Integration of Transitory Image Sequences

John Weng and Yuntao Cui
Computer Science Department
Michigan State University
East Lansing, MI 48824

Narendra Ahuja
Beckman Institute
University of Illinois
Urbana, IL 61801

Ajit Singh
Siemens Corporate Research
755 College Road East
Princeton, NJ 08540

Abstract

A transitory image sequence is one in which no scene element is visible through the entire *sequence. This article deals with some major theoretical and algorithmic issues associated with the task of estimating structure and motion from transitory image sequences. Two representations, world-centered (WC) and camera-centered (CC), behave very differently with a transitory sequence. The asymptotical error properties derived in this article indicate that one representation is significantly superior to the other, depending on whether one uses camera-centered or world-centered estimates. Rigorous experiments were conducted with real-image sequences taken by a fully calibrated camera system. The comparison of the experimental results with the ground truth has demonstrated that a good accuracy can be obtained from transitory image sequences.*

1 Introduction

So far, most works deal with non-transitory image sequences, and successful improvements have been achieved in this type of fusion (*e.g.*,Ayache and Faugeras [2], Matthies and Shafer [4], Kumar *et al* [3]). Experiments for scene construction from transitory image sequence only started recently, and we have seen two efforts by Cui *et al* [1] and Tomasi and Kanade [5], respectively. In Cui *et al* [1], some relative accuracy was reported from a transitory image sequence. Tomasi and Kanade [5] conducted experiments with transitory image sequences and discussed how to expand the measurement matrix by filling in "hallucinated" projections.

The work reported here addresses following new issues:

1. It is shown that from a transitory sequence it is inherently not possible to get better estimates with a longer sequence. We establish asymptotic behavior of error with respect to the number of frames.
2. This article introduces different techniques for two different usages of the result: global and local. It is demonstrated that different representations result in very different stabilities.
3. Rigorous experiments have been conducted with a fully calibrated camera system. The algorithm is fully automatic, including feature selection, stereo matching, temporal matching and tracking, 3-D structure integration, and motion and pose estimation.

We first introduce in the next section some basic concepts related to transitory and non-transitory sequences and analyze the stability and asymptotic error behavior for different representations, which motivated our new cross-frame methods presented in Section 3. The experimental results are presented in Section 4. Section 5 gives concluding remarks.

2 Transitory Image Sequence

We consider a rigid scene and a sensing system, which undergo a motion relative to each other. If the system of reference is placed on the scene, the representation is called world-centered (WC). If the system is placed on the camera system, the representation is called camera-centered (CC).

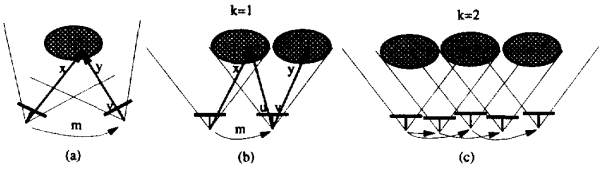

Figure 1: Transitory and non-transitory sequences. (a) non-transitory. (b) simple transitory. (c) general transitory.

A view of a 3-D feature point x is a 2-vector in monocular case and a 4-vector in stereo case. The covariance matrix of a 3-D point from a monocular view can be represented by Γ_x.

First, we examine the error from determination of the pose m of a camera system in a system of reference, where m is a 6-vector. The pose is estimated from x, a set of 3-D points, represented in that reference system, and the view u of x. We can express the error covariance as:

$$\Gamma_m = \frac{\partial m}{\partial x}\Gamma_x\frac{\partial m^\top}{\partial x} + \frac{\partial m}{\partial u}\Gamma_u\frac{\partial m^\top}{\partial u} \qquad (1)$$

assuming that the correlation between x and u is negligibly small.

Next, we investigate the error in determining 3-D position of a set of 3-D points y visible from a camera system whose estimated pose is m. Its covariance matrix is

$$\Gamma_y = \frac{\partial y}{\partial m}\Gamma_m\frac{\partial y^\top}{\partial m} + \frac{\partial y}{\partial v}\Gamma_v\frac{\partial y^\top}{\partial v} \qquad (2)$$

where v is the view and we assume that the correlation between error in m and v is negligibly small.

Plug (1) into (2), we have

$$\Gamma_y = A\Gamma_x A^\top + B\Gamma_u B^\top + C\Gamma_v C^\top \quad (3)$$

where A, B and C are the appropriate Jacobians.

In a simple transitory sequence, each scene point is visible in two consecutive frames, as shown in Fig.1(b). We can use (3) recursively, we can get the error covariance of the structure x at frame n

$$\Gamma_{x_n} = \sum_{t=1}^{n} (B_t \Gamma_{u_t} B_t^\top + C_t \Gamma_{v_t} C_t^\top)$$

where, B_t and C_t are the products of the appropriate Jacobians. Now, we assume a uniformity in which the difference among the terms under the summation is neglected. Thus,

$$\Gamma_{x_n} = n(B_t \Gamma_{u_t} B_t^\top + C_t \Gamma_{v_t} C_t^\top) \quad (4)$$

The general situation with a transitory sequence is shown in Fig.1(c), where a point can be visible in any number of frames. Because we are interested in asymptotic error behavior, we may make some assumption about uniformity. Assume that every feature point is visible in $2k$ frames. Thus, we regard the entire sequence $F = \{f_t \mid t = 1, 2, ...n\}$ as k subsequences $F_l = \{f_{pk+l} \mid p > 0 \text{ is an integer}\}$, $l = 1, 2, ..., k$, so that in each F_l each point is visible by two frames. k is called the visibility span. The entire sequence consists of k subsequences and each is a simple transitory sequence of n/k long. According to the result of simple-transitory case with the uniformity assumption, the error covariance matrix of the linear minimum variance estimate is proportional to the length n/k:

$$\Gamma_{x_n} = \frac{n}{k}(B_t \Gamma_{u_t} B_t^\top + C_t \Gamma_{v_t} C_t^\top) \quad (5)$$

where the term in summation should be that for a simple transitory subsequence. On the other hand, we have k subsequences, each gives an independent observation of structure x_t. Thus, we can use the result for ideal non-transitory sequence, which says the error covariance matrix is reduced by a factor of $1/k$:

$$\Gamma_{x_n} = \frac{n}{k^2}(B_t \Gamma_{u_t} B^\top + C_t \Gamma_{v_t} C_t^\top) \quad (6)$$

where the order n/k^2 is called the asymptotic rate.

Other cases are similarly derived. The asymptotic error rates are shown in Table 1. In the table, n is current time, k the visibility span, and b is the batch size $b \leq k$. The pose is the camera pose with respect to the WC system of reference. The CC structure is with respect to the CC system for the currently visible scene. As can be seen from the table, with a general transitory sequence, for global structure representation, the WC representation is better, but for the camera-centered local structure, the CC representation is superior.

Table 1: Asymptotic rate for error covariance matrix due to integration

Estimate	Non	Simple	General
WC structure	$1/n$	n	n/k^2
WC pose	1	n	n/k^2
CC structure (local)	$1/n$	1	$1/b$
CC pose	1	n	n/k^2

3 Cross-frame Approach with CC and WC

The above analysis motivated our method of keeping two representations, WC for global measurements and CC for local measurements. To be specific, we assume a stereo camera system. The method can be directly extended to monocular case without any major modification.

Let X_p denote the 3-D positional vector of a point represented in the CC system at frame p. Point X_q represented in the CC system at frame q is moved to X_p in the CC system at time p: $X_p = R_{p,q}X_q + T_{p,q}$,where $R_{p,q}$ and $T_{p,q}$ are a rotation matrix and a translation vector, respectively. Let $m_{p,q}$ which is a function of $R_{p,q}$ and $T_{p,q}$, denote the relative pose from q to p. With a batch at frame p, the current active cross-frame motion set is denoted by

$$W(p) = \bigcup_{i=p-K-1}^{p-1} \{m_{p,i}(R_{p,i}, T_{p,i})\}.$$

where K is the batch size. The essence of our new cross-frame approach is to estimate cross-frame motion set $W(p)$ as a whole, eliminating error accumulation by conventional frame-by-frame methods (e.g., Kalman filtering).

Let N be the total number of feature points being considered; $x_{i,s}$ denote the three dimensional local structure of i-th point in s-th camera-centered system; $u_{i,j,s}$ be the 2-D image coordinate vector of i-th point on the j-th side (left, right) at the s-th frame. Assuming that the noise in the observations ($u_{i,j,s}$) is uncorrelated and has the same variance (σ_u^2, σ_v^2) in the two image coordinates, the objective function can be written:

$$\min_{\forall x_{i,p}, \forall m \in W(p)} f(m, x_{i,p}) = A + B \quad (7)$$

where

$$A = \sum_{i=1}^{N} \{S_i^\top (R_{p,p-K-1} \Gamma_{x^*_{i,p-K-1}} R_{p,p-K-1}^\top)^{-1} S_i\}$$

with $S_i = (x_{i,p} - X(m_{p,p-K-1}, x^*_{i,p-K-1}))$ and

$$B = \sum_{s=p-K}^{p} \sum_{j=L}^{R} \sum_{i=1}^{N} C_{s,j,i}^\top \begin{bmatrix} \sigma_u^{-2} & 0 \\ 0 & \sigma_v^{-2} \end{bmatrix} C_{s,j,i}$$

with $C_{s,j,i} = \hat{u}_{i,j,s} - u(m_{s,p}, x_{i,p})$.

In the above expression, $X(m_{s,p}, x_{i,p})$ is the transformation function to transform the point $x_{i,p}$ from camera coordinate system at frame p to frame s based on the motion parameters $m_{s,p}$. Function $u(m_{s,p}, x_{i,p})$ is the noise-free projection computed from $m_{s,p}$ and $x_{i,p}$, which includes transformation and projection.

The WC representation follows a similar derivation. The difference is that the structure does not move in WC system. Thus, the structure integrated in the WC system up to any time can be used directly for later WC integration.

4 Experiments

Experiments with synthetic and real word images were conducted in order to experimentally exam the error rates listed in Table 1 and compare the WC and CC representations.

4.1 Simulation Results

3-D feature points were generated randomly for each trial, between depth 2000mm and depth 3000mm, with a uniform distribution. The entire scene is covered by 31 frames and the distance between consecutive frames is roughly 200mm. Data were obtained through 100 random trials each with a different set of 3-D points.

The simulated camera system has a resolution of 512 × 480 pixels, just like the real cameras we used. Measurement error was simulated by pixel round-off error. The camera's global orientation is determined by a rotation matrix ($R_{i,1}$) and the position by translation vector ($T_{i,1}$). The error of a matrix or vector is measured as the Euclidean norm of the difference between the estimated and true one. In the WC representation, the global structure is directly estimated but its local structure needs to be computed via the estimated global pose of the camera. The situation is just the opposite in the CC representation, where the local structure is directly estimated while the global structure must be computed via camera's global pose.

Fig. 6 shows the current camera position error ($R_{i,1}$, $T_{i,1}$) for different frames, where i is the index for time. It can be seen from the figure that the batch size of our cross-frame method has more impact in the CC representation than WC. This is because in the CC representation, the reference frame moves, which introduces more nonlinearity than the WC case when the old observation is transformed into the current CC reference system. A larger batch size is more appropriate for such a nonlinear transform, because covariance for error modeling is based on a linear approximation for nonlinear systems. Fig. 6 clearly shows that the WC representation is a little better for camera pose estimates, which is consistent with Table 1.

4.2 Experiments with a Real Setup

The setup for our image acquisition consists of a Denning MRV-3 mobile robot and a pair of stereo cameras, mounted on a custom-designed stereo positional setup. A stereo image sequence of 151 frames was acquired from the moving mobile robot. A temporally subsampled (one sample every 5 frames) subsequence of 31 frames was used for motion and structure estimation.

Figure 2: A few stereo frames in the 151-frame sequence. (a) frame 0; (b) frame 50; (c) frame 100; (d) frame 150; all left views.

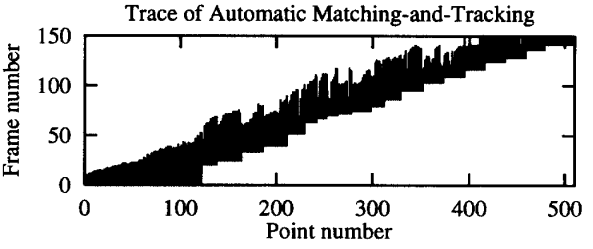

Figure 3: The record of automatic matching and tracking of the feature points through the 151 frames in the sequence. If a point k is successfully tracked from frame i to frame j, a vertical line is shown at point number k from frame i to frame j. (Due to the limit of the printer resolution, all the individual lines form a black region in the plot.)

A feature point detector has been developed for this project to automatically detect feature points from images. Stereo matching was done using an image matching algorithm by Weng *et al* [6], which provides a dense displacement field with a disparity vector for every pixel. This temporal matching and tracking method was very successful. The trace record of the entire sequence is shown in Figure 3. Figure 4 presents an example of stereo and temporal matching.

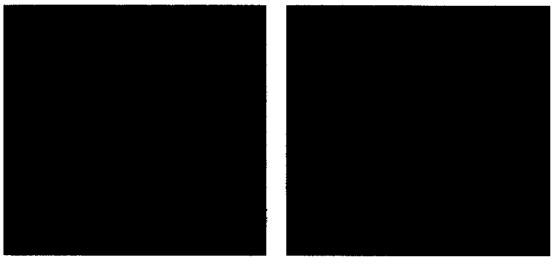

Figure 4: Stereo matching and temporal matching-and-tracking. A white needle is draw from the feature point to its position in the target frame. (a) An example of stereo matching (frame 0). Note that due to camera vergence, the stereo disparities are not large and not always horizontal. (b) An example of temporal matching and tracking (frame 24 to 69).

The measurement of the real setup is similar to the simulation. To verify the accuracy of structure estimates as well as camera pose estimates, the global coordinates of a set of test points were carefully measured to within an error of 1mm. The selection of test points

were based on ease of measurement and was not based on automatically selected features.

Fig. 7 shows the camera position error and the global error of the test points visible at the current time. As we predicted for any transitory image sequence, the error increases with the time. But the estimates appear good. After traveling about 3000mm, the estimated camera global position error is less than 60mm in the most unreliable component Z (less than 2.3%). This seems to indicate that reasonable results can be obtained with a fully automatic algorithm, even with a difficult transitory image sequence.

Fig. 5 shows the reconstructed 3D surface.

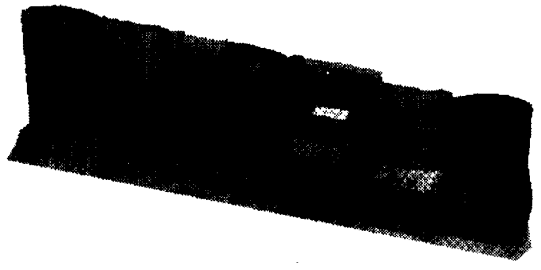

Figure 5: Reconstructed 3D surface shown with original intensity viewed from an arbitrary direction.

5 Conclusions

In this article, we introduced the concept of transitory image sequence for structure and motion estimation from long image sequences. It has been shown that integration for transitory sequence has very different asymptotic error rates from those for a non-transitory image sequence. The analytical results listed in Table 1 indicates that the WC representation is better for global estimates and the CC representation is superior for local estimates.

References

[1] N. Cui, J. Weng, and P. Cohen, Extended structure and motion analysis from monocular image sequences, in *Proc. 3rd Int'l Conf. on Computer Vision*, Osaka, Japan, pp. 222-229, 1990.

[2] N. Ayache and O. Faugeras, Building, registration, and fusing noisy visual maps, in *Proc. 1st Int'l Conf. on Computer Vision*, England, pp. 73-82, 1987.

[3] R. Kumar, H. S. Sawhney and A. R. Hanson, 3D model acquisition from monocular image sequences, in *Proc. IEEE Conf. CVPR*, Champaign, IL, pp. 209-215, 1992.

[4] L. Matthies and S. Shafer, Error Modeling in Stereo Navigation, *IEEE J. of Robotics and Automation*, vol. 3, no. 3, pp. 239-248, 1987.

[5] C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: a factorization method, *Int'l J. of Computer Vision*, vol. 9, no. 2, pp. 137-154, 1992.

[6] J. Weng, N. Ahuja, and T. S. Huang, Two-view matching, in *Proc. 2nd Int'l Conf. on Computer Vision*, pp. 64-73, 1988.

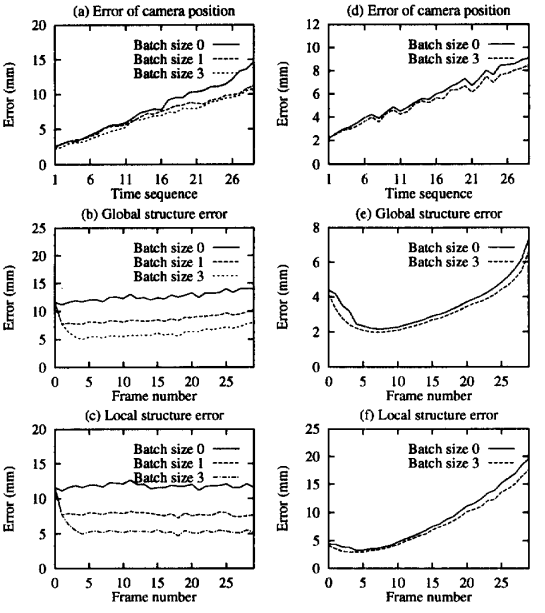

Figure 6: Simulation results. Let column: CC; right column: WC. (a) and (d) Global pose. (b) and (e) Global structure error. (c) and (f) Local structure error.

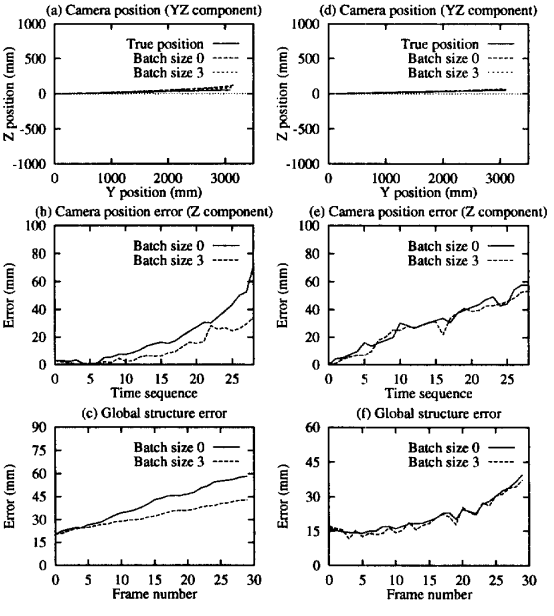

Figure 7: Real setup: camera position and structure error. Let column: CC; right column: WC. (a) and (d) Global pose. (b) and (e) Pose error (z-component). (c) and (f) Global structure error.