

# ROBUST ESTIMATION OF DEPTH AND MOTION USING STOCHASTIC APPROXIMATION

Amit K. Roy Chowdhury, Rama Chellappa

Center for Automation Research  
Department of Electrical and Computer Engineering  
University of Maryland, College Park  
MD 20742, USA  
{amitrc,rama}@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 and image correspondences, 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. 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 and can reconstruct the scene to arbitrary accuracy given a sufficient number of frames. Experimental results are reported to support these claims.

## 1. INTRODUCTION

The problem of structure from motion 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 and poor image correspondences, thus necessitating multi-frame algorithms (MFSfM). We describe a recursive fusion algorithm to estimate the 3D structure and camera motion from a sequence of images, given the estimates from every consecutive pair

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.

of images. The technique uses ideas from stochastic optimization to obtain a robust estimate. With this method, it is possible to reconstruct the scene to an arbitrary accuracy given a sufficiently large number of frames. The basic SfM equations we will consider throughout this paper are [3]:

$$\begin{aligned} p(x, y) &= (x - x_f)h(x, y) + xy\omega_x - (1 + x^2)\omega_y + y\omega_z \\ q(x, y) &= (y - y_f)h(x, y) + (1 + y^2)\omega_x - xy\omega_y - x\omega_z, \end{aligned} \quad (1)$$

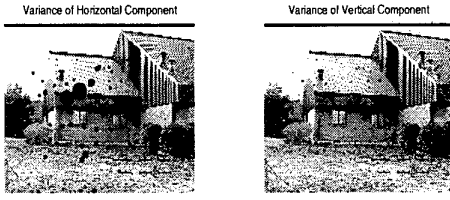
where  $p(x, y)$  and  $q(x, y)$  are the horizontal and vertical velocity fields of a point  $(x, y)$  in the image plane,  $\mathbf{V} = [v_x, v_y, v_z]$  and  $\mathbf{m} = [\omega_x, \omega_y, \omega_z]$  are the translational and rotational motion vectors respectively,  $(x_f, y_f) = (\frac{v_x}{v_z}, \frac{v_y}{v_z})$  is the *focus of expansion* (FOE) and  $h(x, y) = \frac{t_x}{Z(x, y)}$ , where  $Z(x, y)$  is the scene depth<sup>1</sup>.

## 2. STATISTICAL ANALYSIS OF TWO-FRAME RECONSTRUCTION

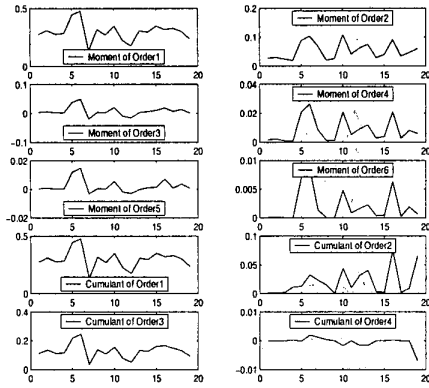
**Observation Statistics** In Fig. 2, we plot the estimates of the first six moments and the first four cumulants of the two-frame depths values. For Gaussian random variables, all odd central moments are identically zero and all cumulants greater than two are zero, which is not the case as seen from the figure. The estimated skewness is -0.25 and the kurtosis is 1.9, which indicates that the distribution is far from Gaussian [4]. We also plot the estimated variances of the different features along the horizontal and vertical directions separately in Figure 1<sup>2</sup>. The features were selected and tracked across 50 frames using the standard KLT tracker [5]. The variances were computed separately for the horizontal and vertical components using *bootstrap*-

<sup>1</sup>All linear dimensions are normalized in terms of the focal length  $f$  of the camera.

<sup>2</sup>The variance of the vertical components is small since the motion was approximately restricted to the horizontal plane.



**Fig. 1.** Plot of the variances of the image correspondences for different features in an outdoor sequence. The diameter of the circle is proportional to the variance of that feature point.



**Fig. 2.** Plot of estimates of the moments and cumulants of the two-frame depth for the outdoor house sequence of Fig. 1 against the feature points. Skewness = 1.1; Kurtosis = 3.2  $\Rightarrow$  right skewed and peaked distribution.

*ping* techniques with 200 bootstrap samples [6]. Analysis of the figures suggests that **the uncertainty in the image correspondences is a function of the feature point.**

**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 [7]. Experiments have shown that the LMedS estimator is very robust to outliers due to both bad localization and false matches [8]. Also, because of the non-Gaussianity of the noise, the LMedS will be an efficient estimator [4].

We will now derive an expression relating the covariance of the structure and motion estimates to the covariance in the image correspondences. Recall equation (1). Consider that the FOE is known<sup>3</sup>. Then following the notation of [2] (re-

<sup>3</sup>In most video sequences, the FOE does not change appreciably over a few frames; thus it can be estimated from the first two/three frames and

fer here for the details of the notation), equation (1) can be written, for  $N$  points, as

$$\mathbf{A}\mathbf{z} = \mathbf{u}. \quad (2)$$

where

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} h \\ \mathbf{m} \end{bmatrix} \\ \mathbf{h} &= (h_1, h_2, \dots, h_N)' \\ \mathbf{u} &= (p_1, q_1, p_2, q_2, \dots, p_N, q_N)' \\ \mathbf{m} &= (w_x, w_y, w_z) \end{aligned} \quad (3)$$

Let  $\mathbf{z} = \psi(\mathbf{u})$ . Expanding  $\psi$  in a Taylor series around  $E[\mathbf{u}]$ , and denoting  $D_\psi(\mathbf{x}) = \frac{\partial \psi}{\partial \mathbf{x}}$ , we can write up to a first order approximation

$$\begin{aligned} \mathbf{R}_z &= E[(\psi(\mathbf{u}) - E[\psi(\mathbf{u})])(\psi(\mathbf{u}) - E[\psi(\mathbf{u})])'] \\ &= E[D_\psi(E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}])'(D_\psi(E[\mathbf{u}]))'] \\ &= D_\psi(E[\mathbf{u}])\mathbf{R}_u D_\psi(E[\mathbf{u}])' \end{aligned} \quad (4)$$

where  $\mathbf{R}_u$  is the covariance matrix of  $\mathbf{u}$ . Now consider the cost function

$$\begin{aligned} C &= \frac{1}{2} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|^2 \\ &= \frac{1}{2} \sum_{i=1}^{n=2N} C_i^2(u_i, \mathbf{z}) \end{aligned} \quad (5)$$

Then using the *implicit function theorem* [9],

$$D_\psi(\mathbf{u}) = -\mathbf{H}^{-1} \frac{\partial \phi}{\partial \mathbf{u}}. \quad (6)$$

where

$$\phi = \frac{\partial C'}{\partial \mathbf{z}}, \quad \text{and} \quad \mathbf{H} = \frac{\partial \phi}{\partial \mathbf{z}}. \quad (7)$$

Thus (4) becomes

$$\mathbf{R}_z = \mathbf{H}^{-1} \frac{\partial \phi}{\partial \mathbf{u}} \mathbf{R}_u \frac{\partial \phi'}{\partial \mathbf{u}} \mathbf{H}'^{-1} \quad (8)$$

and after some simple calculations (similar to that outlined in [8]) and assuming that each feature point as well as the components of the motion vector at each feature point are independent of each other, equation (8) becomes

$$\mathbf{R}_z = \mathbf{H}^{-1} \left( \sum_i \frac{\partial C_i'}{\partial \mathbf{z}} \frac{\partial C_i}{\partial \mathbf{u}} \mathbf{R}_u \frac{\partial C_i'}{\partial \mathbf{u}} \frac{\partial C_i}{\partial \mathbf{z}} \right) \mathbf{H}'^{-T}, \quad (9)$$

where  $\mathbf{R}_u = \text{diag}[R_{u1o}, R_{u1e}, \dots, R_{uNo}, R_{uNe}]$ , and which gives a precise relationship between the uncertainty of the image correspondences  $\mathbf{R}_u$  and the uncertainty of the depth assumed known thereafter.

and motion estimates  $\mathbf{R}_z$ <sup>4</sup>. Defining  $\bar{i} = \lceil i/2 \rceil$ , representing the upper ceiling of  $i$ , ( $\bar{i}$  will then represent the feature points  $N$  and  $i = 1, \dots, n = 2N$ ).

$$\begin{aligned} A_{\bar{i}o} &= [-(x_i - x_f)\mathbf{1}_{\bar{i}} | -\mathbf{r}_i] = [A_{\bar{i}oh} | A_{\bar{i}em}] \\ A_{\bar{i}e} &= [-(y_i - y_f)\mathbf{1}_{\bar{i}} | -\mathbf{s}_i] = [A_{\bar{i}eh} | A_{\bar{i}em}] \end{aligned}$$

where  $\mathbf{1}_n$  denotes a  $\mathbf{1}$  in the  $n^{\text{th}}$  position of the array and zeros elsewhere and  $\mathbf{r}_i = [x_i y_i, -(1 + x_i^2), y_i]$  and  $\mathbf{s}_i = [1 + y_i^2, -x_i y_i, -x_i]$ . Substituting our cost function from (5), we get

$$\frac{\partial C_i}{\partial \mathbf{z}} = \begin{cases} A_{\bar{i}o}, & i \text{ odd} \\ A_{\bar{i}e}, & i \text{ even} \end{cases}, \quad (10)$$

as a  $1 \times (N+3)$  dimensional vector and

$$\begin{aligned} \frac{\partial C_i}{\partial \mathbf{u}} &= \begin{bmatrix} \frac{\partial C_i}{\partial p_1} & \frac{\partial C_i}{\partial q_1} & \dots & \frac{\partial C_i}{\partial p_N} & \frac{\partial C_i}{\partial q_N} \end{bmatrix}, \\ &= \mathbf{1}_i, \end{aligned} \quad (11)$$

as a  $1 \times 2N$  dimensional array. Hence the Hessian from (9) becomes

$$\mathbf{H} = \sum_{i \text{ odd}} A_{\bar{i}o}' A_{\bar{i}o} + \sum_{i \text{ even}} A_{\bar{i}e}' A_{\bar{i}e} \quad (12)$$

and

$$\mathbf{R}_z = \mathbf{H}^{-1} \left( \sum_{i \text{ odd}} A_{\bar{i}o}' A_{\bar{i}o} R_{u\bar{i}o} + \sum_{i \text{ even}} A_{\bar{i}e}' A_{\bar{i}e} R_{u\bar{i}e} \right) \mathbf{H}^{-T} \quad (13)$$

A similar expression can be derived when the FOE is unknown by redefining

$$\begin{aligned} A_{\bar{i}o} &= [-(x_i - x_f)\mathbf{1}_{\bar{i}} \quad h_{\bar{i}} \quad 0 \quad | -\mathbf{r}_i], \\ &= [A_{\bar{i}oh} | A_{\bar{i}em}], \\ A_{\bar{i}e} &= [-(y_i - y_f)\mathbf{1}_{\bar{i}} \quad 0 \quad h_{\bar{i}} \quad | -\mathbf{s}_i]. \\ &= [A_{\bar{i}eh} | A_{\bar{i}em}] \end{aligned}$$

Because of the partition of  $\mathbf{z} = [\mathbf{h} | \mathbf{m}]$ ,  $\mathbf{R}_z$  can be partitioned as

$$\mathbf{R}_z = \begin{bmatrix} \mathbf{R}_h & \mathbf{R}_{hm} \\ \mathbf{R}_{hm}^T & \mathbf{R}_m \end{bmatrix}. \quad (14)$$

### 3. THE ESTIMATION TECHNIQUE

#### 3.1. Problem Formulation

In our notational convention, subscripts will refer to feature points; superscripts will refer to frame number. Thus  $x_i^j$  refers to the variable  $x$  for the  $i$ -th feature point in the  $j$ -th frame.

<sup>4</sup> $\mathbf{A}^T$  and  $\mathbf{A}'$  both represent the transpose of  $\mathbf{A}$ .

When either of them is omitted, it means that the expressions are valid irrespective of the omitted feature point or frame number. 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,  $i = 1, \dots, K$ , ( $K+1$ ) 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.  $R$  is the rotation matrix corresponding to  $\Omega$ . We represent the motion components by the vector  $\mathbf{m} = [\omega_x, \omega_y, \omega_z, \frac{v_x}{v_z}, \frac{v_y}{v_z}]$  if FOE is unknown or  $\mathbf{m} = [\omega_x, \omega_y, \omega_z]$  if the FOE is known. If  $\{d^i\}$  is the transformed sequence of depth values with respect to a common frame of reference, then the optimal value of the depth at the point under consideration is obtained as

$$u^* = \arg \min_u (\text{median}(d^i - u)^2) \quad (15)$$

#### 3.2. Depth Estimation

The Robbins-Monro stochastic approximation (RMSA) algorithm is a stochastic search technique for finding the root  $\theta^*$  to  $g(\theta) = 0$  based on noisy measurements of  $g(\theta)$ , i.e.  $Y_k(\theta) = g(\theta) + e_k(\theta)$ ,  $k = 1, \dots, K$ , where  $e_k(\theta)$  is assumed to be an **arbitrary 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 [10]. Recall equation (15). For each point, we compute  $X^i(u) = (d^i - u)^2$ ,  $u \in \mathcal{U}$ . Then we need to compute the median (say  $\theta$ ) of  $X^0, \dots, X^K$  with an unknown distribution  $F_X$ , i.e. obtain  $\theta$  such that  $g(\theta) = F_X(\theta) - 0.5 = 0$ . Defining  $Y^k(\hat{\theta}^k) = s^k(\hat{\theta}^k) - 0.5$  and  $s^k(\hat{\theta}^k) = \mathbf{I}_{[X^k \leq \hat{T}^k(\hat{\theta}^k)]}$  ( $\mathbf{I}$  represents the indicator function and  $\hat{T}^k$  is the estimate of the camera motion), the the Robbins-Monro (RM) recursion for the problem is (refer to [11] for details of the algorithm):

$$\hat{\theta}_{k+1} = \hat{T}_k(\hat{\theta}_k) - a_k (s_k(\hat{\theta}_k) - 0.5) \quad (16)$$

#### 3.3. Camera Motion Tracking

Our discrete-time dynamical model of the camera motion is:

$$\begin{aligned} \mathbf{m}^i &= \mathbf{m}^{i-1} + \mathbf{w}^i, \\ \mathbf{y}^i &= \mathbf{m}^i + \mathbf{v}^i. \end{aligned} \quad (17)$$

$\mathbf{w}$  is modeled as a zero mean white noise process with  $E[\mathbf{w}^i \mathbf{w}^j] = \mathbf{Q}^i \delta(i, j)$ . The observations of the camera motion (output of the two-frame algorithm),  $\mathbf{y}^i$  are corrupted by a zero mean, noise process  $\mathbf{v}^i$ , with a diagonal covariance matrix  $\mathbf{V}^i$ .  $\mathbf{v}$

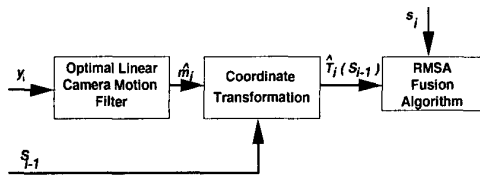


Fig. 3. Block diagram of the multi-frame fusion algorithm.

and  $w$  are assumed to be mutually uncorrelated across all instants of time, i.e.  $E[v^i w^j] = 0$ , for all  $(i, j)$  and are also independent of the parameter  $m^i$  at all time instants. We are interested in designing a linear mean square error (LMSE) estimator of the camera motion  $m^t$  based on the observations  $y = [y^t, y^{t-1}, \dots, y^{t-k+1}]'$ . Let  $\hat{m}^{t|s}$  denote the estimate of  $m^t$  based on the observations  $[y^1, \dots, y^s]$  and  $\Sigma^{t|s} = E[(m^t - \hat{m}^{t|s})(m^t - \hat{m}^{t|s})']$ . Then the LMSE estimate can be obtained from the Kalman filtering algorithm as follows. Re-indexing the observation vector  $y$  as  $[y^k, \dots, y^1]$ , the Kalman filter is given by the following recursion [12]

$$\begin{aligned} \hat{m}^{k|k} &= \hat{m}^{k|k-1} + K^k (y^k - \hat{m}^{k|k}) \\ \hat{m}^{k|k-1} &= \hat{m}^{k-1|k-1} \\ K^k &= \Sigma^{k|k-1} [V^k + \Sigma^{k|k-1}]^{-1} \\ \Sigma^{k|k-1} &= \Sigma^{k-1|k-1} + Q^k. \end{aligned} \quad (18)$$

Then  $\Sigma_{y^k} = E[(y^k - E[y^k])(y^k - E[y^k])'] = E[(m^k + v^k - \mu_m)(m^k + v^k - \mu_m)'] = E[(m^k - \mu_m)(m^k - \mu_m)'] + V^k = R_m^k$ , where  $\mu_m = E[m^i] = E[m^{i-1}] = E[y^i]$ <sup>5</sup>. Thus the observation noise covariance can be obtained from (14) and the camera motion filter derived.

Figure (3) depicts a block diagram of the multi-frame fusion algorithm.

#### 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. 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 using the MATLAB Graphics toolbox. From this model, we synthesized views which are not part of the original image sequence (Fig. 4).



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.

#### 5. CONCLUSION

In this paper we have presented a recursive algorithm for fusing two-frame depth estimates over time using stochastic approximation techniques and tracking the camera motion using an optimal motion filter. Our method is independent of the underlying two-frame algorithm and takes into consideration the statistics of the errors in the intermediate reconstructions. 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 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*, vol. http://www.neci.nj.nec.com/homepages/oliensis/, 2000.
- [2] S. Shridhar, "Extracting structure from optical flow using fast error search technique," *CJFAR Technical Report, University of Maryland, CAR-TR-893*, 1998.
- [3] Vishvijit Nalwa, *A Guided Tour of Computer Vision*, Addison Wesley Publishing Company, 1993.
- [4] Jun Shao, *Mathematical Statistics*, Springer, 1998.
- [5] C. Tomasi and J. Shi, "Good features to track," in *CVPR94*, 1994, pp. 593-600.
- [6] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993.
- [7] P.J. Rousseeuw, "Least median of square regression," *Journal of the American Statistical Association*, vol. 79, pp. 871-880, 1984.
- [8] Z.Y. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *IJCV*, vol. 27, no. 2, pp. 161-195, March 1998.
- [9] Rudin Walter, *Principles of Mathematical Analysis, 3rd Edition*, McGraw-Hill Inc., 1976.
- [10] Lenart Ljung and Torsten Soderstrom, *Theory and Practice of Recursive Identification*, MIT Press, 1987.
- [11] A. Roy Chowdhury and R.Chellappa, "A robust algorithm for fusing noisy depth estimates using stochastic approximation," in *Proc. IEEE ICASSP-01*, 2001.
- [12] H.V. Poor, *An Introduction to Signal Detection and Estimation*. Springer Verlag, New York, 1988.

<sup>5</sup>The same notation is used for the random variable and its realization.