ROBUST ESTIMATION OF DEPTH AND MOTION USING STOCHASTIC APPROXIMATION

Amit K. Roy Chowdhury, R. Chellappa

Center for Automation Research
Department of Electrical and Computer Engineering
University of Maryland, College Park
MD 20742, USA
{amitrc,chella}@cfar.umd.edu

ABSTRACT

The problem of structure from motion (SFM) is to extract the three-dimensional model of a moving scene from a sequence of images. Though two images are sufficient to produce a 3D reconstruction, they usually perform poorly because of errors in the estimation of the camera motion, thus motivating the need for multiple frame algorithms. One common approach to this problem is to determine the estimate from pairs of images and then fuse them together. Data fusion techniques, like the Kalman filter, require estimates of the error in modeling and observations. The complexity of the SFM problem makes it difficult to reliably estimate these errors and makes the multi-frame algorithm dependent on the two-frame one. This paper describes a new recursive algorithm to estimate the camera motion and scene structure by fusing the two-frame estimates, using stochastic approximation techniques. The method does not require estimates of the error in the two-frame case, is independent of the underlying two-frame algorithm and can reconstruct the scene to arbitrary accuracy given a sufficient number of frames. Experimental results are reported to justify the claims.

1. INTRODUCTION

The problem of structure from motion (SFM) is to extract the three-dimensional model of a moving scene from a sequence of images. Traditional SFM algorithms [1], [2] recover a 3D scene structure from two images. However, these algorithms often produce inaccurate reconstructions of the scene, mainly due to incorrect estimation of camera motion, thus necessitating multi-frame algorithms (MFSFM).

One obvious strategy in MFSFM algorithms is the *integration over time* approach [3]. However, this method can be potentially unstable if the initial estimate of the structure is inaccurate. An alternative is to obtain a structure estimate from the most recent pair of images, using a two-frame algorithm, which is then fused with the previous estimate [4]. It is desirable that such an algorithm be independent of the underlying two-frame algorithm. Moreover, fusion techniques require a reliable estimate of the error, which is difficult to obtain for many two-frame algorithms and even when possible, will be dependent on that particular method.

Our approach tries to address all these issues. We describe a recursive algorithm to estimate the 3D structure and camera motion from a sequence of images, given the estimates from every consecutive pair of images. The technique uses ideas from stochastic optimization (stochastic gradient and stochastic Newton search) to obtain a robust estimate independent of the underlying two-frame algorithm. With this method, it is possible to reconstruct the scene to an arbitrary accuracy given a sufficiently large number of frames. We also estimate the number of frames required for the estimation by recursively computing the Fisher information.

2. PROBLEM FORMULATION

It is assumed that the camera is moving in an unknown, fixed environment, consisting of isolated 3D points. The goal is to determine the locations of the 3D points and the motion of the camera in some coordinate system. Before we venture to describe the algorithm, a few important points are worth noting.

Observation Statistics Assumptions of normally distributed, independent observations are abundantly used in many estimation problems because of the central limit theorem and mathematical tractability. However, in many natural situations these assumptions are not valid and can give highly erroneous results. In Fig. 1, we plot the estimates of the first six moments and the first four cumulants of the two-frame depths values. For Gaus-

Prepared through collaborative participation in the Advanced Sensors Consortium (ASC) sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0001.

sian random variables, all odd central moments are identically zero and all cumulants greater than two are zero, which is not the case as seen from the figure. Regarding independence, since we use the same algorithm for every pair of images, there is every reason to believe that the errors will actually be dependent.

Outliers In order to make our algorithm robust to outliers, we use the least median of squares (LMedS) cost function rather than the least mean square (LMS). The LMedS method has a high breakdown point which makes it robust to outliers [5].

Tracking Camera Motion The camera motion parameters will change for every pair of frames and need to be tracked. Thus we will have to incorporate time varying dynamics into the stochastic approximation framework.

Two-Frame Algorithm As we do not require explicit distribution statistics for the error in the estimates, the method is independent of the underlying two-frame algorithm.

We now describe our problem formally. The modeling of the depth is done for each 3D point separately. Let s_i represent the depth computed from the i and $(i+1)$-th frame, $i = 1, ..., K$, $(K+1)$ being the total number of frames. As the camera moves to a new position, the fused structure S_i is transformed to the new coordinate system as $T_i(S_i) = R_i S_i + V_i$; and the problem at stage $(i+1)$ is to fuse s_{i+1} and $T_i(S_i)$, where R_i and V_i represent the rotation and translation of the camera between the i and $(i+1)$-th frames. (Note that the motion is not for every point, but for the entire object in the frame). Denoting the rotational component of the camera motion as $\Omega = [\omega_{\mathbf{x}}, \omega_{\mathbf{y}}, \omega_{\mathbf{z}}]'$ and the translational component $V = [v_x, v_y, v_z]'$, $R = \hat{\Omega} + I$. [1] Since we can estimate only the direction of the translational motion (due to the scale ambiguity), we represent the motion components by the vector $\mathbf{m} = [\omega_{\mathbf{x}}, \omega_{\mathbf{y}}, \omega_{\mathbf{z}}, \frac{\mathbf{v_x}}{\mathbf{v_z}}, \frac{\mathbf{v_y}}{\mathbf{v_z}}]$. We thus model the motion components as

$$\begin{aligned} m_{i+1} &= m_i + w_i \\ y_i &= m_i + e_i \end{aligned} \tag{2}$$

where m represents each component of the vector $\mathbf{m}$, w and e are noise processes with unknown distribution and y is the observation from the two-frame algorithm. If $\{d_i\}$ is the transformed sequence of depth values with respect to a common

[1] For any vector $\mathbf{a} = [\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}]$, there exists a unique skew-symmetric matrix

$$\hat{a} = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \tag{1}$$

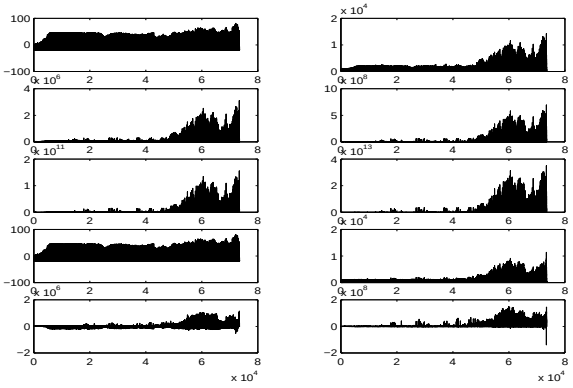

Fig. 1. The top six figures plot the estimates of the first six moments of the observation vector and the bottom four figures plot the first four cumulants. The horizontal axis represents the pixel number. The first column represents the odd central moments/cumulants and the second column the even ones.

frame of reference, then the optimal value of the depth at the point under consideration is obtained as

$$u^* = \arg\min_u \left(median(d_i - u)^2\right) \tag{3}$$

Note that we have addressed all the issues that were raised above.

3. THE ESTIMATION TECHNIQUE

3.1. Depth Estimation

The Robbins-Monro stochastic approximation (RMSA) algorithm is a stochastic search technique for finding the root θ^* to $g(\theta) = 0$ based on noisy measurements of $g(\theta)$, i.e. $Y_k(\theta) = g(\theta) + e_k(\theta), k = 1, ..., K$, where $e_k(\theta)$ is assumed to be the noise term, K is the number of observations and $E[Y(\theta, e)] = g(\theta)$ (E denotes expectation over e). The RMSA algorithm obtains the estimate by the following recursion, $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k Y_k(\hat{\theta}_k)$, where a_k is an appropriately chosen sequence [6]. Given a sequence of depth values $s_1, ..., s_K$ corresponding to a particular 3D point, we compute $X_i(u) = (s_i - u)^2$, where u (representing the true depth that we wish to determine) belongs to a predefined search set $\mathcal{U}$. Then we need to compute the median (say θ) of $X_0, ..., X_K$ with an unknown distribution F_X, i.e. obtain θ such that $g(\theta) = F_X(\theta) - 0.5 = 0$. Defining $Y_k(\hat{\theta}_k) = s_k(\hat{\theta}_k) - 0.5$ and $s_k(\hat{\theta}_k) = \mathbf{1}_{[X_k \leq \hat{T}_k(\hat{\theta}_k)]}$ ($\mathbf{1}$ represents the indicator func-

tion and $\hat{T}_k$ is the estimate of the camera motion),

$$\begin{aligned} E[Y_k(\hat{\theta}_k)|\hat{\theta}_k] &= E[s_k(\hat{\theta}_k)|\hat{\theta}_k] - 0.5 \\ &= E[\mathbf{1}(X_k \leq \hat{T}_k(\hat{\theta}_k))] - 0.5 \\ &= P(X_k \leq \hat{T}_k(\hat{\theta}_k)) - 0.5 \\ &= F_X(\hat{T}_k(\hat{\theta}_k)) - 0.5 = g(\hat{\theta}_k). \end{aligned}$$

Then the Robbins-Monro (RM) recursion for the problem is:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k(s_k(\hat{\theta}_k) - 0.5) \tag{4}$$

The choice of the gain sequence a_k is determined by the convergence properties of the algorithm [6]. [2] It is required that

$$a_k \geq 0, \quad \sum_{k=1}^{\infty} a_k = \infty, \quad \sum_{k=1}^{\infty} a_k^2 < \infty. \tag{5}$$

3.2. Camera Motion Tracking

We use the Stochastic Newton algorithm to track the time-varying camera motion. [6]. Denoting, as before, each component of $\mathbf{m}$ as m, the tracking equations become

$$\begin{aligned} \hat{m}_i &= \hat{m}_{i-1} + \gamma_i C_i^{-1}(y_i - \hat{m}_{i-1}) \\ C_i &= C_{i-1} + \gamma_i(1 - C_{i-1}). \end{aligned} \tag{6}$$

Since we want to track time-varying dynamics, we must relax the conditions on γ as stated in (5). We let γ_i tend to a small positive number γ_0, which is chosen as a tradeoff between tracking capability (γ_0 large) and noise sensitivity (γ_0 small).

3.3. Convergence Properties

It is well known that the RMSA estimate is strongly consistent and the error in the estimate converges in distribution to a normal with zero mean and suitable covariance matrix which depends on the Jacobian of $g(\theta)$ and a_k [6]. Thus given a suitably large number of frames, the estimate of the depth obtained by our recursion can be arbitrarily close to the true value.

3.4. Estimating the Number of Frames

We evaluate the importance of the consecutive observations by recursively estimating the Fisher information [7]. Given the observations denoted by $\mathbf{Y}$, the Fisher information matrix is

$$J(\theta) = E_\theta[(\bigtriangledown_\theta \ln(f_\theta(\mathbf{Y})))(\bigtriangledown_\theta \ln(f_\theta(\mathbf{Y})))^T] \tag{7}$$

where θ is the parameter to be estimated given the observations, [3] We estimate the Fisher information using simultaneous perturbation for the gradient approximation and averaging for the expectation operation [8]. For the observation model $X = \theta + N, N \sim f_N(n)$, [4] where N is a random variable with a density f_N denoting the noise in the observations, we can write

$$\begin{aligned} \frac{d}{d\theta}\log f_X(x) &= \frac{d}{d\theta}\log f_N(x-\theta) \\ &= \frac{d}{dt}\log f_N(t)\frac{dt}{d\theta}, \quad t = x - \theta \\ &= -\frac{1}{f_N(t)}\frac{df_N(t)}{dt}. \end{aligned}$$

The estimate of the gradient of $f(t)$ with respect to $t \in \mathcal{R}^p$:

$$\hat{g}(t) = \frac{f(t+\Delta) - f(t-\Delta)}{2}\begin{bmatrix} \Delta_1^{-1} \\ \vdots \\ \Delta_p^{-1} \end{bmatrix} \tag{8}$$

where $\Delta = (\Delta_1, ..., \Delta_p)$ and the components of Δ are independent Bernoulli random variables. The steps in computing the Fisher information are:

Step 1 Given $\hat{\theta}_k$ in (4), generate a set of k pseudo measurements according to the empirical distribution of the observations. Denote these by $x_{pseudo}(k)$. Calculate the gradient according to (8). It may be necessary to average several gradient estimates with independent values of Δ. Compute the term within the expectation operator in the definition of Fisher information (7).

Step 2 Repeat Step 1 a large number of times. Average the estimates obtained. This is the estimate of the Fisher information, $\hat{F}_k(\hat{\theta}_k)$.

We can evaluate the relative importance of the observations by analyzing the increase in Fisher information (see Fig. 2).

3.5. The Algorithm

Assume that we have the fused 3D structure S_i obtained from i frames and the 2-frame depth map s_{i+1} computed from the i and $(i+1)$-th frames. The main steps of the algorithm are

Track Estimate the camera motion according to (6).

Transform Transform the previous model S_i to the new reference frame.

Update Update the transformed model using s_{i+1} to obtain S_{i+1} using (4).

Compute Fisher Information Compute the Fisher Information.

[2] We used the commonly chosen gain sequence $a_k = 0.1/(k+1)^{.501}$.

[3] E_θ represents expectation with respect to θ and $\bigtriangledown_\theta$ represents the gradient with respect to θ.

[4] x is the realization of a random variable X

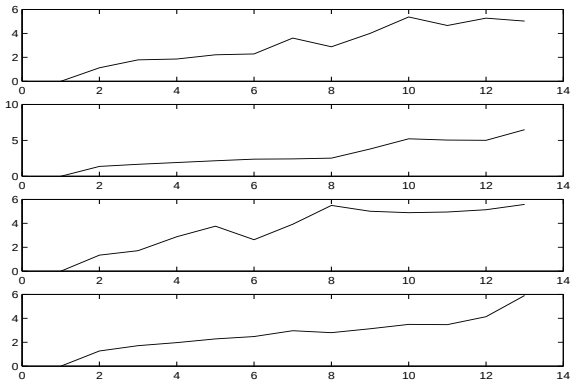

Fig. 2. The figure shows the variation of the Fisher information (FI) over increasing frames.

Fig. 3. The first two columns show the first and last frames used to compute the depth. The last two columns represent views from camera positions not part of the original sequence.

Iterate If increase in Fisher information is small, stop. Else set $i = i + 1$ and go back to Track.

4. RESULTS AND ANALYSIS

We applied our algorithm for 3D modeling of human faces from 2D images. Given a sequence of images, we used the two frame algorithm described in [2] to obtain the depth map and the motion estimates. In this method, a fast partial search is used to compute the motion and structure. Given a set of hypotheses for the focus of expansion, the least squares error of the system is computed using Fourier transform techniques. The algorithm takes $\mathcal{O}(N^2 \log N)$ operations for a $N \times N$ flow field. The two-frame depths obtained from this algorithm were then fused by the method described above. At each stage, the fused estimate was transformed to the new coordinate system of the next pair of images, using the estimates of the camera motion tracked till that frame. A 3D model was created by interpolating the values at the pixels at which the depth was not obtained. From this model, we synthesized views which are not part of the original image sequence (Fig. 3). To illustrate the point that fusion improves upon the individual observations, we plot the two frame and fused depth maps in Fig. 4.

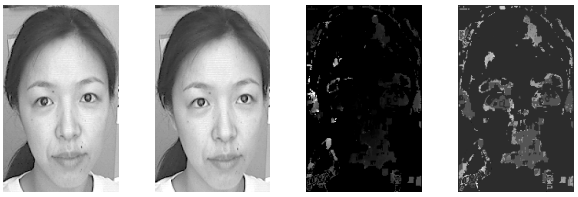

Fig. 4. The first two columns show the first and last frames used to compute the depth. The third column shows the depth map from two frames and the last figure represents fused depth map.

5. CONCLUSION

In this paper we have presented a recursive algorithm for fusing two-frame depth estimates over time using stochastic approximation techniques. We also demonstrate how to incorporate time-varying dynamics to track the camera motion. Our method is independent of the underlying two-frame algorithm and does not require separate computation of the two-frame error. The method is robust to stray erroneous values in the depth and the estimate converges to the true value given a sufficiently large number of frames. The number of frames is estimated by recursively computing the Fisher information of the observations. The work was applied to the modeling of human faces and results have been presented.

6. REFERENCES

[1] J. Oliensis, "A critique of structure from motion algorithms," *NECI TR*, 1997.

[2] S. Shridhar, "Extracting structure from optical flow using fast error search technique," *CfAR Technical Report, University of Maryland, CAR-TR-893*, 1998.

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

[4] J.Inigo Thomas and J. Oliensis, "Dealing with noise in multiframe structure from motion," *Computer Vision and Image Understanding*, vol. 76(2), pp. 109–124, 1999.

[5] P.J.Rousseeuw, "Least median of square regression," *Journal of the American Statistical Association*, vol. 79, pp. 871–880, 1984.

[6] Lenart Ljung and Torsten Söderstörm, *Theory and Practice of Recursive Identification*, MIT Press, 1987.

[7] R.E. Blahut J.A.O'Sullivan and D.L. Snyder, "Information theoretic image formation," *IEEE Trans. on Information Theory*, vol. 44(6), 1998.

[8] J.C.Spall, "Resampling-based calculation of the information matrix for general identification problems," in *Proc. of the American Control Conf., PA*, 1998.