Transitory Image Sequences, Asymptotic Properties, and Estimation of Motion and Structure

John (Juyang) Weng, *Member, IEEE*, Yuntao Cui, and Narendra Ahuja, *Fellow, IEEE*

Abstract—A transitory image sequence is one in which no scene element is visible through the entire sequence. When a camera system scans a scene which cannot be covered by a single view, the image sequence is transitory. This article deals with some major theoretical and algorithmic issues associated with the task of estimating structure and motion from transitory image sequences. It is shown that integration with a transitory sequence has properties that are very different from those with a non-transitory one. Two representations, world-centered (WC) and camera-centered (CC), behave very differently with a transitory sequence. The asymptotic error rates derived in this article indicate that one representation is significantly superior to the other, depending on whether one needs camera-centered or world-centered estimates. To establish the tightness of these error rates, it has been shown that these reachable error rates are in fact the lowest possible given by a theoretical lower error bound, the Cramér-Rao error bound. Based on these results, we introduce an efficient "cross-frame" estimation technique for the CC representation. For the WC representation, our analysis indicates that a good technique should be based on camera global pose instead of interframe motions. In addition to testing with synthetic data, 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.

Index Terms— Motion analysis, struction from motion, image sequences, optimal estimation, Cramér-Rao bound, optical flow.

1 Introduction

If a system need to sense a large 3D rigid scene which cannot be covered by single view, the system may actively move and scan the scene. In general, during a dynamic sensing process, any component of the scene is visible only in a subsequence, and thus the resulting image sequence is transitory as we defined in the abstract.

Issues with the transitory nature of scene components have mostly not yet been investigated. Most works deal with non-transitory image sequences, and successful improvements have been achieved in their fusion (e.g., [4], [6], [8], [1]). Experiments for scene construction from transitory image sequence only started recently, and we have so far seen two efforts by Cui et al. [2] and Tomasi and Kanade [9], respectively. In Cui et al. [2], some relative accuracy was reported from a transitory image sequence, which indicated that the accuracy was not further reduced once incoming and exiting feature points are comparable. Tomasi and Kanade [9] conducted experiments with transitory image sequences and discussed how to expand the measurement matrix by filling in "hallucinated" projections. The results showed that the object structure and camera pose constructed from two transitory sequences "Ball" and "Hand" contained larger error than that from the nontransitory sequence "Hotel" [9].

Most questions related to the integration of transitory sequences are still open. The work reported in this article addresses these new issues. The new contribution of this work includes:

1) It is shown that from a transitory sequence, it is inherently not possible to get better estimates with a longer sequence.
2) Techniques are introduced for two different usages: global and local (e.g., visual map generation and global pose determination belong to the former and obstacle avoidance and object manipulation belong to the latter).
3) It is demonstrated that different representations result in very different stabilities. In general, world-centered (WC) is better for a global usage, and camera-centered (CC) is superior for a local usage.
4) We establish asymptotic error rates with respect to the number of frames, which indicates how the error in the estimates evolves with time and how to minimize the pace of error accumulation.
5) We establish that the asymptotic error rates are, in fact, the lowest possible based on the Cramér-Rao error bound.
6) In order to provide actual accuracy with a real system setup, careful experiments have been conducted with a fully calibrated camera system.

- *J. Weng and Y. Cui are with the Computer Science Department, Michigan State University, East Lansing, MI 48824. E-mail: weng@cps.msu.edu; cui@scr.siemens.com.*
- *N. Ahuja is with the Beckman Institute, University of Illinois, Urbana, IL 61801. E-mail: ahuja@vision.ai.uiuc.edu.*

Manuscript received. 20 Sept 1995; revised 2 Jan. 1996. Recommended for acceptance by A. Singh.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P97007.

The results have been compared with ground truth for complete position of test points on the scene and the pose of the camera system. The algorithm is automatic, including feature selection, stereo matching, temporal matching and tracking, 3D structure integration, and motion and pose estimation.

The work presented here seems to be the first to introduce the concept of transitory sequence and provide analytical results of this type of sequence. It reports fully verified accuracy with a real transitory image sequence. The experiment in Cui et al. [2] used a transitory sequence, but no particular attention was paid to the nature of transitory sequence. The work reported here does a systematic study of transitory sequences. Tomasi and Kanade [9] conducted experiments using a few nontransitory and transitory sequences under the orthography assumption. In [9], camera orientation and relative structure error were reported only for nontransitory image sequence "Hotel," but no accuracy was reported for the transitory image sequences.

2 Basic Concepts

We consider a rigid scene and a sensing system (we will call it a camera system). No matter which is actually moving, or both are moving, what we need to consider for the kinematics here is just the relative motion between the two.

We first define the system of reference. Because we are considering two entities: the scene and the camera system, it does not help us to place the system of reference on any object other than these two. If the system of reference is placed on the scene, the representation with respect to this system is called WC (also called object-centered). If the system is placed on the camera system, the representation is called CC. Fig. 1 shows these two representations. In the WC representation, the camera is moving with respect to a static scene, while in the CC representation, the scene is moving relative to a static camera. To be specific in discussion, we say that the scene is static and camera is moving. Thus, the world-centered reference system is fixed (with the scene) and the camera-centered reference system is moving (with the camera).

A view u of a 3D feature point x is a two-vector (two dimensional vector) in monocular case and a four-vector in stereo case (left and right views). With random error in the image measurement u, the 3D position of the point x determined from u becomes a probability distribution whose extent can be characterized by its error covariance matrix Γ_x. The covariance matrix of a 3D point from a monocular view can be represented by

$$\Gamma_x = H\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \sigma_2^2 & 0\\ 0 & 0 & \sigma_3^2\end{bmatrix}H^t$$

where, σ_3 is an extremely large number or infinity and the orthogonal matrix H specifies the orientation of the major axes of the covariance. By using a covariance matrix also for monocular view, we can treat monocular and multi-ocular cases in a unified way. Our analysis is applicable to both perspective and orthographic projections. As a notation, we write a small perturbation of a vector v by δ_v and the error covariance matrix of a vector v by Γ_v.

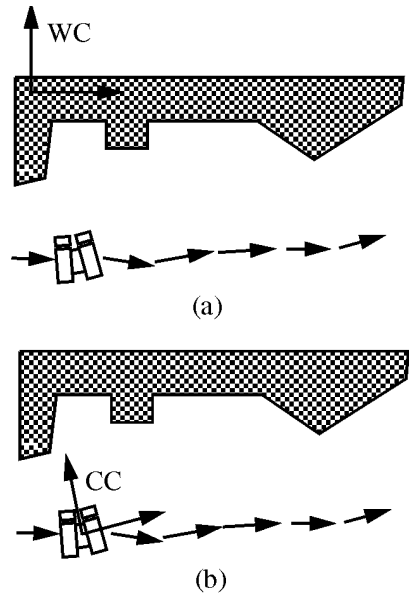


Fig. 1. Two systems of reference. (a) World-centered. (b) Camera-centered.

First, we examine the error from determination of the pose m of a camera system in a system of reference, where p is a six-vector. For example,

$$m = (\theta_x, \theta_y, \theta_z, p_x, p_y, p_z) \tag{1}$$

where p_x, p_y, p_z specifies the position of the camera projection center and θ_x, θ_y, θ_z specifies the orientation of the pose represented by a rotation matrix $R(\theta_x, \theta_y, \theta_z)$. The pose is estimated from x, a set of 3D points, represented in that reference system and u, their image observations in the camera. Therefore, the pose is a function of x and u: $m(x, u)$. We can express the error in m in terms of that in x and u:

$$\delta_m = \frac{\partial m}{\partial x}\delta_x + \frac{\partial m}{\partial u}\delta_u \tag{2}$$

and for its covariance matrix:

$$\Gamma_m = \frac{\partial m}{\partial x}\Gamma_x\frac{\partial m^t}{\partial x} + \frac{\partial m}{\partial u}\Gamma_u\frac{\partial m^t}{\partial u} \tag{3}$$

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

Next, we investigate the error in determining 3D position of a set of 3D points y visible by a camera system whose estimated pose is m. These points in y correspond to a set of image points v. The estimated 3D position of points y in the above system of reference is then a function $y(m, v)$. We can express the error in the estimated y by that of m and v as

$$\delta_y = \frac{\partial y}{\partial m}\delta_m + \frac{\partial y}{\partial v}\delta_v \tag{4}$$

and for its covariance matrix:

$$\Gamma_y = \frac{\partial y}{\partial m}\Gamma_m\frac{\partial y^t}{\partial m} + \frac{\partial y}{\partial v}\Gamma_v\frac{\partial y^t}{\partial v} \tag{5}$$

assuming that the correlation between error in m and v is negligibly small. The above equation indicates that the error covariance of the 3D points has two components, one is caused by the error in the pose estimate, the other results from error in the feature measurements.

Now, we use the above result to analyze pose determination from x and the use of estimated pose m to determine y. We consider two cases:

1) x and y correspond to the same set of scene points, as shown in Fig. 2a. Thus, u and v are the same in (2) and (4), which gives

$$\delta_y = \frac{\partial y}{\partial m}\frac{\partial m}{\partial x}\delta_x + \frac{\partial y}{\partial m}\frac{\partial m}{\partial u}\delta_u + \frac{\partial y}{\partial v}\delta_v = \frac{\partial y}{\partial m}\frac{\partial m}{\partial x}\delta_x + \left(\frac{\partial y}{\partial m}\frac{\partial m}{\partial u} + \frac{\partial y}{\partial v}\right)\delta_v \tag{6}$$

which gives

$$\Gamma_y = A\Gamma_x A^t + D\Gamma_v D^t \tag{7}$$

where A and D are the appropriate Jacobians.

2) x and y correspond to different scene points as shown in Fig. 2b. Substituting Γ_m in (3) for that in (5), it follows that

$$\Gamma_y = A\Gamma_x A^t + B\Gamma_u B^t + C\Gamma_v C^t \tag{8}$$

where A, B, and C are the appropriate Jacobians. The first term is caused by the error in the 3D structure x from which the pose is computed. The second term is due to error in u, the observation of x. The third term results from error in the observation of y.

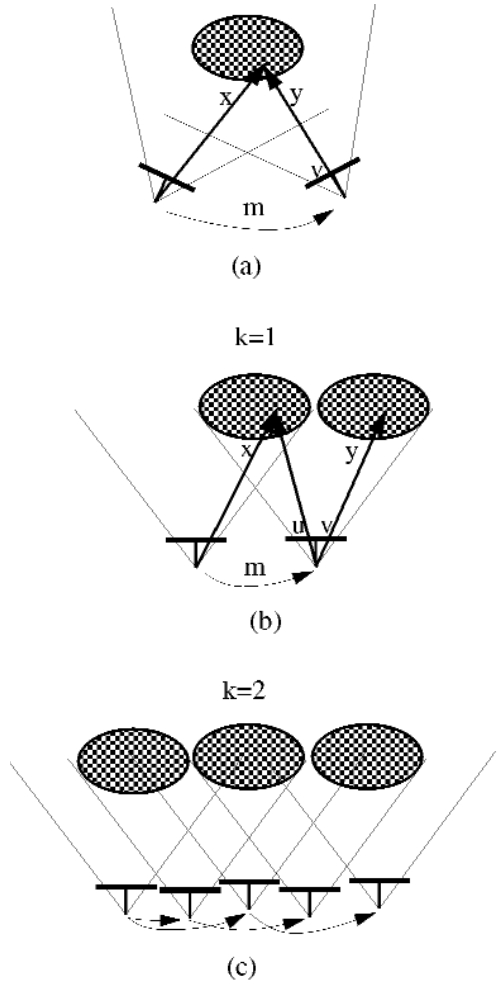


Fig. 2. Transitory and nontransitory sequences. The camera model is represented by the projection center, the image (a thick line), and the field of view. (a) Nontransitory. (b) Simple transitory. (c) General transitory.

3 Asymptotic Error Properties of Different Integrations

In this section, we derive how the amount of error in the estimate changes with integration of various sequences. We assume that the algorithm obtains a linear minimum variance estimate in the sense of Gauss-Markov [5], which is the minimum variance estimate with Gaussian noise.

In order to investigate the best possible result, the processing method is assumed to be batch unless stated otherwise. This means that all the observed data are available for processing and the estimate is computed with all the data as a single batch. In contrast to batch processing is recursive processing [5] where data items are used one at a time, each giving an updated estimate for the result, and once an data item has been used for updating it is discarded. In other words, recursive processing imposes a restriction on the way data are available. Thus, recursive processing may have a worse asymptotic error behavior than the batch processing, unless the problem is actually linear [5].

3.1 World-Centered Representation

In the WC representation, every new observation about object structure is transformed into the WC system of reference using the estimated camera pose. Then all the transformed structure observations are fused together according to each's error covariance matrix.

3.1.1 Ideal Nontransitory Sequence

Consider that a set of feature points y is visible in all the views in the image sequence, as shown in Fig. 2a. Suppose that from $t = 1$ to $t = n$, n observations are made for structure y:

$$y = y_t + \delta_{y_t} \tag{9}$$

Without loss of generality, we can assume that the pose m is relative to the pose at $t = 1$. The correlation of error in δ_{y_t} between different ts is weak because error is random. According to the Gauss-Markov Theorem [5], the linear minimum variance estimator of z in the linear equation $Az = b + \delta$, where the noise term δ has a covariance matrix Γ_δ, is $z = \left(A^t\Gamma_\delta^{-1}A\right)^{-1}A^t\Gamma_\delta^{-1}b$ with an error covariance matrix $\Gamma_z = \left(A^t\Gamma_\delta^{-1}A\right)^{-1}$. Thus, the minimum variance linear estimate for y in (9) is

$$y = \left(\sum_{t=1}^{n}\Gamma_{y_t}^{-1}\right)^{-1}\left(\sum_{t=1}^{n}\Gamma_{y_t}^{-1}y_t\right) \tag{10}$$

where Γ_{y_t} is given in (5). The error covariance matrix of y in (9) is given by

$$\Gamma_y = \left(\sum_{t=1}^{n} \Gamma_{y_t}^{-1} \right)^{-1} \tag{11}$$

It can be shown that if A and B are real symmetric positive definite matrices, then $A - (A^{-1} + B^{-1})^{-1} = A(A + B)^{-1}A$, a positive definite matrix. Using this result, we know from (11) that any observation y_t decreases the expected error in the structure. In order to give a concise and intuitive expression about error covariance matrix, we need to assume some uniformity in the sense that the difference in the error covariance matrix from each view is neglected and each is replaced the average error covariance matrix. Here, we assume that difference of Γ_{y_t} among different t is neglected. Thus,

$$\Gamma y = \left(n\Gamma_{y_t}^{-1} \right)^{-1} = \frac{1}{n}\Gamma_{y_t} = O(1 / n) \tag{12}$$

Thus, it is clear that the expected error variance in the structure is *inversely proportional* to the number of frames n. We call the factor $1/n$ *error rate*.

3.1.2 Simple Transitory Sequence

In a simple transitory sequence, each scene point is visible in two consecutive frames. In this case, the pose m estimated from point set $x_t = x$ and its observation $u_t = u$ is used to estimate the new structure $x_{t+1} = y$ whose observation is $u_{t+1} = v$, as shown in Fig. 2b. From (8), we can estimate the error covariance of the structure x_t:

$$\Gamma_{x_t} = A_t \Gamma_{x_{t-1}} A_t^t + B_t \Gamma_{u_t} B_t^t + C_t \Gamma_{v_t} C_t^t \tag{13}$$

Thus, using the above expression recursively, we get

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

where B_t and C_t are the products of the appropriate Jacobian and we have neglected error in the reestimated structure represented in the WC reference system, just as we did in the last subsection. Now, we assume a uniformity in which the difference among the terms under the summation is neglected. Thus,

$$\Gamma_{x_n} = n\left(B_t \Gamma_{u_t} B_t^t + C_t \Gamma_{v_t} C_t^t \right) \tag{14}$$

In other words, the error covariance in the structure is *proportional* to the number of frames. This implies that error is accumulated with the number of frames.

3.1.3 General Transitory Sequence

The general situation with a transitory sequence is shown in Fig. 2c, where a point can be visible in any number of frames (except the entire sequence). Detailed formulation for this general case is tedious and the resulting complex expression will not give us an insight. Because we are interested in the 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 $F = \{f_t \mid t = 0, 1, 2, \cdots, n\}$ as k subsequences $F_l = \{f_{pk+l} \mid p \geq 0 \text{ is an integer}\}$, $l = 0, 1, 2, \cdots, k-1$, so that in each F_l each point is visible by two frames and each F_l is then a simple transitory sequence. k is called visibility span. The entire sequence consists of k subsequences each is a simple transitory sequence and is 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 based on each F_l is proportional to the length n/k:

$$\Gamma_{x_n} = \frac{n}{k}\left(B_t \Gamma_{u_t} B_t^t + C_t \Gamma_{v_t} C_t^t \right) = O(n / k) \tag{15}$$

where the factor in the parentheses 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 nontransitory sequence we obtained when we derive (12), which says that the error covariance matrix is reduced by a factor of $1/k$:

$$\Gamma_{x_n} = \frac{n}{k^2}\left(B_t \Gamma_{u_t} B^t + C_t \Gamma_{v_t} C_t^t \right) = O\left(n / k^2\right) \tag{16}$$

which gives an error rate n/k^2 for error covariance matrix Γ_{x_n}. This is a very interesting rate. If the temporal sampling density is increased by a factor of two with the same scan trajectory, n and k are both doubled, and the error covariance matrix is reduced by a factor of 1/2! We can also see that when $k = 1$, the general rate n/k^2 becomes the rate for the simple transitory case and when $k = n$ it gives that for the nontransitory case.

The error rate n/k^2 in (16) implies that the later a part of scene enters the view, the larger the number n, and thus the larger the variance of the error in its position with respect to the WC reference system fixed at the first view.

In the above particular subsequence decomposition gives a reachable error rate. Of course, this decomposition is not necessarily what is done by an actual estimation algorithm. It is to make derivation of error rate more concise and simpler. This does not affect the asymptotic error rate n/k^2, because the best estimate is still derived by processing the entire set of structure observations as a single batch.

3.1.4 Global Pose Error

In a nontransitory case, the error covariance matrix is given in (3) which is almost independent of n.

Now, consider the transitory case. According to (3), the error variance of camera pose estimate is the sum of two terms, that from Γ_u and that from $\Gamma_{x_{n-l}}$ where $n - l$ is the past time frame that shares sufficient features with the current view at n. Therefore, we have

$$\Gamma_{x_n} = \frac{n-l}{l^2}\left(B_t \Gamma_{u_t} B^t + C_t \Gamma_{v_t} C_t^t \right) + \frac{\partial m}{\partial u}\Gamma_u \frac{\partial m^t}{\partial u} \tag{17}$$

Since l is on the same order of k, the asymptotic error rate is n/k^2. Denote the last term in the above equation by $O(1)$ indicating it is caused by a single view of u vector. Thus, the pose error with a transitory sequence has the same asymptotic error rate as that of the structure estimate:

$$\Gamma_{x_n} = O\left(n / k^2\right) + O(1) \tag{18}$$

3.2 Camera-Centered Representation

In the CC representation, object structure is represented in the camera reference system. In other words, every previous observation about object structure must be transformed into the camera reference system at the current frame and be fused according to the Gauss-Markov Theorem.

An important difference between the WC and CC representations is the following. In the WC representation, every part of the scene that has been observed but is not currently visible does not need to be updated with the current view, because the WC reference system does not change with respect to the scene. In the CC representation, every part of the scene that has been observed but is not currently visible must be transformed to the current camera centered system because the CC reference system moves with respect to the scene.

With the CC representation, the pose m to be computed is from the past time $t - p$ to the current time t, $p = 1, 2, 3, \cdots, t - 1$. After fusing all the past views with the current view at t, the resulting structure is called the CC structure. Theoretically, the structure error should be the same as that with the WC representation if all the past frames are treated in a batch fashion. Thus the behavior of the error covariance matrix for the CC structure is the same as that of the WC structure estimated with the WC representation, except that time t is now reversed: the older the frame, the worse the structure accuracy in the CC representation.

However, the local structure, i.e., that is visible in the current frame, does not have the above transitory problem, simply because it is visible at current time n. Therefore, it can take the advantage of the situation enjoyed by the ideal nontransitory sequence. If the CC structure only takes past b frames into account as a batch, and those b frames share a considerable number of features with the current view at n. Then, according to the result (12) derived with uniform ideal non-transitory sequence, the CC structure of the currently visible part is of order $\Gamma_{x_n} = \frac{1}{b}\Gamma_{y_t}$ where Γ_{y_t} is the error covariance matrix of the past structure transformed to frame n, and b is the batch size. For the above expression to hold true, b should be small enough so that the past b frames share the structure x_n with the view at time n.

Now, we are ready to summarize the asymptotic error rates using Table 1. In Table 1, n is current time (or frame number), k the visibility span, and b is the batch size $b \le k$. All the structure error is that for the visible part at the current nth frame. The camera pose error in CC representation should be zero in all the cases, because it is defined directly in the camera system itself instead of being measured. In the table, "0" is used to indicate this fact.

TABLE 1
ASYMTOTIC RATE FOR ERROR COVARIANCE MATRIX IN INTEGRATION

Representation	Estimate	Nontransitory	Simple transitory	General transitory
WC	structure	$O(1/n)$	$O(n)$	$O(n/k^2)$
WC	pose	$O(1)$	$O(n)$	$O(n/k^2)+O(1)$
CC	structure	$O(1/n)$	$O(1)$	$O(1/b)$
CC	pose	0	0	0

As can be seen from Table 1, with a general transitory sequence, for global structure representation which is necessary for extended scene reconstruction, one should increase the visibility span k as much as possible. For the camera-centered local structure which is useful in grasping or collision avoidance, one should increase the batch processing size $b \le k$ for the best possible accuracy.

3.3 The Tightness of the Error Rates

The error rates we obtained in Table 1 are achievable How tight are those rates? Are those rates the best one can possibly achieve?

In general, the observation model of our problem can be expressed as

$$\hat{u} = u(\alpha) + \delta_u \tag{19}$$

where $\hat{u}$ is a vector of image-plane observations, contaminated by noise vector δ_u, and $u(\alpha)$ is the noise-free image plane vector which depends on the parameter vector α. In our problem, u consists of image coordinates of all the features in all the image frames. δ_u is the error vector which takes into account a wide variety of errors. The vector α is the parameter vector one wants to estimate, such as structure of the currently visible scene, camera pose, motion parameters, etc.

Suppose that $\hat{\alpha}$ is an unbiased estimator of α from $\hat{u}$ in (19), the noise vector δ_u has a zero mean and covariance matrix Γ_u, and the probability distribution density of the noise factor is $p(u, \alpha)$. In reality our estimator is not exactly unbiased and the noise mean does not have to be exactly zero. We assume that the absolute bias and the noise mean are negligibly small. The multidimensional version of the Cramér-Rao error bound [7], [12] gives

$$E(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^t \ge F^{-1} = \mathrm{CRB}(\alpha) \tag{20}$$

where E denotes expectation operator, and F is the Fisher information matrix:

$$F = \left[\frac{\partial \ln p(u,\alpha)}{\partial \alpha}\right]^t \left[\frac{\partial \ln p(u,\alpha)}{\partial \alpha}\right] \tag{21}$$

The inequality in (20) means that the difference matrix of the two sides is nonnegative definite. In particular, the diagonal elements and the trace of a nonnegative definite matrix are all nonnegative. Therefore Cramér-Rao bound gives a lower error bound for the error covariance of every component of the parameter vector α. As indicated in (21), such a bound is evaluated with noise-free observation u and true parameter vector α. It is worth noting that the bound is *algorithm independent*. It indicates that no matter what algorithm is used to estimate α, the resulting error covariance matrix of α cannot be lower than that specified by the bound.

Next, we investigate the Cramér-Rao bound of the global pose of the camera system in WC representation. We consider a general transitory sequence of length n, $F = \{f_t \mid t = 0, 1, 2, \cdots, n-1\}$ with a visibility span k. Since we are investigating asymptotic behavior in which n goes to infin-

ity, without loss of generality, we consider n to be an integral multiple of k, i.e., $n = (j + 1)k$, for some positive integer j. $j + 1$ is the length of k subsequences $F_l = \{f_{pk+l} \mid p = 0, 1, \cdots, j\}$, $l = 0, 1, 2, \cdots, k - 1$, each of them is a simple transitory sequence.

3.3.1 The Simple Transitory Case

Consider the subsequence F_0, of length j. As explained in (1), the global position of the camera at the ith frame of F_0, with respect to its global position at 0th frame F_0, can be specified by a column vector

$$m(i) = (p_x(i), p_y(i), p_z(i), \theta_x(i), \theta_y(i), \theta_z(i))^t$$

where $p(i) = (p_x(i), p_y(i), p_z(i))^t$ and $\theta(i) = (\theta_x(i), \theta_y(i), \theta_z(i))^t$ specify the global position and orientation, respectively. Define incremental interframe displacement

$$d(i) = m(i) - m(i - 1) \tag{22}$$

$i = 1, 2, \cdots, j$. From (22), we have the relation $m(i) = \sum_{t=1}^{i} d(t) + m(0)$, or

$$\begin{bmatrix} m(0) \\ m(1) \\ \vdots \\ m(j) \end{bmatrix} = \begin{bmatrix} I & 0 & \cdots & 0 \\ I & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ I & I & \cdots & I \end{bmatrix} \begin{bmatrix} m(0) \\ d(1) \\ \vdots \\ d(j) \end{bmatrix} \tag{23}$$

with I denoting the identity matrix. Alternatively, we write

$$m_{j,0} = M_j d_{j,0}$$

where we denote the left side of (23) by $m_{j,0}$ and the right side by the product of the matrix M_j and vector $d_{j,0}$. Geometrically, $m_{j,0}$ is the global attitude trajectory of the camera system while $d_{j,0}$ is the interframe displacement vector, plus the initial attitude $m(0)$. According to the definition of the Cramér-Rao bound, we have

$$\mathrm{CRB}\left(d_{j,0}\right) = \left\{ \left[\frac{\partial \ln p\left(u, d_{j,0}\right)}{\partial d_{j,0}} \right]^t \left[\frac{\partial \ln p\left(u, d_{j,0}\right)}{\partial d_{j,0}} \right] \right\}^{-1}$$

and

$$\mathrm{CRB}\left(m_{j,0}\right) = \left\{ \left[\frac{\partial \ln p\left(u, m_{j,0}\right)}{\partial m_{j,0}} \right]^t \left[\frac{\partial \ln p\left(u, m_{j,0}\right)}{\partial m_{j,0}} \right] \right\}^{-1}$$

Since

$$\frac{\partial \ln p\left(u, d_{j,0}\right)}{\partial d_{j,0}} = \frac{\partial \ln p\left(u, m_{j,0}\right)}{\partial m_{j,0}} \frac{\partial m_{j,0}}{\partial d_{j,0}} = \frac{\partial \ln p\left(u, m_{j,0}\right)}{\partial m_{j,0}} M_j$$

it follows that

$$\mathrm{CRB}\left(m_{j,0}\right) = M_j \mathrm{CRB}\left(d_{j,0}\right) M_j^t \tag{24}$$

For our purpose of investigating the asymptotic behavior of the Cramér-Rao bound, we need the uniformity condition of the motion sequence as we did earlier, since the behavior of an otherwise arbitrarily changing motion trajectory can depend more on a particular local motion instead of the temporal trend of the error behavior. Now, we assume a uniformity with which the differences among the interframe motions $d(i)$, $i = 1, 2, \cdots, j$ are neglected. In other words, the Cramér-Rao bound (CRB) of interframe motions $\mathrm{CRB}(d_{j,0})$, which is a symmetric matrix, is now a band matrix:

$$CRB\left(d_{j,0}\right) = \begin{bmatrix} C_0 & C_1 & C_2 & & 0 \\ C_1 & C_0 & C_1 & \ddots & \\ C_2 & C_1 & C_0 & \ddots & C_2 \\ & \ddots & \ddots & \ddots & C_1 \\ 0 & & C_2 & C_1 & C_0 \end{bmatrix} \tag{25}$$

In other words, denoting $\mathrm{CRB}(d_{j,0}) = [C_{pq}]$, then $C_{pq} = C_{qp} = 0$ whenever $| p - q | \geq h$, for some constant h. The unfirmity condition requires that the error bounds for estimating interframe motions d_i and d_j, respectively, are not correlated when the interframe motions are farther than h frames apart. This is a reasonable condition because an interframe motion depends mostly on the two image frames that defined the interframe motion. Although the information about the scene structure may contribute to the estimation of interframe motion to some degree, two far apart interframe motions do not share any common scene element when h is large enough in a general transitory sequence.

Without loss of generality, we can consider $h = 2$ for a simple transitory sequence. Thus, (24) and (25), give

$$\mathrm{CRB}\left(m_{j,0}\right) = \begin{bmatrix} I & 0 & \cdots & 0 \\ I & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ I & I & \cdots & I \end{bmatrix} \begin{bmatrix} C_0 & C_1 & & 0 \\ C_1 & C_0 & \ddots & \\ & \ddots & \ddots & C_1 \\ 0 & & C_1 & C_0 \end{bmatrix} \begin{bmatrix} I & 0 & \cdots & 0 \\ I & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ I & I & \cdots & I \end{bmatrix}^t \tag{26}$$

The last element in the bottom row of $\mathrm{CRB}(m_{j,0})$ is the CRB for the global pose of the camera at frame j of F_0, which gives

$$\mathrm{CRB}\left(m_{j,0}\right) = \begin{bmatrix} I\, I\, \cdots\, I \end{bmatrix} \begin{bmatrix} C_0 & C_1 & & 0 \\ C_1 & C_0 & \ddots & \\ & \ddots & \ddots & C_1 \\ 0 & & C_1 & C_0 \end{bmatrix} \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} = \tag{27}$$

$$(j + 1)C_0 + 2jC_1 = O(j)$$

In other words, we have proved the following theorem:

THEOREM 1. *Under the uniformity condition, the* CRB *of the global pose error at frame j in a simple transitory sequence is of the order* $O(j)$.

Similarly, the CRB of the global position of the structure has the same order in error rate as the global pose.

3.3.2 The General Transitory Case

First, we extend the above result for F_0 to the other subsequences F_l. We extend our notation from $m_{j,0}$ and $d_{j,0}$ to $m_{j,l}^{(l)}$ and $d_{j,l}^{(l)}$, respectively, to denote the corresponding trajectories of F_l, starting from frame f_l to frame f_{jk+l}, $l = 0, 1, \cdots,$

$k-1$. Given F, the above discussion still holds for F_l, except that the meaning of C_i in (25) is the CRB of the corresponding component based on the entire F instead of just F_0. Therefore, the CRB of the error rate of the global pose $m_{j,l}^{(l)}$ is still of order $O(j) = O(n/k)$:

$$\mathrm{CRB}\left(m_{j,l}^{(l)}\right) = O(n\ /\ k) \tag{28}$$

Consider each scene element x_n that is visible from f_{jp+l}, the last frame of F_l, $l = 0, 1, \cdots, k-1$. Since CRB is a lower bound and Table 1 means that $\mathrm{CRB}(x_n) \le O(n/k^2)$. To establish $\mathrm{CRB}(x_n) = O(n/k^2)$, all we need to prove is $\mathrm{CRB}(x_n) \ge O(n/k^2)$. To do the latter, we can neglect some errors without affecting the order. For $l = 1, 2, \cdots, k-1$, we neglect the interframe pose error between frame f_0 and f_l, and the error in the process of constructing x_n from frame f_{jp+l}. Thus, x_n is determined by the camera global position at f_{jp+l} by a function g:

$$x_n = g\left(m_{j,0}^{(0)}, m_{j,1}^{(1)}, \ \ldots, \ m_{j,k-1}^{(k-1)}\right) = g\left(m_j\right)$$

where we define $m_j = \left(m_{j,0}^{(0)}, m_{j,1}^{(1)}, \ \ldots, \ m_{j,k-1}^{(k-1)}\right)^t$. Since all subsequences F_l are independent with each other, the CRB of m_j is a block diagonal matrix:

$$\mathrm{CRB}\left(m_j\right) = diag\left\{\mathrm{CRB}\left(m_{j,0}^{(0)}\right), \mathrm{CRB}\left(m_{j,1}^{(1)}\right), \ \ldots, \ \mathrm{CRB}\left(m_{j,k-1}^{(k-1)}\right)\right\} \tag{29}$$

Using the variable change technique as we used before, we have

$$\begin{aligned}
&CRB\left(x_n\right) = \\
&\left\{\left[\frac{\partial \ln p\left(u, x_n, m_j\right)}{\partial x_n}\right]^t \left[\frac{\partial \ln p\left(u, x_n, m_j\right)}{\partial x_n}\right]\right\}^{-1} \\
&\ge \left\{\left[\frac{\partial m_j}{\partial x_n}\right]^t \left[\frac{\partial \ln p\left(u, x_n, m_j\right)}{\partial m_j}\right]^t \left[\frac{\partial \ln p\left(u, x_n, m_j\right)}{\partial m_j}\right]\left[\frac{\partial m_j}{\partial x_n}\right]\right\}^{-1} \\
&= \left\{\left[\frac{\partial m_j}{\partial x_n}\right]^t \mathrm{CRB}\left(m_j\right)^{-1}\left[\frac{\partial m_j}{\partial x_n}\right]\right\}^{-1}
\end{aligned} \tag{30}$$

Since $\mathrm{CRB}(m_j)$ is block diagonal, then so is its inverse. The above inequality gives

$$\mathrm{CRB}\left(x_n\right) \ge \left\{\sum_{l=0}^{k-1}\left[\frac{\partial m_{j,l}^{(l)}}{x_n}\right]^t \mathrm{CRB}\left(m_{j,l}^{(l)}\right)^{-1}\left[\frac{\partial m_{j,l}^{(l)}}{x_n}\right]\right\}^{-1}$$

Under the uniformity condition, $\dfrac{\partial m_{j,l}^{(l)}}{x_n}$ and $\mathrm{CRB}\left(m_{j,l}^{(l)}\right)$ are treated as constant with respect to l. Thus

$$\begin{aligned}
\mathrm{CRB}\left(x_n\right) &\ge \left\{k\left[\frac{\partial m_{j,0}^{(0)}}{x_n}\right]^t \mathrm{CRB}\left(m_{j,0}^{(0)}\right)^{-1}\left[\frac{\partial m_{j,0}^{(0)}}{x_n}\right]\right\}^{-1} \\
&= \left[\left[\frac{\partial m_{j,0}^{(0)}}{x_n}\right]^t\right]^{-1} \frac{1}{k}\,\mathrm{CRB}\left(m_{j,0}^{(0)}\right)\left[\frac{\partial m_{j,0}^{(0)}}{x_n}\right]^{-1} \\
&= O\left(n\ /\ k^2\right)
\end{aligned} \tag{31}$$

the last equation used the result in (28). Therefore, we have $\mathrm{CRB}(x_n) = O(n/k^2)$.

The CRB for global pose can be directly derived from that of the global position of the structure. The derivation for the order of CRB in nontransitory case is simple and is omitted.

In summary, we have established the following result:

THEOREM 2. *The asymptotic error rates in Table* 1 *are not only reachable but also the theoretical lowerest possible specified by the Cramér-Rao lower bound. This is true for any distribution as long as the uniformity condition is satisfied.*

These error rates are determined by the nature of the transitory sequence. Although we have used the uniformity condition so that the rate can be expressed simply, the uniformity condition can be applied to *ensemble average* in terms of random process. Passing without a rigorous proof, the rates stated in Theorems 1 and 2 are probably true for general random motion sequences in the sense that they are average rates as long as the uniformity is true on average.

4 METHODS AND ALGORITHMS

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.

We first consider estimation with a nonlinear observation function f. Suppose that an observation vector y is related to a parameter vector m by a nonlinear equation $y = f(m) + \delta_y$ where δ_y is a pairwise uncorrelated random noise vector with zero mean, and covariance matrix $\Gamma_y = E\delta_y\delta_y^t$. The maximum likelihood estimate with Gaussian noise δ_y or minimum variance linear estimate with a general noise distribution calls for minimizing

$$\left(y - f(m)\right)^t \Gamma_y^{-1}\left(y - f(m)\right) \tag{32}$$

with respect to m. In other words, the optimal parameter vector m is the one that minimizes the matrix-weighted discrepancy between the computed observation $f(m)$ and the actual observation y. At the solution that minimizes (32), the estimated $\hat{m}$ has a covariance matrix

$$\Gamma_{\hat{m}} = E\left(\hat{m} - m\right)\left(\hat{m} - m\right)^t \simeq \left\{\frac{\partial f\left(\hat{m}\right)^t}{m}\Gamma_y^{-1}\frac{\partial f\left(\hat{m}\right)}{m}\right\}^{-1} \tag{33}$$

One of the advantages of this minimum variance criterion is that we do not need to know the exact noise distribution.

4.1 Cross-Frame Approach With CC Representation

Let X_p denote the 3D 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.

All the structure observed in the past needs to be transformed to the CC system at frame p and properly fused. There are two basic approaches in the fusion of the past structure.

1) Recursive method: frame by frame. The fused structure at previous frame is transformed to the current frame p and fused with the new observation at p according to the estimated interframe motion $m_{p,p-1}$.
2) Batch method: cross-frame. For each $q \in \{p-1, p-2, \cdots, p-b+1\}$, estimate the cross-frame motion $m_{p,q}$ and transform X_q to frame p and fuse with the new observation at p.

The first method involves two frames at a time, $p-1$ and p. As we discussed before, the fused structure has an error covariance matrix of $\left(\left(\Gamma_m + \Gamma_{x_{p-1}}\right)^{-1} + \Gamma_p^{-1}\right)^{-1}$ where Γ_p is the error covariance matrix of single observation at p and $\Gamma_{x_{p-1}}$ is due to the error in interframe motion estimate. A structure estimate at from $p-l$ will undergo l such deteriorations under the frame-by-frame recursive method and thus, when $l > 1$, the old structure estimate is hardly useful in the fusion with that in view p.

Under the second cross-frame method, each previous structure estimate at $p-l$ is directly transformed to p under one transformation. Thus the transformed structure deteriorates by the motion error only once. Fig. 3 graphically explains the advantage of using cross-frame motions.

In practice, we define a number K, called extra batch size, to be the number of extra image (stereo) frames that are processed as a batch in additional to the last two. Thus, at current frame number p, the image frame batch consists of frames from $p-K-1$ to p. According to our discussion about non-transitory and transitory image sequences, it is useful for K to span a subsequence that is nearly nontransitory. With a batch at frame p, the current active cross-frame motion set is denoted by

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

The cross-frame motion set completely defines the motion between any two frames within the batch. When $K = 0$, we have just an interframe motion in $W(p)$.

Let N be the total number of feature points being considered; $x_{i,s}$ denote the 3D local structure of ith point in sth camera-centered system; $u_{i,j,s}$ be the 2D image coordinate vector of ith point on the jth side (left, right) at the sth frame. Assuming that the noise in the observations ($u_{i,j,s}$) is uncorrelated and has the same variance $\left(\sigma_u^2, \sigma_v^2\right)$ in the two image coordinates, expression (32) that is to be minimized can be written the following form

$$\min_{\forall x_{i,p}, \forall m \in W(p)} f\left(m, x_{i,p}\right) = A + B \tag{34}$$

where

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

with

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

and

$$B = \sum_{s=p-K}^{p} \sum_{j=L}^{R} \sum_{i=1}^{N} \left(\hat{u}_{i,j,s} - u\left(m_{s,p}, x_{i,p}\right)\right)^t \begin{bmatrix} \sigma_u^{-2} & 0 \\ 0 & \sigma_v^{-2} \end{bmatrix} \left(\hat{u}_{i,j,s} - u\left(m_{s,p}, x_{i,p}\right)\right)$$

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. This objective function has two terms. The first term, A, reflects the integrated 3D structure in the past up to time $p-K-1$. The second term, B, is used to minimize the image plane error of the frames within the batch from from $p-K$ up to p. The summation bound for i can be modified to include only those points that are visible in each frame so that a point does not have to be visible through the entire batch.

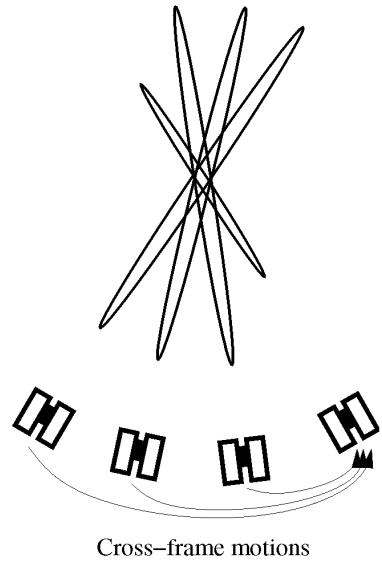


Fig. 3. Using cross-frame motions to integrate many views. Each elongated ellipse indicates the uncertainty in 3D point position transformed from a single previous stereo view to the current view. The integrated uncertainty is greatly reduced using the multiple cross-frame motions instead of interframe motions.

4.1.1 Minimization of the Objective Function

The objective function in (34) is neither linear nor quadratic in terms of cross-frame motion parameters, m, and 3D feature points, x. An iterative algorithm is required to search for the solution of m and x. The dimension of the unknown parameters is intractably huge due to a typically large N. Thus, a direct optimization is impractical. Our two procedures play a central role in resolving this problem:

First, a (suboptimal) closed-form solution for interframe motion from $p-1$ to p is first computed. This interframe motion is used together with previous pose estimate to compute a preliminary estimate for all the cross-frame motions needed.

The second is to eliminate iteration on the structure. The gradient-based search is only applied to cross-frame motions, because given each candidate set of cross-frame motions the best structure for (34) can be directly computed in a closed-form. To show how, let us examine the objective function (34). The second term of the objective function corresponds to minimizing the image vector error within the batch. An alternative way to approximate this is to use the matrix-weighted discrepancy of $x_{i,p} - X(m_{p,j}, x_{i,j})$, the 3D position difference, to give the total discrepancy

$$\min_{x_{i,p}} \sum_{s=p-K}^{p} \sum_{i=1}^{N} \left(x_{i,p} - X\left(m_{p,s}, x^*_{i,s}\right)\right)^t \left(R_{p,s}\Gamma_{x^*_{i,s}} R^t_{p,s}\right)^{-1} \left(x_{i,p} - X\left(m_{p,s}, x^*_{i,s}\right)\right) \tag{35}$$

where $x^*_{i,s}$ is computed from the triangulation at frame s, $\Gamma_{x^*_{i,s}}$ is the estimated covariance matrix of $x^*_{i,s}$ for triangulation. Substituting the second term of objective function (34) with (35), we minimize

$$\min_{x_{i,p}} f\left(x_{i,p}\right) = \sum_{s=p-K-1}^{p} \sum_{i=1}^{N} \left(x_{i,p} - X\left(m_{p,s}, x^*_{i,s}\right)\right)^t \left(R_{p,s}\Gamma_{x^*_{i,s}} R^t_{p,s}\right)^{-1} \left(x_{i,p} - X\left(m_{p,s}, x^*_{i,s}\right)\right) \tag{36}$$

given any $W(p)$. The above is a linear minimization problem, for which we just need to solve the following linear equation [3],

$$\left\{\sum_{s=p-K-1}^{p} \left[R_{p,s}\Gamma^{-1}_{x_{i,s}} R^t_{p,s}\right]\right\} x_{i,p} = \sum_{s=p-K-1}^{p} \left\{\left(R_{p,s}\Gamma^{-1}_{x_{i,s}} R^t_{p,s}\right) X\left(m_{p,s}, x_{i,s}\right)\right\} \tag{37}$$

which gives

$$x_{i,p} = \left\{\sum_{s=p-K-1}^{p} \left[R_{p,s}\Gamma^{-1}_{x_{i,s}} R^t_{p,s}\right]\right\}^{-1} \left\{\sum_{s=p-K-1}^{p} \left\{\left(R_{p,s}\Gamma^{-1}_{x_{i,s}} R^t_{p,s}\right) X\left(m_{p,s}, x_{i,s}\right)\right\}\right\}$$

Its error covariance matrix is estimated by [3]

$$\Gamma_{x_{i,p}} = \left[\sum_{s=p-K-1}^{p} \Gamma^{-1}_{i,s}\right]^{-1}$$

where

$$\Gamma_{i,s} = \left(\frac{\partial X\left(m_{p,s}, x^*_{i,s}\right)}{\partial m}\right) \Gamma_{m_{p,s}} \left(\frac{\partial X\left(m_{p,s}, x^*_{i,s}\right)}{\partial m}\right)^t + \left(\frac{\partial X\left(m_{p,s}, x^*_{i,s}\right)}{\partial x^*_{i,s}}\right) \Gamma_{x^*_{i,s}} \left(\frac{\partial X\left(m_{p,s}, x^*_{i,s}\right)}{\partial x^*_{i,s}}\right)^t$$

4.2 World-Centered Representation

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.

4.2.1 Objective Function

Without loss of generality, let the world coordinate system coincide with the camera-centered coordinate system of the first frame.

$$M(m, n) = \bigcup_{i=m}^{n} \left\{m_{i,1}\right\} = \bigcup_{i=m}^{n} \left\{R_{i,1}, T_{i,1}\right\}$$

is the collection of all the global motions, where $\left(R_{i,1}, T_{i,1}\right)$ is the rotation matrix and translation vector from frame 1 to i. For each feature points i, we have structure G_i corresponding to the world coordinate system. Now slightly modifying the equation (34), we get the appropriate objective function for the WC representation:

$$\min_{\forall G_i, \forall m \in M(p-K-1,p)} f(G_i, m) = A + B \tag{38}$$

where

$$A = \sum_{i=1}^{N} \left(G_i - G^*_i\right)^t \Gamma^{-1}_{G^*_i} \left(G_i - G^*_i\right) \tag{39}$$

and

$$B = \sum_{s=p-K}^{p} \sum_{j=L}^{R} \sum_{i=1}^{N} \left(\hat{u}_{i,j,s} - u\left(m_{s,1}, G_i\right)\right)^t \begin{bmatrix} \sigma_u^{-2} & 0 \\ 0 & \sigma_v^{-2} \end{bmatrix} \left(\hat{u}_{i,j,s} - u\left(m_{s,1}, G_i\right)\right) \tag{40}$$

In the objective function, $u(m_{s,1}, G_i)$ is the noise-free projection computed from $m_{s,1}$ and G_i. The essence of the above objective function is that newly updated global structure G_i

takes into account the old observation G_i^* integrated up to frame $p - K - 1$, but it considers all the observations in the batch as image vectors, all properly weighted in the sense of Gauss-Markov.

Similar to computation for the CC representation, no iteration is needed for the structure part, and a suboptimal closed-form solution is computed first for motion and structure which is used as the initial guess for minimization. The following equation gives the closed form solution for structure parameters G_i when the motion parameters $M(m, n)$ are given:

$$G_i = \left(\sum_{s=p-K-1}^{p} \Gamma_{G_{i,s}}^{-1} \right)^{-1} \left(\sum_{s=p-K-1}^{p} \Gamma_{G_{i,s}}^{-1} G_{i,s} \right) \tag{41}$$

where $G_{i,s}$ is the estimation based on the single frame s. The estimated error covariance matrix of the newly updated the structure is $\Gamma_{G_i} = \left(\sum_{s=p-K-1}^{p} \Gamma_{G_{i,s}}^{-1} \right)^{-1}$. This WC based objective function is in essence similar to those of [4] and [6]. The differences are:

1) a batch parameter K is used to better deal with the transitory sequence;
2) the image-plane discrepancy is minimized to automatically take into account nonsymmetrical nature of error distribution in 3D point positions; and
3) the algorithm can automatically handle leaving points and coming points which is required with transitory sequences.

5 Experiments

We conducted experiments with synthetic and real word images in order to experimentally examine the error rates listed in Fig. 1 and compare the WC and CC representations.

5.1 Simulation

The 3D feature points were generated randomly for each trial, between depth 2,000 mm and depth 3,000 mm, with a uniform distribution. The entire scene is covered by 31 frames and the distance between consecutive frames is roughly 200 mm. A small rotation is added between each pair of two consecutive frames. Fig. 4 illustrates the simulation environment. This environment is similar to the real setup to show later. The average errors we will show were obtained through 100 random trials each with a different set of 3D points. With our setup, in order to let the first and the last frame in the batch share at least 30 percent of the scene, the batch size should not be larger than three.

5.1.1 Results

Fig. 5 shows the current camera position error ($R_{i,1}$, $T_{i,1}$) for different frames. It can be seen from the figure that the batch size 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. The structure error is shown as the average error of all the visible

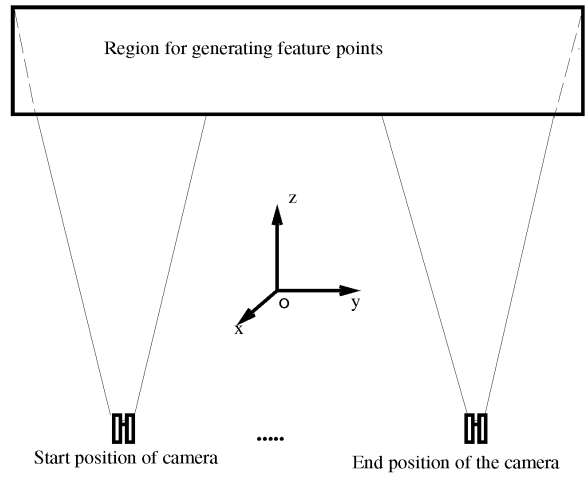


Fig. 4. Simulation environment, where 7,000 mm distance is covered by the 31 frames.

feature points at the current frame. The result indicates that a larger batch size is very effective to reduce both the local and global structure errors, for CC representation, as we predicted in Section 4.1. A larger batch size does not improve much for WC representation due to dominantly linear nature of the WC structure fusion. The figure also shows that the WC representation performs better for estimating the global structure while the CC representation does better for local structure, as predicted by Table 1. A point worth noting here is the fact the local structure error with the CC representation is constant, while that with the WC representation grows with time, also a property predicted by Table 1.

5.2 Experiments With a Real Setup

A challenging task facing the area of motion and structure analysis is to provide data from rigorous experiments that verified the actual accuracy of the results with an automatic algorithm, so that we can evaluate whether passive structure sensing is possible and reliable in real world. The result reported here is an effort toward this goal.

The setup used for our image acquisition is a Denning MRV-3 mobile robot and a pair of stereo cameras, 265 mm apart, mounted on a custom-designed stereo positional setup that allows step-motor controlled pan and tile for each camera from a computer, as shown in Fig. 6. The stereo camera system was calibrated with distortion compensation using an algorithm from Weng et al. [11]. The field of view of each camera is about 36 degrees diagonally, and each digitized image has 512×480 pixels. An image sequence of 151 frames was acquired from the moving mobile robot. It contains a left-view sequence and a right-view sequence. The entire stereo sequence was used for automatic feature extraction, matching and tracking. A temporally subsampled (one sample every five frames) subsequence of 31 frames was used for motion and structure estimation with a consideration that this subsequence is dense enough for estimation and yet enables cross-frame motions to cover more original frames with a relatively small batch size. Fig. 6 shows a few images in the 151-frame sequence.

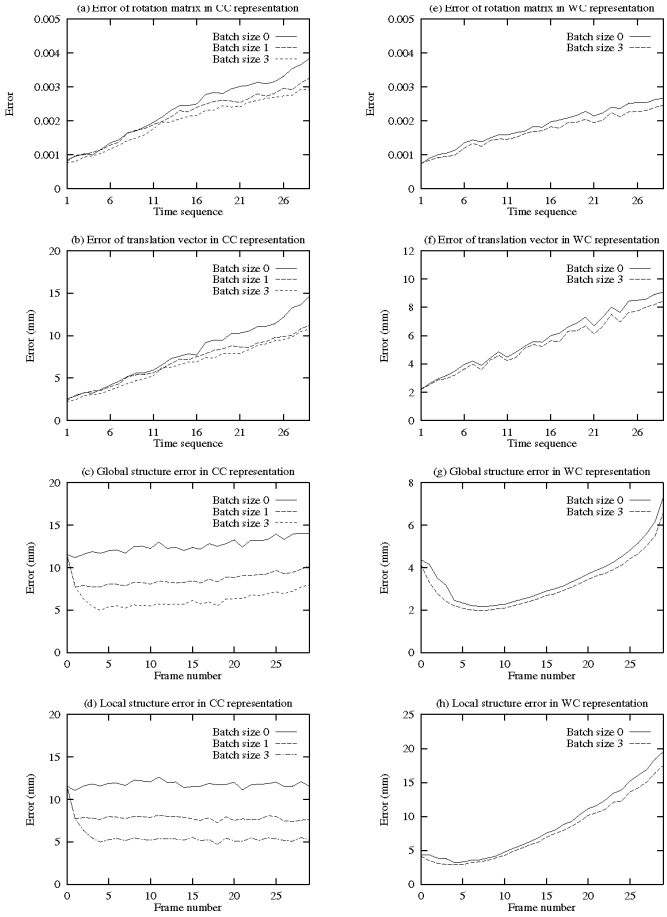


Fig. 5. The error versus time for the simulation result. The CC representation on the left column and WC on the right.

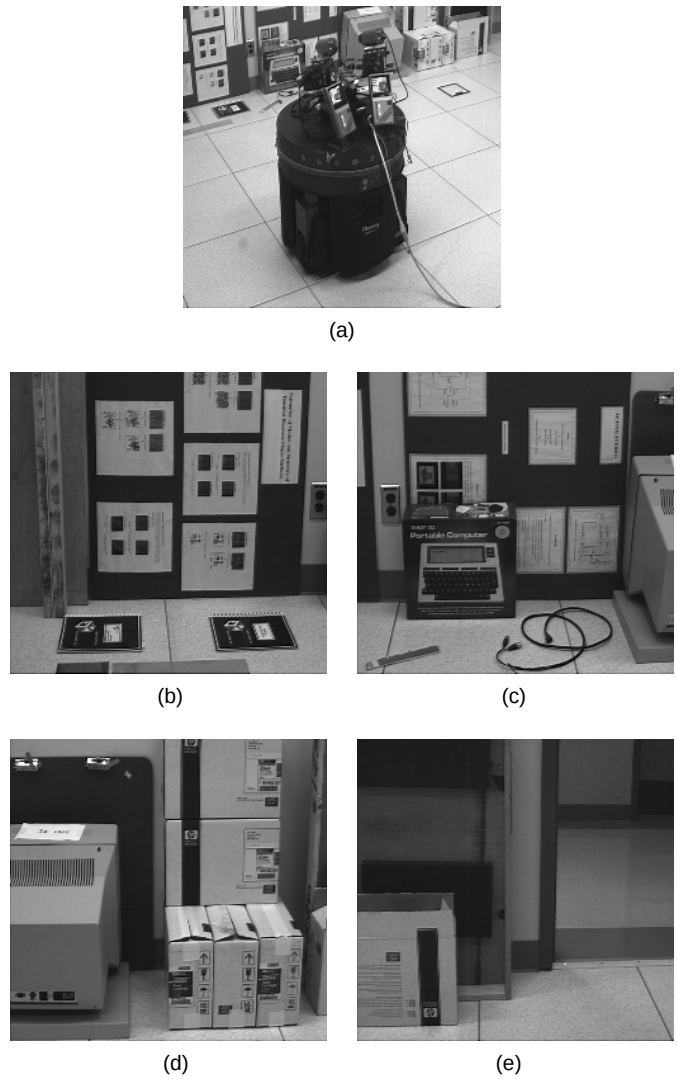


Fig. 6. The robot and a few stereo frames in the 151-frame sequence. (a) Robot. (b) Left image of frame 0. (c) Left image of frame 50. (d) Left image of frame 100. (e) Left image of frame 150.

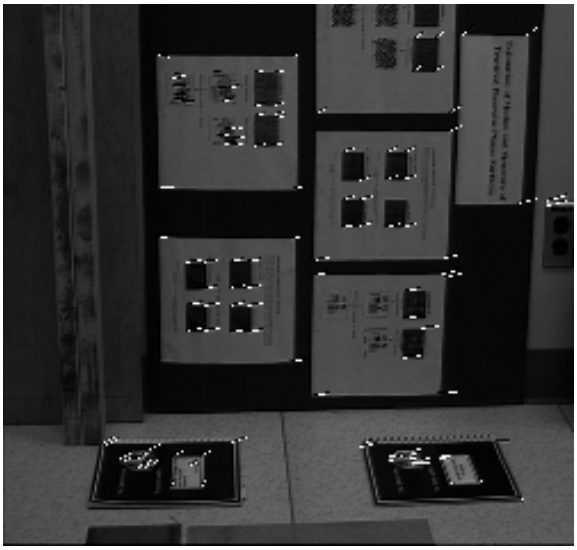
A feature point detector has been developed for this project to automatically detect feature points from images. The feature detector first computes the cornerness measure (the degree a point looks like at a corner) at every pixel. Then the local peaks of this cornerness measure are detected to form a peak histogram, ranked with the cornerness measure. The program automatically determines the threshold so that the required number of features are given from top rankings. An area-directed analysis is incorporated into the scheme so that the detected feature points evenly spread across the entire image.

Stereo matching was done using the image matching algorithm from Weng et al. [10], which provides a dense displacement field with a disparity vector for every pixel. The disparity vector at every feature point is extracted from this field.

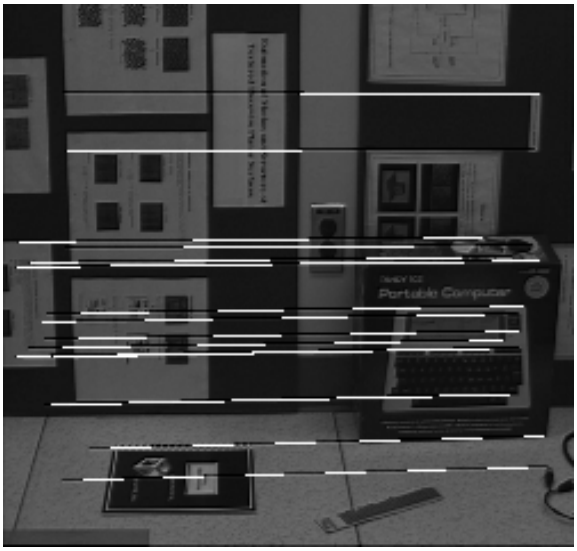
For efficiency, the algorithm uses a corner tracking mechanism as much as possible. Only when the tracking is not successful based on the closeness measure used by tracking, is the matcher called. The trace record of the entire sequence is shown in Fig. 7. About 100 feature points were automatically kept at any time. Since some points may go out of view and some points may become inactive, the number of active points may fall below a tolerable number (90 in our experiment). If this happens, the feature detector is called which provides additional points from the image and then the stereo matcher is called to give stereo matching for these new points. The time when the feature detector was automatically called can be clearly identified in Fig. 7. Fig. 8 presents an example of temporal matching-and-tracking. A careful visual inspection of entire point trace indicates that there was no visible errors.

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 1 mm. The selection of test points were based on ease of measurement and was not based on automatically selected features. Thus, each test point is not necessarily a part of the feature points used for the automatic algorithms, although many of them are. The image coordinates of the test points were manually measured from digital images. The accuracy of the reconstructed structure error was measured by the following steps:

1) Compute the WC and CC representations for feature point position and camera pose using the fully automatic algorithm described above.
2) Manually measure the image coordinates of the test points.



(a)



(b)

Fig. 8. Stereo matching and temporal matching-and-tracking. (a) An example of stereo matching (frame 0). (b) An example of temporal matching and tracking (frame 24 to 69). A needle is drawn from the feature point to its position in the target frame. Due to camera vergence, the orientation of the needles in (a) is correct.

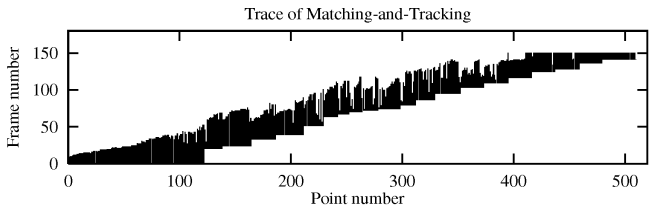


Fig. 7. The tracking record of the feature points through the 151 frames in the sequence. If a point *k* is successfully tracked from frame *i* to frame *j*, a vertical line is shown at point number *k* from frame *i* to frame *j*. (Due to the limit of the printer resolution, lines are merged in the plot.)

3) Perform a multiframe triangulation to get the 3D position of the test points. The number of frame used is according to the corresponding batch size.
4) Measure the global position error as the difference between the true and estimated test points.

This way of measuring error tests not only the pose of the camera, but also some of the reconstructed feature points, if they are also test points. Table 2 lists some data of the real setup. Fig. 9 shows the test points of one frame.

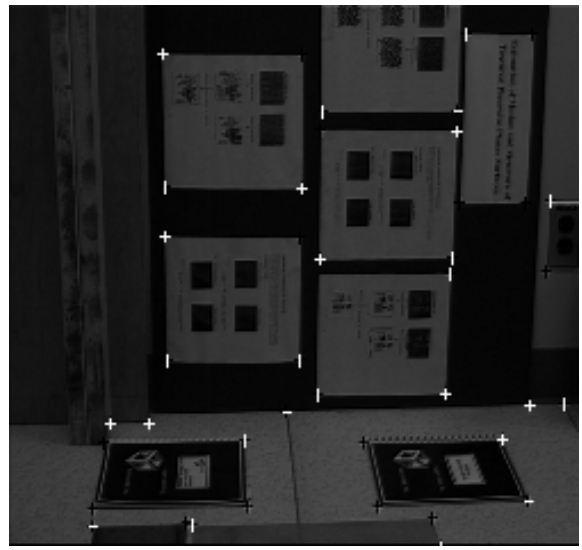


Fig. 9. Sample test points on one frame. Each cross shows the location of a test point.

TABLE 2
SOME DATA FOR THE REAL SETUP

Number of frame	31	Distance traveled (mm)	3,097
Number of feature points	387	Number of test points	85

TABLE 3
AVERAGE IMAGE PLANE RESIDUAL

Representation	batch size 0	batch size 3
WC	0.76 pixel	0.68 pixel
CC	0.45 pixel	0.51 pixel

First, to show how well the estimated structure and camera pose agree with the automatically detected feature points, the estimated 3D feature points were projected onto the image plane according to the pose. The average distance difference between every projected point and actually detected feature point is called image plane residual and is shown in Table 3. The values are around a fraction of a pixel for both representations. These numbers also indicate that camera distortion compensation in the calibration was very effective.

Fig. 10 shows the actual camera orientation error. Although the image sequence here is transitory, the pitch and row errors are comparable with those in the nontransitory "Hotel" sequence experiment by Tomasi and Kanade [9] over the entire sequence. The visibility span of our setup is about four. At frame four, the yaw error has the same magnitude as that in [9]. After frame four, the error tends to increase due to the transitory nature of the sequence. It is interesting that roll and pitch errors did not increase

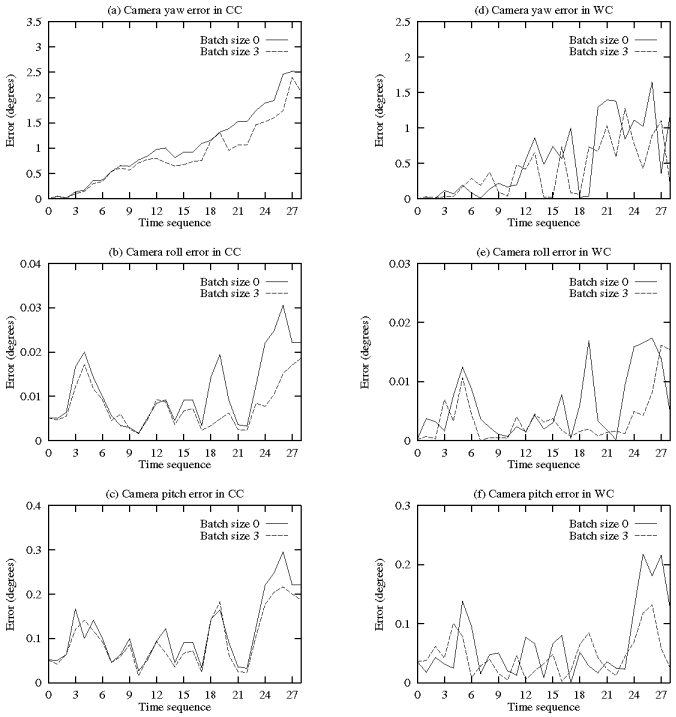


Fig. 10. Camera rotation error versus time.

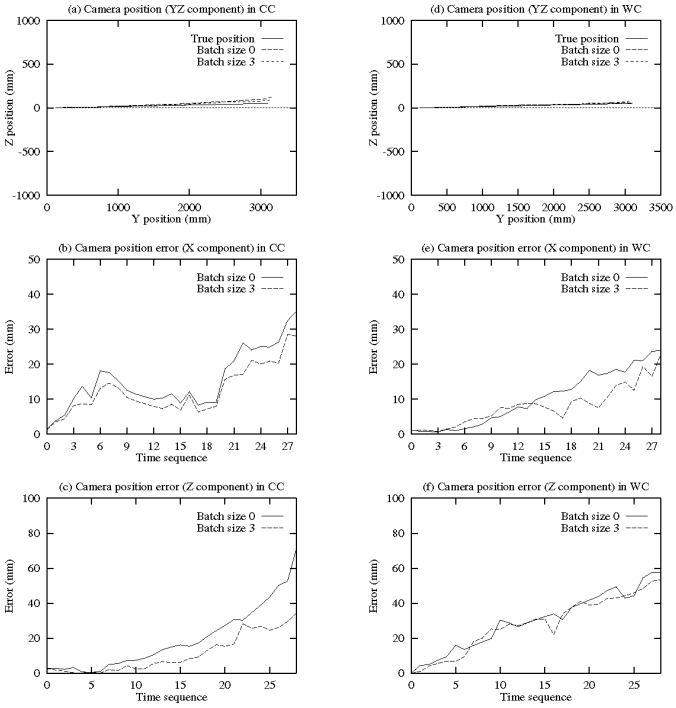


Fig. 11. Camera position error versus time.

quickly with time. After traveling about 3,000 mm, the total orientation error is not more than 0.02° in roll, 0.23° in pitch, and 2° in yaw with the WC representation.

Fig. 11 shows the camera position error and Fig. 12 presents the global error of the test points visible at the current time. As we predicted, the error increases with the time. But the estimates appear good. After traveling about 3,000 mm, the estimated camera global position error is less than 60 mm in depth Z (less than 2.3 percent), about 43 mm horizontally and under 25 mm vertically with the WC representation. This

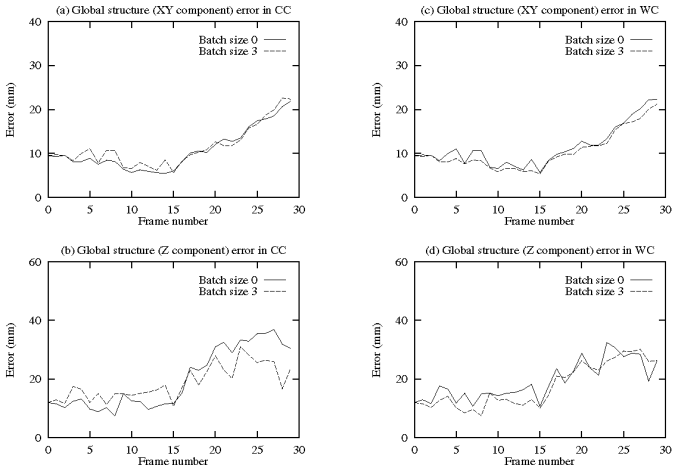


Fig. 12. Structure error versus time.

Fig. 13. Reconstructed 3D surface integrated from many partial views in the sequence, shown with original intensity viewed from an arbitrary direction.

seems to indicate that reasonable results can be obtained with a fully automatic algorithm, even with a transitory image sequence. Fig. 13 shows the reconstructed 3D surface.

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 asymptotic error rates that are very different from those with a nontransitory one. The theoretical error rates listed in Table 1 indicates that the WC representation is better for global estimates and the CC representation is superior for local estimates.

The verified accuracy in our experiment appears to indicate that with off-the-shelf cameras, one can automatically determine the scene structure and pose of the camera with a good accuracy, although the image sequence here is of a more difficult transitory type (compared with nontransitory ones).

ACKNOWLEDGMENTS

The authors would like to thank Ajit Singh for illuminating discussions and Li-an Tang for improving a data lot. The work was supported in part by a research initiation grant from Michigan State University.

REFERENCES

[1] T.J. Broida and R. Chellappa, "Estimation of Object Motion Parameters From Noisy Images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 90-99, 1986.

[2] N. Cui, J. Weng, and P. Cohen, "Extended Structure and Motion Analysis From Monocular Image Sequences," *CVGIP: Image Unserstanding*, vol. 59, no. 2, pp. 154-170, Mar. 1994.

[3] T.F. Elbert, *Estimation and Control of Systems.* New York: Van Nostrand Reinhold, 1984.

[4] N. Ayache and O. Faugeras, "Building, Registration, and Fusing Noisy Visual Maps," *Proc. First Int'l Conf. Computer Vision*, England, pp. 73-82, 1987.

[5] D.G. Luenberger, *Optimization by Vector Space Methods.* New York: John Wiley, 1969.

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

[7] C.R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley, 1973.

[8] H. Shariat and K.E. Price, "Motion Estimation With More Than Two Frames," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 5, pp. 417-434, 1990.

[9] C. Tomasi and T. Kanade, "Shape and Motion From Image Streams Under Orthography: A Factorization Method," *Int'l J. Computer Vision*, vol. 9, no. 2, pp. 137-154, 1992.

[10] J. Weng, N. Ahuja, and T.S. Huang, "Matching Two Perspective Views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 806-825, Aug. 1992.

[11] J. Weng, P. Cohen, and M. Herniou, "Camera Calibration With Distortion Models and Accuracy Evaluations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 10, pp. 965-980

[12] J. Weng, N. Ahuja, and T.S. Huang, "Optimal Motion and Structure Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 864-884, Sept. 1993.

John (Juyang) Weng (S'85-M'88) received the BS degree from Fudan University, Shanghai, China, in 1982 and the MS and PhD degrees from University of Illinois, Urbana-Champaign, USA, in 1985 and 1989, respectively, all in computer science. From September 1984 to December 1988, he was a research assistant at the Coordinated Science Laboratory, University of Illinois, Urbana-Champaign. In the summer of 1987 he was employed at the IBM Los Angeles Scientific Center. Since January 1989, he has been a researcher at Centre de Recherche Informatique de Montreal, Montreal, Quebec, Canada, while adjunctively with Ecole Polytechnique de Montreal. From October 1990 to August 1992, he held a visiting assistant professor position at University of Illinois, Urbana-Champaign. Currently, he is an assistant professor at the Department of Computer Science, Michigan State University, East Lansing. He is a coauthor of the book *Motion and Structure from Image Sequences* (Springer-Verlag, 1993) and is on the editorial board of *IEEE Transactions on Image Processing.* His current research interests include computer vision, learning for humans and machines; human-machine interface using vision, speech, gesture, and action; multimedia understanding by machines, and autonomous learning robots.

Yuntao Cui received his BS degree from Shanghai Jiao Tong University in 1988, MS degree from Southeastern Massachusetts University in 1991, and PhD from Michigan State University in 1996, all in computer science. He is currently with Siemens Corporate Research, Inc., Princeton, N.J. His research interests include pattern recognition, computer vision, and machine learning techniques.

Narendra Ahuja (S'79-M'79-SM'85-F'92) received the BE degree with honors in electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 1972, the ME degree with distinction in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1974, and the PhD degree in computer science from the University of Maryland, College Park, USA, in 1979.

From 1974 to 1975 he was a scientific officer in the Department of Electronics, Government of India, New Delhi. From 1975 to 1979, he was at the Computer Vision Laboratory, University of Maryland, College Park. Since 1979 he has been with the University of Illinois at Urbana-Champaign, where he is currently a professor in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute. His interests are in computer vision, robotics, image processing, image synthesis, sensors, and parallel algorithms. He has been involved in teaching, research, and development in these areas. His current research emphasizes integrated use of multiple image sources of scene information to construct 3D descriptions of scenes; the use of integrated image analysis for realistic image synthesis; parallel architectures and algorithms and special sensors for computer vision; use of the results of image analysis for a variety of applications, including visual communication, image manipulation, video retrieval, robotics, and scene navigation.

He was selected as Beckman Associate in the University of Illinois Center for Advanced Study for 1990-1991. He received University Scholar Award (1985), Presidential Young Investigator Award (1984), National Scholarship (1967-1972), and President's Merit Award (1966). He has coauthored the books *Pattern Models* (Wiley, 1983), and *Motion and Structure From Image Sequences* (Springer-Verlag, 1992) and coedited the book *Advances in Image Understanding* (IEEE Press, 1996). He is fellow of the Institute of Electrical and Electronics Engineers, American Association for Artificial Intelligence, International Association for Pattern Recognition, Association for Computing Machinery, and International Society for Optical Engineering. He is a member of the Optical Society of America. He is on the editorial boards of the journals *IEEE Transactions on Pattern Analysis and Machine Intelligence; Computer Vision, Graphics, and Image Processing; Journal of Mathematical Imaging and Vision;* and *Journal of Information Science and Technology*, and a guest coeditor of the *Artificial Intelligence Journal* special issue on vision.