$ and $\mathbf{n}^{\dot{m}}$ is assumed to be Gaussian for the rotational velocities and uniform for the translational directional angles and scale. We also assume a IID gaussian state noise $w_i \sim \mathsf{N}(0, Q_i^l)$ with very low variance ($Q_i^l \approx 10^{-3}$) in β_i. The observation equation for $i = 1 \dots N$ can be written as

$$\mathbf{p}_i^t = \mathbf{q}_i^t - (\mathbf{q}_i^t - \mathbf{p}_i^t) + \mathbf{n}_i^p = h(\mathbf{m}^t, \widehat{Z_i}) - \beta_i^t C(\mathbf{m}^t, \widehat{Z_i}) + \mathbf{n}_i^p \tag{8}$$

where we assume the observation noise for each feature point to be distributed as $\mathbf{n}_i^p \sim \mathsf{N}(0, \sigma_p^2)$. For each feature point i, $\mathbf{q}_i^t$ is a non-linear function h of current motion $\mathbf{m}^t$ and reference depths $\widehat{Z_i}$. Similarly, $(\mathbf{q}_i^t - \mathbf{pvp}_i^t)$ is a non-linear function

C of current motion and reference depths. Thus our state space model is of the form of diagonal model as in [11] and the marginalized particle filter described in [11] can be used to compute the posterior distributions of motion and parallax magnitudes. The filtering procedure then follows *Algorithm* 1 in [11] and we refer the reader to [11] for further details. In next, we show how to get the prior for parallax values.

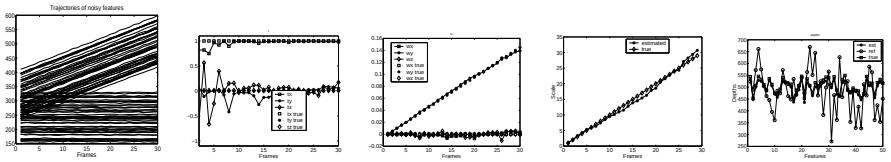
Use of prior depth information In general, the range sensor will provide some estimate of depth values along with their uncertainties (mean and covariances). For each feature point i, let $\widehat{Z_i}$ be distributed as $\widehat{Z}_i \sim \mathsf{N}(Z_i + m_i, \sigma_i^2)$, where Z_i is the true depth value. Thus the reference depth is assumed to follow a Gaussian density around the true depth with mean m_i and variance σ_i^2. Using (6), the prior distribution on the parallax magnitude β_i will be $\beta_i \sim \mathsf{N}(-m_i/Z_i, \sigma_i^2/Z_i^2)$. In practical scenarios, since Z_i is not known, we can use the given reference depth value. Thus we can assume $\beta_i \sim \mathsf{N}(-m_i/\widehat{Z}_i, \sigma_i^2/\widehat{Z}_i^2)$. For scenes where depth variability is low, one can use the above formulation starting with a flat depth map as the reference depths As before, the initial variance of the parallax magnitude can be set to few pixels.

3 Experiments

In all experiments, we use a particle filter with 2500 particles and estimates refer to Maximum A-Posteriori (MAP) estimates. Kanade-Lucas-Tomasi (KLT) feature tracker was used to track features.

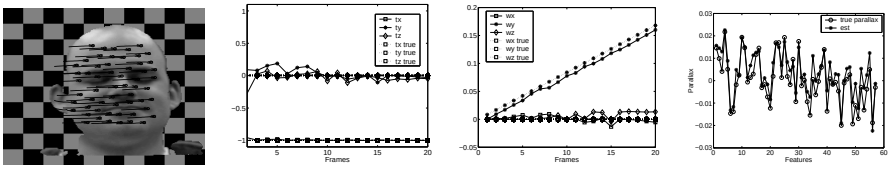
Synthetic Data A random cloud of 50 feature points was generated and their 2D projections were taken in 30 frames. Gaussian random noise of variance 1 pixel was added to feature trajectories. The camera was translated along the X axis with rotation about the Y axis. Fig. 1(a) shows the trajectories of the X and Y coordinates of all the noisy feature points with frames. For each feature point, the reference depth was chosen to be randomly distributed around the true depth with $\sigma = 0.2$ times the depth value. Thus the initial mean and variance of parallax magnitude for all feature points was set as $\mu^i_{0|-1} = 0, P^i_{0|-1} = \sigma^2$. Figs. 1(b), 1(c) and 1(d) shows the estimates of the translation direction, rotational velocities and scale with frames respectively along with ground truth. Fig. 1(e) shows the plot of initial reference depths, true depths and estimated depths for all feature points in the last frame using the estimated parallax. The estimated depths are close to ground truth.

Face Sequence The face texture images and range data were downloaded from http://sampl.eng.ohio-state.edu/ sampl/data/3DDB/RID/minolta/faces-hands.1299. 20 frames of the face sequence were generated from a virtual camera with the ground truth focal length. Fig. 2(a) shows the tracked features overlayed on the first frame. The initial (reference) depths were chosen to be equal for all feature points as shown in Fig. 2(e). Figs. 2(b) and 2(c) show the estimates of translation direction and rotational velocities. Figs. 2(d) and 2(e) show the comparison of estimated β and depths with ground truth for all feature points for the last frame. The estimates are very close to the true values. Figs. 2(f), 2(g) and 2(h) shows novel views of texture mapped 3D model.

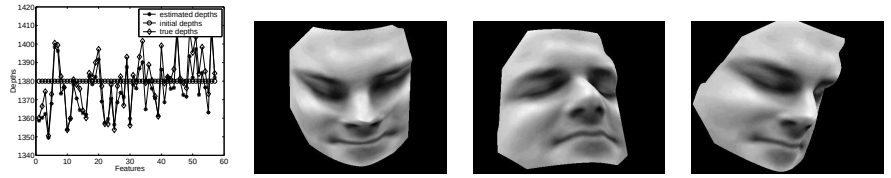


(a) Noisy feature trajectories (b) Translation direction estimates (c) Rotational velocities estimates (d) Scale estimates (e) True, reference and estimated depths

Fig. 1. Synthetic Example



(a) Tracked features (b) Translation direction estimates (c) Rotational velocities estimates (d) True and estimated β



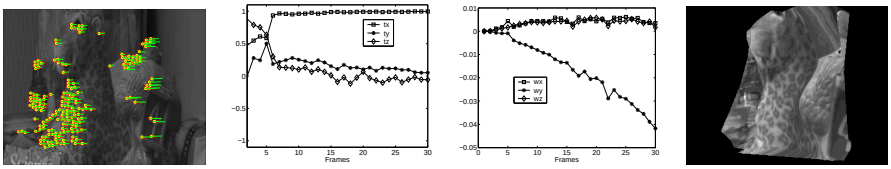
(e) True, reference and estimated depths (f) Novel View (g) Novel View (h) Novel View

Fig. 2. Face Sequence

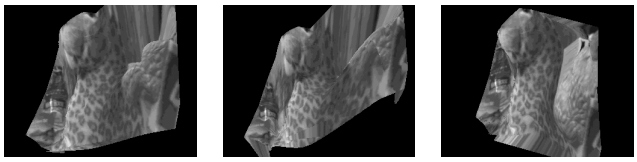
Indoor Sequence A video sequence of several toy objects was taken in a lab. The dominant camera motion was in the X direction. We choose the reference depths to be equal (500 units) for all feature points. The initial mean and variance for the parallax magnitudes were set to 0 and 1 pixel respectively. A total of 200 feature points were tracked for 30 frames with trajectories shown overlayed on first frame in Fig. 3(a). The estimates of translation direction and rotational velocities are shown in Fig. 3(b) and Fig. 3(c) respectively. The final depth map is obtained by interpolating the estimated depths. Fig. 3(d) to 3(g) shows novel views of texture mapped 3D model.

References

1. Broida, T., Chellappa, R.: Estimating the kinematics and structure of a rigid object from a sequence of monocular images. IEEE Trans. Pattern Anal. Machine Intell. **13** (1991) 497–513
2. Azarbayejani, A., Pentland, A.: Recursive estimation of motion, structure, and focal length. IEEE Trans. Pattern Anal. Machine Intell. **17** (1995) 562–575



(a) Tracked features (b) Translation direction estimates (c) Rotational velocities estimates (d) Novel View



(e) Novel View (f) Novel View (g) Novel View

Fig. 3. Indoor Sequence

3. Qian, G., Chellappa, R., Zheng, Q.: Robust structure from motion estimation using inertial data. J. Optical Society of America A (2001) 2982–2997
4. Kitagawa, G.: Monte carlo filter and smoother for non-gaussian nonlinear state space models. J. Computational and Graphical Statistics **5**(1) (1996) 1–25
5. Gordon, N., Salmond, D., Smith, A.: Novel approach to nonlinear/non-gaussian bayesian state estimation. In: IEE Proc. Radar, Sonar and Navig. (1993) 107–113
6. Isard, M., Blake, A.: Condensation – conditional density propagation for visual tracking. Int'l J. Computer Vision **29**(1) (1998) 5–28
7. Qian, G., Chellappa, R.: Structure from motion using sequential monte carlo methods. Int'l J. Computer Vision (2004) 5–31
8. Khan, Z., Balch, T., Dellaert, F.: A rao-blackwellized particle filter for eigentracking. In: Proc. Conf. Computer Vision and Pattern Recognition. (2004)
9. Qian, G., Chellappa, R.: Bayesian self-calibration of a moving camera. (to appear in *CVIU*)
10. Soatto, S., Perona, P.: Reducing structure from motion: A general framework for dynamic vision part 1: Modeling. IEEE Trans. Pattern Anal. Machine Intell. **20**(9) (1998) 933–942
11. Schon, T., Gustafsson, F.: Marginalized particle filters for mixed linear/nonlinear state-space models. (to appear in *IEEE Trans. Signal Processing*)
12. Khan, Z., Balch, T., Dellaert, F.: Efficient particle filter based tracking of multiple interacting targets using an mrf-based motion model. In: IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems. (2003)
13. Casella, G., Robert, C.: Rao-blackwellisation of sampling schemes. Biometrika **83**(1) (1994) 81–94
14. Anandan, P., Kumar, R., Hanna, K.: Shape recovery from multiple views: A parallax based approach. In: Proc. Image Understanding Workshop. (1994)
15. Sawhney, H.: 3D geometry from planar parallax. In: Proc. Conf. Computer Vision and Pattern Recognition. (1994) 929–934