A ROBUST ALGORITHM FOR FUSING NOISY DEPTH ESTIMATES USING STOCHASTIC APPROXIMATION

Amit K. Roy Chowdhury and Rama 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. Most of the algorithms which work by fusing the two-frame depth estimates (observations) assume an underlying statistical model for the observations and do not evaluate the quality of the individual observations. However, in real scenarios, it is often difficult to justify the statistical assumptions. Also, outliers are present in any observation sequence and need to be identified and removed from the fusion algorithm. In this paper, we present a recursive fusion algorithm using Robbins-Monro stochastic approximation (RMSA) which takes care of both these problems to provide an estimate of the real depth of the scene point. The estimate converges to the true value asymptotically. We also propose a method to evaluate the importance of the successive observations by computing the Fisher information (FI) recursively. Though we apply our algorithm in the SFM problem by modeling of human face, it can be easily adopted to other data fusion applications.

1. INTRODUCTION

The SfM algorithms extract the 3-D 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. Recently, techniques have been developed that use multiple images for scene reconstruction, achieving greater robustness and accuracy by fusing the two-frame estimates (multi-frame SfM or MFSfM) [3], [4], [5], [6], [7].

One obvious strategy in MFSfM algorithms is the *integration over time* approach [7]. 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 [6]. However, 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.

This paper describes a new data fusion algorithm applied to multi-frame structure from motion using stochastic approximation (SA). The method can be easily extended to other data fusion applications also. We assume that the 2-frame estimates (observations) are available from a suitable 2-frame SfM algorithm. [1]

We propose a recursive strategy for estimating the true depth given the two-frame observations. The method does not require the statistics of the error in the observations and takes care of eliminating outliers by using the least median of squares estimator instead of the more conventional least mean square. We use the Robbins-Monro stochastic approximation [8] technique as a solution to the problem. The estimates obtained by this method asymptotically converge to the true value. We also propose a method for evaluating the importance of successive observations (i.e. the number of frames to consider) by evaluating the Fisher Information (FI) recursively. The results of our algorithm are demonstrated by applying them on image sequences of a human face.

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 in some coordinate system. Before we venture to describe the algorithm, a few

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.

[1] The particular two-frame algorithm chosen here is the one described in [2] because of its speed of computation, but our fusion algorithm is not specific to this particular method.

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 their application 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 Gaussian 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.

Robust Estimators Fig. (2) plots the depth values across 50 frames for four randomly chosen points in an image. It can be seen that there are isolated outliers in all the four cases. Application of least squares estimation techniques in the presence of such outliers will severely affect the estimation technique. One bad datum is sometimes enough to perturb least squares completely. 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) [9]. The median is a preferred estimator as it has high breakdown point. Also, the efficiency of LMedS is poor in the presence of Gaussian noise. Since the noise in the structure estimates deviates appreciably from Gaussianity, as discussed above, it is a good choice for our application.

Two-frame Algorithm The SfM estimate is an extraordinarily complex function of image data and to expect a single error analysis to work in all domains would be naive. It is thus desirable that we design algorithms for specific problem domains. Our algorithm will work well under the conditions where the error analysis of the underlying two-frame algorithm conforms approximately to the experimental observations outlined above. It is more general than [6] in the sense that it does not require us to compute the error covariance for every two-frame algorithm separately; in fact, since it does not require an explicit expression for the error, it can be applied even when such a computation is very difficult or impossible. Another point that needs to be borne in mind is that fusion is essentially no more robust than the few-image intermediate reconstructions it is based on. Though fusion can improve the result of reasonably accurate intermediate reconstructions, it can also fail miserably when they are not.

On the basis of the observations outlined above, the cost function we optimize is

$$u^* = \arg\min_u \left(median(d_i - u)^2 \right), \qquad (1)$$

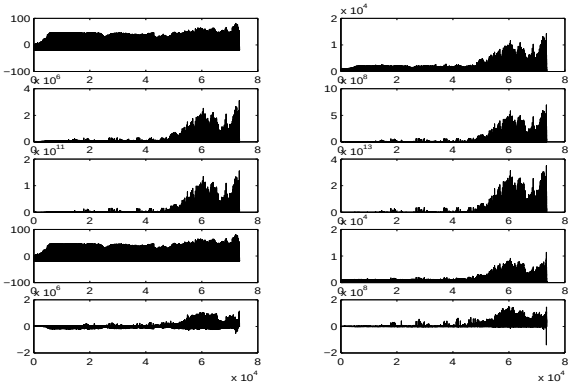

Fig. 1. The top six figures plot the estimates of the first six moments of the observation vector and the bottom four figures plot the first four cumulants. The horizontal axis represents the pixel number. The first column represents the odd central moments/cumulants and the second column the even ones.

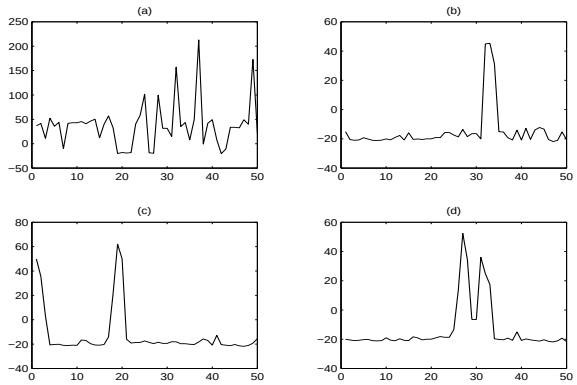

Fig. 2. A plot of the depth values across 50 frames for four randomly chosen points. It can be seen that there are isolated outliers in all the four cases.

where $\{d_i\}$ is the sequence of depth values, u belongs to a predefined search set $\mathcal{U}$ consisting of possible values of the true depth at that point and u^* is the optimal value obtained after the minimization. The disadvantage of this method is that we no longer have a closed form solution for the optimal estimate as we had for the mean-square error criterion.

3. A RECURSIVE ALGORITHM USING STOCHASTIC APPROXIMATION

3.1. The Robbins-Monro Algorithm

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 as-

sumed 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,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k Y_k(\hat{\theta}_k). \tag{2}$$

where a_k is an appropriately chosen sequence. Details of the algorithm can be found in [8]. We will outline the method for obtaining the solution for our specific problem. Suppose that $F_X(x)$ is the unknown distribution of a sequence of observations $X_0, X_1, \ldots$ and we are interested in finding the root of the equation $g(\theta) = F_X(\theta) - 0.5 = 0$, i.e. the median of the distribution. Since

$$\begin{aligned} E[Y_k(\hat{u}_k)|\hat{u}_k] &= E[s_k(\hat{u}_k)|\hat{u}_k] - 0.5 \\ &= E[\mathbf{I}_{[X_k \le \hat{u}_k]}] - 0.5 \\ &= P(X_k \le \hat{u}_k) - 0.5 \\ &= F_X(\hat{u}_k) - 0.5 = g(\hat{u}_k), \end{aligned}$$

the Robbins-Monro recursion is as follows [8]:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k(s_k(\hat{\theta}_k) - 0.5) \tag{3}$$

where $s_k(\hat{\theta}_k) = \mathbf{I}_{[X_k \le \hat{\theta}_k]}$ (**I** represents the indicator function). The choice of the gain sequence a_k is determined by the convergence properties of the algorithm [8]. [2]

The sequence in consideration in our case is $X_i(u) = (d_i - u)^2$. The minimization is carried out over a predetermined search set $\mathcal{U}$ and the number of frames is determined by analyzing the Fisher information of the observations. Since the depth observations $\{d_i\}$ are the result of a 2-frame SfM algorithm, they are corrupted by noise whose distribution is in general unknown. However, since RMSA solves for the estimate in the situation where the distribution of the observations is unknown, it is robust enough to deal with our particular problem.

3.2. Convergence Properties of the Algorithm

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 [8]. 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.

3.3. Estimating the Fisher Information

We evaluate the importance of the consecutive observations by recursively estimating the Fisher information [10]. Given the observations denoted by $\mathbf{Y}$, the Fisher information matrix is

$$J(\theta) = E_\theta[(\bigtriangledown_\theta \ln(f_\theta(\mathbf{Y})))(\bigtriangledown_\theta \ln(f_\theta(\mathbf{Y})))^T] \tag{4}$$

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 [11]. 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

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

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

$$\hat{g}(t) = \frac{f(t+\Delta) - f(t-\Delta)}{2}\begin{bmatrix} \Delta_1^{-1} \\ \vdots \\ \Delta_p^{-1} \end{bmatrix} \tag{5}$$

where $\Delta = (\Delta_1, \ldots, \Delta_p)$ 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 (3), 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 (5). 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 (4).

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, and hence the number of frames needed for the recursion, by looking at increase in the Fisher information (see Fig. 3).

4. RESULTS AND ANALYSIS

We applied our algorithm for 3-D modeling of human faces from 2-D images. Given a sequence of images, we used the two frame algorithm described in [2] to obtain the depth map. In this method, a fast partial search is used to compute the motion and structure. The least squares error of the system is computed using Fourier techniques and the focus of expansion is estimated in $\mathcal{O}(N^2 \log N)$ operations for a $N \times$

[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

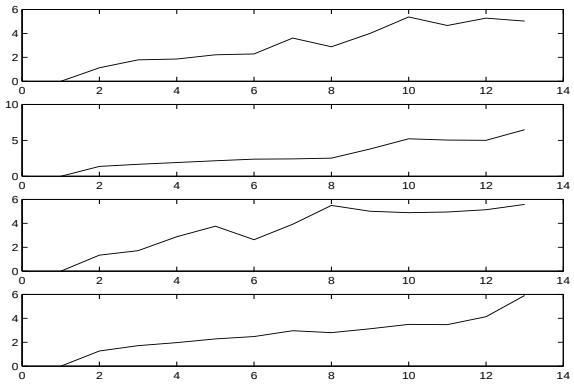

Fig. 3. The figure shows the variation of the Fisher information (FI) over increasing frames.

Fig. 4. 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.

N flow field. The two-frame depths were then fused by the method described above. A 3-D 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. 4). To illustrate the point that fusion improves upon the individual observations, we plot the two frame and fused depth maps in Fig. 5.

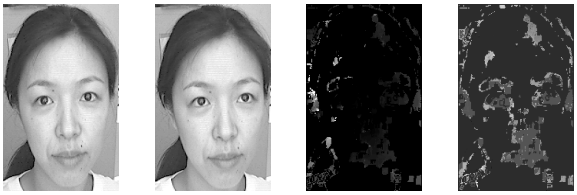

Fig. 5. The first two columns show the first and last frames used to compute the depth. The third column shows the depth map from two frames and the last figure represents fused depth map.

5. CONCLUSION

In this paper we have presented a recursive algorithm for fusing two-frame depth estimates over time using stochastic approximation techniques. Our method is applicable to a large class of two-frame algorithms as it does not require separate computation of the 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] T.J. Broida and R. Chellappa, "Estimating the kinematics and structure of a rigid object from a sequence of monocular images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13(6), pp. 497–513, 1991.

[4] T. Kanade L. Matthies and R. Szeliski, "Kalman filtering algorithms for estimating depth from image sequences," *International Journal of Computer Vision*, vol. 3, pp. 209–236, 1989.

[5] R.Szeliski and S.B.Kang, "Recovering 3d shape and motion from image streams using non-linear least squares," *Journal of Visual Computation and Image Representation*, vol. 5(1), pp. 10–28, 1994.

[6] 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.

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

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

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

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

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