A Sequential Factorization Method for Recovering Shape and Motion From Image Streams

Toshihiko Morita and Takeo Kanade, *Fellow, IEEE*

Abstract—We present a sequential factorization method for recovering the three-dimensional shape of an object and the motion of the camera from a sequence of images, using tracked features. The factorization method originally proposed by Tomasi and Kanade produces robust and accurate results incorporating the singular value decomposition. However, it is still difficult to apply the method to real-time applications, since it is based on a batch-type operation and the cost of the singular value decomposition is large. We develop the factorization method into a sequential method by regarding the feature positions as a vector time series. The new method produces estimates of shape and motion at each frame. The singular value decomposition is replaced with an updating computation of only three dominant eigenvectors, which can be performed in $O(P^2)$ time, while the complete singular value decomposition requires $O(FP^2)$ operations for an $F \times P$ matrix. Also, the method is able to handle infinite sequences, since it does not store any increasingly large matrices. Experiments using synthetic and real images illustrate that the method has nearly the same accuracy and robustness as the original method.

Index Terms—Shape from motion, singular value decomposition, feature tracking, 3D object reconstruction, image understanding, real-time vision.

1 Introduction

Recovering both the 3D shape of an object and the motion of the camera simultaneously from a stream of images is an important task and has wide applicability in many tasks, such as navigation and robot manipulation. Tomasi and Kanade [1] first developed a factorization method to recover shape and motion under an orthographic projection model, and obtained robust and accurate results. Poelman and Kanade [2] have extended the factorization method to scaled-orthographic projection and paraperspective projection. This method closely approximates perspective projection in most practical situations so that it can deal with image sequences which contain perspective distortions.

Although the factorization method is a useful technique, its applicability is, so far, limited to off-line computations for the following reasons. First, the method is based on a batch-type computation; that is, it recovers shape and motion after all the input images are given. Second, the singular value decomposition, which is the most important procedure in the method, requires $O(FP^2)$ operations for P features in F frames. Finally, it needs to store a large measurement matrix whose size increases with the number of frames. These drawbacks make it difficult to apply the factorization method to real-time applications.

This report presents a sequential factorization method that considers the input to the system as a vector time series of feature positions. The method produces estimates of shape and motion at each input frame. A covariance-like matrix is stored, instead of feature positions, and its size remains constant as the number of frames increases. The singular value decomposition is replaced with a computation, updating only three dominant eigenvectors, which can be performed in $O(P^2)$ time. Consequently, the method becomes recursive.

We first briefly review the factorization method by Tomasi and Kanade. We then present our sequential factorization method in Section 3. The algorithm's performance is tested using synthetic data and real images in Section 4.

2 Theory of the Factorization Method: Review

2.1 Formalization

The input to the factorization method is a measurement matrix W, representing image positions of tracked features over multiple frames. Assuming that there are P features over F frames, and letting $\left(x_{fp}, y_{fp}\right)$ be the image position of feature p at frame f, W is a $2F \times P$ matrix, such that

$$W = \begin{bmatrix} x_{11} & \cdots & x_{1P} \\ \vdots & & \vdots \\ x_{F1} & \cdots & x_{FP} \\ y_{11} & \cdots & y_{1P} \\ \vdots & & \vdots \\ y_{F1} & \cdots & y_{FP} \end{bmatrix}. \qquad (1)$$

- T. Morita is with Fujitsu Laboratories Ltd., Autonomous Systems Lab., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-88, Japan. E-mail: tmorita@flab.fujitsu.co.jp.
- T. Kanade is with the Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave., 234 Smith Hall, Pittsburgh, PA 15213. E-mail: tk@cs.cmu.edu.

Manuscript received 28 July 1995. 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 104066.

Each column of W contains all the observations for a single point, while each row contains all the observed x-coordinates or y-coordinates for a single frame.

Suppose that the camera orientation at frame f is represented by orthonormal vectors $\boldsymbol{i}_f$, $\boldsymbol{j}_f$, and $\boldsymbol{k}_f$, where $\boldsymbol{i}_f$ corresponds to the x-axis of the image plane, and $\boldsymbol{j}_f$ to the y-axis. The vectors $\boldsymbol{i}_f$ and $\boldsymbol{j}_f$ are collected over F frames into a motion matrix $M \in R^{2F\times 3}$ such that

$$M = \begin{bmatrix} \boldsymbol{i}_1^T \\ \vdots \\ \boldsymbol{i}_F^T \\ \boldsymbol{j}_1^T \\ \vdots \\ \boldsymbol{j}_F^T \end{bmatrix}. \tag{2}$$

Let $\boldsymbol{s}_p$ be the location of feature p in a fixed world coordinate system, whose origin is set at the center-of-mass of all the feature points. These vectors are then collected into a shape matrix $S \in R^{3\times P}$, such that

$$S = \left[\boldsymbol{s}_1 \ \dots \ \boldsymbol{s}_P\right]. \tag{3}$$

Note that

$$\sum_{p=1}^{P} \boldsymbol{s}_p = 0. \tag{4}$$

With this notation, the following equation holds by assuming an orthographic projection.

$$W = MS \tag{5}$$

Tomasi and Kanade [1] pointed out the simple fact that the rank of W is at most three, since it is the product of the $2F \times 3$ motion matrix M and the $3 \times P$ shape matrix S. Based on this rank theory, they developed a factorization method that robustly recovers the matrices M and S from W.

2.2 Subspace Computation

The actual procedure of the factorization method consists of two steps. First, the measurement matrix is factorized into two matrices of rank three using the singular value decomposition. Assume, without loss of generality, that $2F \ge P$. By computing the singular value decomposition of $W \in R^{2F\times P}$, we can obtain orthogonal matrices $U \in R^{2F\times 3}$, and $V \in R^{P\times 3}$ such that

$$W = U \Sigma V^T, \tag{6}$$

where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3)$ and $\sigma_1 \ge \sigma_2 \ge \sigma_3 > 0$. In reality, the rank of W is not exactly three, but approximately three. U is made from the first three columns of the left singular matrix of W. Likewise, Σ consists of the first three singular values, and V is made from the first three columns of the right singular matrix. By setting

$$\hat{M} = U \text{ and } \hat{S} = \Sigma V^T \tag{7}$$

we can factorize W into

$$W = \hat{M}\hat{S}, \tag{8}$$

where the product $\hat{M}\hat{S}$ is the best possible rank three approximation to W.

It is well known that the left singular vectors U span the column space of W, while the right singular vectors V span its row space. The span of U, namely *motion space*, determines the motion, and the span of V, namely *shape space*, determines the shape. The rank theory claims that the dimension of each subspace is at most three, and the first step of the factorization method finds those subspaces in the high dimensional input spaces. Both spaces are said to be dual, in the sense that one of them can be computed from the other. This observation helps us to further develop the sequential factorization method.

2.3 Metric Transformation

The decomposition of (8) is not completely unique: It is unique only up to an affine transformation. The second step of the method is necessary to find a 3×3 nonsingular matrix A, which transforms $\hat{M}$ and $\hat{S}$ into the true solutions M and S as follows.

$$M = \hat{M}A \tag{9}$$

$$S = A^{-1}\hat{S} \tag{10}$$

Observing that rows $\boldsymbol{i}_f$ and $\boldsymbol{j}_f$ of M must satisfy the normalization constraints,

$$\boldsymbol{i}_f^T\boldsymbol{i}_f = \boldsymbol{j}_f^T\boldsymbol{j}_f = 1 \text{ and } \boldsymbol{i}_f^T\boldsymbol{j}_f = 0, \tag{11}$$

we obtain the system of $3F$ overdetermined equations, such that

$$\begin{aligned} \hat{\boldsymbol{i}}_f^T L \hat{\boldsymbol{i}}_f &= 1 \\ \hat{\boldsymbol{j}}_f^T L \hat{\boldsymbol{j}}_f &= 1 \\ \hat{\boldsymbol{i}}_f^T L \hat{\boldsymbol{j}}_f &= 0 \end{aligned} \tag{12}$$

where $L \in R^{3\times 3}$ is a symmetric matrix

$$L = AA^T, \tag{13}$$

and $\hat{\boldsymbol{i}}_f$ and $\hat{\boldsymbol{j}}_f$ are the rows of $\hat{M}$. By denoting $\boldsymbol{i}_f^T = \left[i_{f1}, i_{f2}, i_{f3}\right]$, $\boldsymbol{j}_f^T = \left[j_{f1}, j_{f2}, j_{f3}\right]$, and

$$L = \begin{bmatrix} l_1 & l_2 & l_3 \\ l_2 & l_4 & l_5 \\ l_3 & l_5 & l_6 \end{bmatrix}, \tag{14}$$

the system (12) can be rewritten as

$$G\boldsymbol{l} = \boldsymbol{c}, \tag{15}$$

where $G \in R^{3F\times 6}$, $\boldsymbol{l} \in R^6$, and $\boldsymbol{c} \in R^{3F}$ are defined by

$$G = \begin{bmatrix} \boldsymbol{g}^T(\boldsymbol{i}_1, \boldsymbol{i}_1) \\ \vdots \\ \boldsymbol{g}^T(\boldsymbol{i}_F, \boldsymbol{i}_F) \\ \boldsymbol{g}^T(\boldsymbol{j}_1, \boldsymbol{j}_1) \\ \vdots \\ \boldsymbol{g}^T(\boldsymbol{j}_F, \boldsymbol{j}_F) \\ \boldsymbol{g}^T(\boldsymbol{i}_1, \boldsymbol{j}_1) \\ \vdots \\ \boldsymbol{g}^T(\boldsymbol{i}_F, \boldsymbol{j}_F) \end{bmatrix}, \quad \boldsymbol{l} = \begin{bmatrix} \boldsymbol{l}_1 \\ \vdots \\ \boldsymbol{l}_6 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ 1\end{matrix}}\right\} 2F \\ \left.\vphantom{\begin{matrix}0\\ \vdots\\ 0\end{matrix}}\right\} F \end{matrix}, \tag{16}$$

and

$$\boldsymbol{g}^T(\boldsymbol{a}_f, \boldsymbol{b}_f) = [a_{f1}b_{f1} \;\; a_{f1}b_{f2} + a_{f2}b_{f1} \;\; a_{f1}b_{f3} + a_{f3}b_{f1} \;\; a_{f2}b_{f2} \;\; a_{f2}b_{f3} + a_{f3}b_{f2} \;\; a_{f3}b_{f3}]. \quad (17)$$

The simplest solution of the system is given by the pseudoinverse method, such that

$$\boldsymbol{l} = \left(G^T G\right)^{-1} G^T \boldsymbol{c}. \quad (18)$$

The vector $\boldsymbol{l}$ determines the symmetric matrix L, whose eigendecomposition gives A. As a result, the motion M and the shape S are derived according to (9) and (10).

The matrix A is an affine transform which transforms $\hat{M}$ into M in the *motion space,* while the matrix A^{-1} transforms $\hat{S}$ into S in the *shape space*. Obtaining this transform is the main purpose of the second step of the factorization method, which we call *metric transformation*.

3 A Sequential Factorization Method

3.1 Overview

In the original factorization method, there was one measurement matrix W containing tracked feature positions throughout the image sequence. After all the input images are given and the feature positions are collected into the matrix W, the motion and shape are then computed. In real-time applications, however, it is not feasible to use this batch-type scheme. It is more desirable to obtain an estimate at each moment sequentially. The input to the system must be viewed as a vector time series. At frame f, two vectors containing feature positions such that

$$\boldsymbol{x}_f^T = \left[x_{f1}, x_{f2}, \ldots, x_{fP}\right] \text{ and } \boldsymbol{y}_f^T = \left[y_{f1}, y_{f2}, \ldots, y_{fP}\right] \quad (19)$$

are given. Immediately after receiving these vectors, the system must compute the estimates of the camera coordinates $\boldsymbol{i}_f$, $\boldsymbol{j}_f$, and the shape S_f at that frame. At the next frame, new samples $\boldsymbol{x}_{f+1}$ and $\boldsymbol{y}_{f+1}$ arrive, and new camera coordinates $\boldsymbol{i}_{f+1}$ and $\boldsymbol{j}_{f+1}$ are to be computed as well as an updated shape estimate S_{f+1}.

The key to developing such a sequential method is to observe that the shape does not change over time. The *shape space* is stationary, and, thus, it should be possible to derive S_f from S_{f-1} without performing expensive computations.

More specifically, we store the feature vectors $\boldsymbol{x}_f$ and $\boldsymbol{y}_f$ in a covariance-type matrix $Z_f \in R^{P \times P}$ defined recursively by

$$Z_f = Z_{f-1} + \boldsymbol{x}_f \boldsymbol{x}_f^T + \boldsymbol{y}_f \boldsymbol{y}_f^T. \quad (20)$$

As shown later, the rank of Z_f is at most three, and its three dominant eigenvectors Q_f span the *shape space*. Once Q_f is obtained, the camera coordinates at frame f can be computed simply by multiplying the feature vectors and the eigenvectors as follows.

$$\hat{\boldsymbol{i}}_f^T = \boldsymbol{x}_f^T Q_f, \qquad \hat{\boldsymbol{j}}_f^T = \boldsymbol{y}_f^T Q_f \quad (21)$$

This framework makes it possible to estimate camera coordinates immediately after receiving feature vectors at each frame. All information obtained up to the frame is accumulated in Q_f and used to produce the estimates at that frame.

In (20), the size of Z_f is fixed to $P \times P$, which only depends on the number of feature points. Therefore, the algorithm does not need to store any matrices whose sizes increase over time.

The computational effort in the original factorization method is dominated by the cost of the singular value decomposition. In the framework above, we need to compute eigenvectors of Z_f. Note that, however, we only need the first three dominant eigenvectors. Fortunately, several methods exist to compute only the dominant eigenvectors with much less computation necessary to compute all the eigenvectors. Before describing the details of our algorithm, we briefly review these techniques in the following section.

3.2 Iterative Eigenvector Computation

Among the existing methods which can compute dominant eigenvectors of a square matrix, we introduce two methods, the power method and orthogonal iteration [3]. The power method is the simplest, which computes the most dominant eigenvector, i.e., an eigenvector associated with the largest eigenvalue. It provides the starting point for most other techniques, and is easy to understand. The method of orthogonal iteration, which we adopt in our method, is able to compute several dominant eigenvectors.

3.2.1 Power Method

Assume that we want to compute the most dominant eigenvectors of an $n \times n$ matrix B. Given a unit two-norm vector $\boldsymbol{q}^{(0)} \in R^n$, the power method iteratively computes a sequence of vectors $\boldsymbol{q}^{(k)}$:

```
for $k = 1, 2, \ldots$
    $\boldsymbol{y}^{(k)} = B\boldsymbol{q}^{(k-1)}$
    $\boldsymbol{q}^{(k)} = \boldsymbol{y}^{(k)} / \left\|\boldsymbol{y}^{(k)}\right\|_2$
end
```

The second step of the iteration is simply a normalization that prevents $\boldsymbol{q}^{(k)}$ from becoming very large or very small. The vectors $\boldsymbol{q}^{(k)}$ generated by the iteration converge to the most dominant eigenvector of B. To examine the convergence property of the power method, suppose that B is diagonalizable. That is, $X^{-1}BX = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with an orthogonal matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, and $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$. If

$$\boldsymbol{q}^{(0)} = c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + \ldots + c_n\boldsymbol{x}_n \quad (22)$$

and $c_1 \neq 0$, then it follows that

$$\begin{aligned} \boldsymbol{q}^{(k)} = \xi B^k \boldsymbol{q}^{(0)} &= \xi\left(\sum_{j=1}^{n} c_j \lambda_j^k \boldsymbol{x}_j\right) \\ &= \xi c_1 \lambda_1^k \left(\boldsymbol{x}_1 + \sum_{j=2}^{n} \frac{c_j}{c_1}\left(\frac{\lambda_j}{\lambda_1}\right)^k \boldsymbol{x}_j\right) \end{aligned} \quad (23)$$

where ξ is a constant. Since $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$, (23) shows that the vectors $\boldsymbol{q}^{(k)}$ point more and more accurately toward the direction of the dominant eigenvector $\boldsymbol{x}_1$, and the convergence factor is the ratio $r = |\lambda_2/\lambda_1|$.

3.2.2 Orthogonal Iteration

A straightforward generalization of the power method can be used to compute several dominant eigenvectors of a symmetric matrix. Assume that we want to compute p dominant eigenvectors of a symmetric matrix $B \in R^{n\times n}$, where $1 \le p \le n$. Starting with an $n \times p$ matrix Q_0 with orthonormal columns, the method of orthogonal iteration generates a sequence of matrices $Q_k \in R^{n\times p}$:

```
for k = 1, 2, ...
    $Y_k = BQ_{k-1}$
    $Q_k R_k = Y_k$        (QR factorization)
end
```

The second step of the above iteration is the QR factorization of Y_k, where Q_k is an orthogonal matrix and R_k is an upper triangular matrix. The QR factorization can be achieved by the Gram-Schmidt process. This step is viewed as a normalization process that is similar to the normalization used in the power method.

Suppose that $X^T BX = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the eigendecomposition of B with an orthogonal matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, and $|\lambda_1| > |\lambda_2| \ge \ldots \ge |\lambda_n|$. It has been shown in [3] that the subspace range (Q_k) generated by the iteration converges to span $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$ at a rate proportional to $|\lambda_{p+1}/\lambda_p|$, i.e.,

$$\mathrm{dist}\,(\mathrm{range}\,(Q_k),\ \mathrm{range}\,(X_p)) \le \zeta \left|\frac{\lambda_{p+1}}{\lambda_p}\right|^k, \tag{24}$$

where $X_p = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ and ζ is a constant. The function *dist* represents the subspace distance defined by

$$\mathrm{dist}\,(\mathrm{range}\,(Q_k),\ \mathrm{range}\,(X_p)) = \left\| Q_k Q_k^T - X_p X_p^T \right\|_2 \tag{25}$$

The method offers an attractive alternative to the singular value decomposition in situations where B is a large matrix and a few of its largest eigenvalues are needed. In our case, these conditions clearly hold. In addition, when the rank theory of the factorization method [1] holds, the ratio $|\lambda_4/\lambda_3|$ is very small. As a result, we should achieve fast convergence for computing the first three eigenvectors.

3.3 Sequential Factorization Algorithm

As in the original method, the sequential factorization method consists of two steps, sequential *shape space* computation and sequential metric transformation.

3.3.1 A Sequential Shape Space Computation

In the sequential factorization method, we consider the feature vectors $\boldsymbol{x}_f^T$ and $\boldsymbol{y}_f^T$ as a vector time series. Let us denote the measurement matrix in the original factorization method at frame f by W_f. Then, it grows in the following manner:

$$W_1 = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{y}_1^T \end{bmatrix},\ W_2 = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \boldsymbol{y}_1^T \\ \boldsymbol{y}_2^T \end{bmatrix},\ \ldots,\ W_f = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_f^T \\ \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_f^T \end{bmatrix},\ \ldots \tag{26}$$

Now, let us define a matrix time series $Z_f \in R^{P\times P}$ by

$$Z_f = Z_{f-1} + \boldsymbol{x}_f \boldsymbol{x}_f^T + \boldsymbol{y}_f \boldsymbol{y}_f^T. \tag{27}$$

From the definition, it follows that

$$Z_f = W_f^T W_f. \tag{28}$$

Since the rank of W_f is at most three, the rank of Z_f is also at most three. If

$$W_f = U_f \Sigma_f V_f^T \tag{29}$$

is the singular value decomposition of W_f, where $U_f \in R^{2f\times 3}$ and $V_f \in R^{P\times 3}$ are orthogonal matrices, and $\Sigma_f = \mathrm{diag}\,(\sigma_{f,1}, \sigma_{f,2}, \sigma_{f,3})$, then

$$Z_f = \left(U_f \Sigma_f V_f^T\right)^T U_f \Sigma_f V_f^T = V_f \Sigma_f^2 V_f^T. \tag{30}$$

This means the eigenvectors of Z_f are equivalent to the right singular vectors V_f of W_f. Hence, it is possible to obtain the *shape space* by computing the eigenvectors of Z_f.

To compute V_f, we combine orthogonal iteration with updating by (27). Given a $P \times 3$ matrix Q_0 with orthonormal columns and a null matrix $Z_0 \in R^{P\times P}$, the following algorithm generates a sequence of matrices $Q_f \in R^{P\times 3}$:

```
[Algorithm (1)] for f = 1, 2, ....
    1) $Z_f = Z_{f-1} + \boldsymbol{x}_f \boldsymbol{x}_f^T + \boldsymbol{y}_f \boldsymbol{y}_f^T$
    2) $Y = Z_f Q_{f-1}$
    3) $Q_f R = Y$        (QR factorization)
end
```

The index f corresponds to the frame number and each iteration is performed frame by frame. The matrix Q_f generated by the algorithm is expected to converge to the eigenvectors V_f of Z_f. While the original orthogonal iteration works with a fixed matrix, the above algorithm works with the matrix Z_f, which varies from iteration to iteration, incorporating new features. In other words, the sequential factorization method folds the update of Z_f into the orthogonal iteration. If range (V_f) randomly changes over time, no convergence is expected to appear. However, it can be shown that

$$\mathrm{range}\,(V_f) = \mathrm{range}\,(W_f^T) = \mathrm{range}\,(S^T), \text{for all } f. \tag{31}$$

Therefore, range (V_f) is stationary and range (Q_f) converges to range (V_f) as in the orthogonal iteration. Even when noise exists, if the noise is uncorrelated, or the noise space is orthogonal to the signal space range (V_f), then range (V_f) is still equal to range (S^T), and the convergence can be shown. The following convergence rate of the algorithm is deduced from the convergence rate of the orthogonal iteration.

$$\text{dist}\,(\text{range}\,(Q_f),\ \text{range}\,(V_f)) \le c\prod_{k=1}^{f}\left|\frac{\sigma_{k,4}}{\sigma_{k,3}}\right| \tag{32}$$

3.3.2 Stationary Basis for the Shape Space

Algorithm (1), presented in the previous section, produces the matrix Q_f, which converges to the matrix V_f that spans the *shape space*. The true shape and motion are determined from the *shape space* by a metric transformation. It is not straightforward at this point, however, to apply the metric transformation sequentially. The problem is that, even though range (V_f) is stationary, the matrix V_f itself changes as the number of frames increases. This is due to the nature of singular vectors. They are the basis for the row and column subspaces of a matrix, and the singular value decomposition chooses them in a special way. They are more than just orthonormal. As a result, they rotate in the 3D subspace range (V_f). Recall that the matrix A obtained in metric transformation (9) is a transform from $\hat{M}_f$ (or U_f) to M_f in the subspace range $(\hat{M}_f)$. Since V_f changes at each frame, U_f also changes. Consequently, the matrix A also changes frame by frame.

For clarity, let us denote an A matrix at frame f as A_f. The fact that A_f changes at each frame makes it difficult to estimate A_f iteratively and efficiently. Thus, we need to add an additional process to obtain stationary basis for the *shape space* to update matrix A_f.

Let us define a projection matrix $H_f \in R^{P\times P}$ onto range (Q_f) by

$$H_f = Q_f Q_f^T, \tag{33}$$

where Q_f is the output from Algorithm (1). Obviously, the rank of H_f is at most three. Since range (Q_f) (= range $(\hat{M}_f)$) is stationary, the projection matrix H_f must be stationary. It is thus possible to obtain the stationary basis for the *shape space* by replacing Q_f with the eigenvectors of H_f.

An iterative method similar to Algorithm (1) can be used to reduce the amount of computation. Given a $P \times 3$ matrix $\overline{Q}_0$ with orthonormal columns, the iterative method below generates a matrix $\overline{Q}_f \in R^{P\times 3}$, which provides the stationary basis for the *shape space*.

```
[Algorithm (2)] for f = 1, 2, ....
    $H_f = Q_f Q_f^T$
    $Y = H_f \overline{Q}_{f-1}$
    $\overline{Q}_f R = Y$        (QR factorization)
end
```

3.3.3 Sequential Metric Transformation

In the previous section, we derived the *shape space* in terms of $\overline{Q}_f$. Once $\overline{Q}_f$ is obtained, it is possible to compute camera coordinates $\hat{\boldsymbol{i}}_f$ and $\hat{\boldsymbol{j}}_f$ as

$$\hat{\boldsymbol{i}}_f^T = \boldsymbol{x}_f^T\overline{Q}_f, \quad \hat{\boldsymbol{j}}_f^T = \boldsymbol{y}_f^T\overline{Q}_f \tag{34}$$

These coordinates are used to solve the overdetermined equations (12) and the true camera coordinates are recovered in the same way as in the original method. Doing so, however, requires storing all the coordinates $\hat{\boldsymbol{i}}_f$ and $\hat{\boldsymbol{j}}_f$, the number of which may be very large. Instead, we use the following sequential algorithm.

```
[Algorithm (3)] for f = 1, 2, ....
    $\hat{\boldsymbol{i}}_f^T = \boldsymbol{x}_f^T\overline{Q}_f, \quad \hat{\boldsymbol{j}}_f^T = \boldsymbol{y}_f^T\overline{Q}_f$
    $D_f = D_{f-1} + \boldsymbol{g}(\hat{\boldsymbol{i}}_f,\hat{\boldsymbol{i}}_f)\boldsymbol{g}^T(\hat{\boldsymbol{i}}_f,\hat{\boldsymbol{i}}_f) +$
        $\boldsymbol{g}(\hat{\boldsymbol{j}}_f,\hat{\boldsymbol{j}}_f)\boldsymbol{g}^T(\hat{\boldsymbol{j}}_f,\hat{\boldsymbol{j}}_f) + \boldsymbol{g}(\hat{\boldsymbol{i}}_f,\hat{\boldsymbol{j}}_f)\boldsymbol{g}^T(\hat{\boldsymbol{i}}_f,\hat{\boldsymbol{j}}_f)$
    $E_f = E_{f-1} + \boldsymbol{g}(\hat{\boldsymbol{i}}_f,\hat{\boldsymbol{i}}_f) + \boldsymbol{g}(\hat{\boldsymbol{j}}_f,\hat{\boldsymbol{j}}_f)$
end
```

Let G_f and $\boldsymbol{c}_f$ be the matrices G and $\boldsymbol{c}$ at frame f, where G and $\boldsymbol{c}$ are defined in Section 2.3. From the definition, it follows that

$$D_f = G_f^T G_f \tag{35}$$

$$E_f = G_f^T\boldsymbol{c}_f. \tag{36}$$

Assigning (35) and (36) to (18), we have

$$\boldsymbol{l}_f = D_f^{-1}E_f \tag{37}$$

which gives the symmetric matrix L_f. The eigendecomposition of L_f yields the affine transform A_f and, as a result, the camera coordinates and the shape are obtained as follows:

$$\boldsymbol{i}_f^T = \hat{\boldsymbol{i}}_f^T A_f, \quad \boldsymbol{j}_f^T = \hat{\boldsymbol{j}}_f^T A_f \tag{38}$$

$$S_f = \boldsymbol{A}_f^{-1}\overline{Q}_f \tag{39}$$

Algorithm (3) followed by (37), (38), and (39) completes the sequential method. The size of matrices D_f and E_f are fixed to 6×6 and 6×1, and the method does not store any matrices that grow, even in the sequential metric transformation.

4 Experiments

4.1 Synthetic Data

In this section, we compare the accuracy of our sequential factorization method with that of the original factorization method. Since both methods are essentially based on the rank theory, we do not expect any difference in the results. Our purpose here is to confirm that the sequential method has the same accuracy of shape and motion recovery as the original method.

4.1.1 Data Generation

The object in this experiment consists of 100 random feature points. The sequences are created using a perspective projection of those points. The image coordinates of each point are perturbed by adding Gaussian noise, which we assume to simulate tracking error and image noise. The standard deviation of the Gaussian noise is set to two pixels of a 512 × 512 pixel image. The distance of the object center from the camera is fixed to ten times the object size. The focal length is chosen so that the projection of the object covers the whole 512 × 512 image. The camera is rotated as shown in Fig. 1, while the object is translated to keep its image at the image center. Quantization errors are not added, since we assume that we are able to track features with a subpixel resolution.

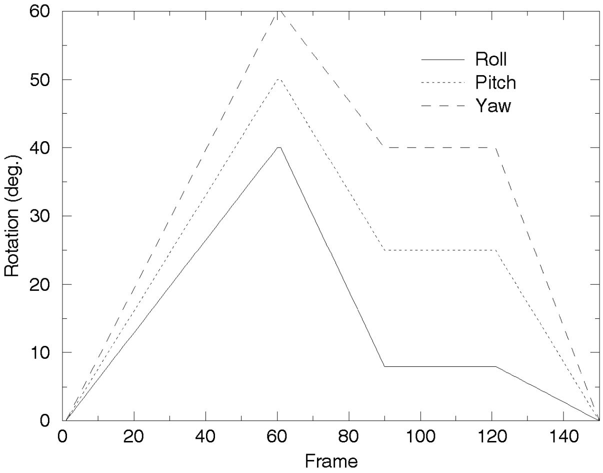


Fig. 1. True camera motion. The camera roll, pitch, and yaw are varied as shown in this figure. The sequence consists of 150 frames.

When discussing the accuracy of the sequential method, one needs to consider its dynamic property regarding the 3D recovery. The accuracy of the recovery at a particular frame by the sequential method depends on the total amount of motion up to that time, since the recovery is made only from the information obtained up to that time. At the beginning of an image sequence, for example, the motion is generally small, so high accuracy cannot be expected. The accuracy generally improves as the motion becomes larger. The original method does not have this dynamic property, since it is based on a batch-type scheme and uses all the information throughout the sequence.

In order to compare both methods under the same conditions, we perform the following computations beforehand. First, we form a submatrix W_f, which only contains the feature positions up to frame *f*. The original factorization is applied to the submatrix, then the results are kept as solutions at frame *f*. They are the best estimates given by the original method. Repeating this process for each frame, we derive the best estimates, with which our results are compared.

4.1.2 Accuracy of the Sequential Shape Space Computation

We first discuss the convergence property of the sequential *shape space* computation. The sequential factorization method starts with Algorithm (1) in Section 3.3.1, iteratively generating the matrix Q_f, which is an estimate for the true *shape space* S^T. Let us represent the estimation error with respect to the true *shape space* by

$$E_s = \text{dist}\,(\text{range}\,(Q_f), \text{range}\,(S^T)) \tag{40}$$

Recall that the function *dist* provides a notion of difference between two spaces. On the other hand, the original method produces the best estimate for the *shape space* by computing the right singular vectors V_f of the submatrix W_f, and its error, with respect to the true *shape space,* is also represented by

$$E_o = \text{dist}\,(\text{range}\,(V_f), \text{range}\,(S^T)) \tag{41}$$

Comparing both errors, Fig. 2 shows that they are almost identical. That is, the errors given by the sequential method are almost equal to those given by the original method.

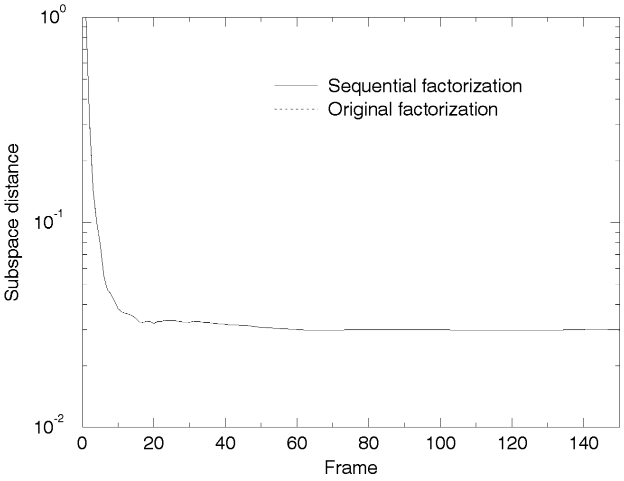


Fig. 2. Shape space errors. Shape space estimation errors by the sequential method (solid line) and the original method (dashed line) with respect to the true shape space. The errors are defined by subspace distance and plotted logarithmically.

At the beginning of the sequence, the amount of motion is small and both errors are relatively large. The ratio of the fourth to third singular values, shown in Fig. 3, also indicates that it is difficult to achieve good accuracy at the beginning. Both errors, however, quickly become smaller as the camera motion becomes larger. After about the 20th frame, constant errors of 3×10^{-2} are observed in this experiment.

The solutions given by the two methods are so close that the graphs are completely overlapped. Thus, we also plot their difference defined by

$$\Delta E = \text{dist}\,(\text{range}\,(Q_f), \text{range}\,(V_f)) \tag{42}$$

in Fig. 4. Although ΔE is relatively large at the beginning, it quickly becomes very small. In fact, after about the 30th frame, ΔE is less than 1×10^{-7}, while E_s and E_o are both 3×10^{-2}.

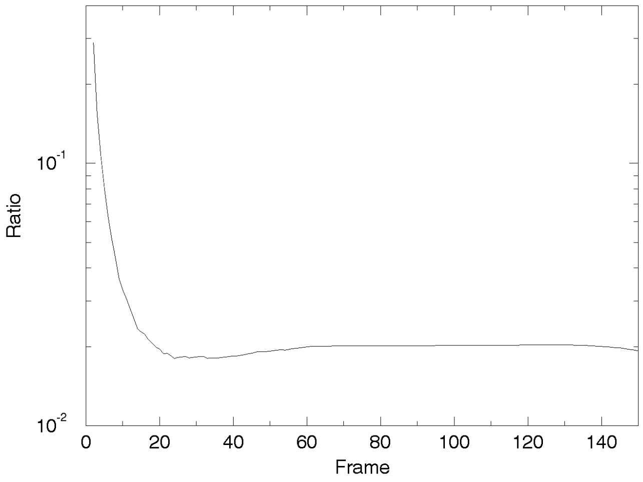


Fig. 3. Singular value ratio. The ratio of the fourth to third singular values, that is σ_4 / σ_3.

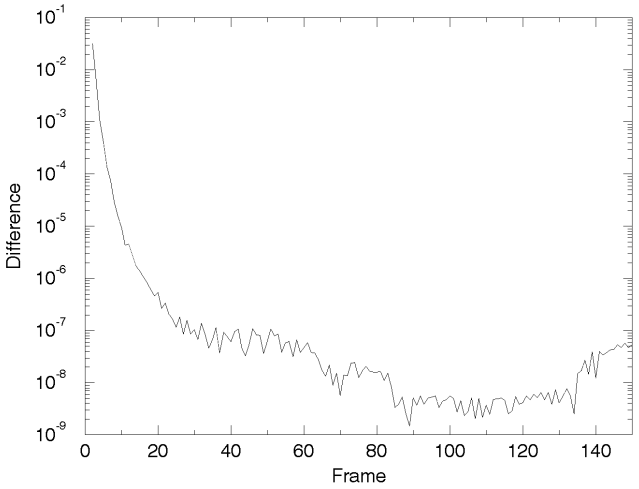


Fig. 4. Difference of shape space errors. The difference of the estimates by the sequential and original methods versus the frame number. The difference is plotted logarithmically.

4.1.3 *Accuracy of the Motion and Shape Recovery*

The three plots of Fig. 5 show errors in roll, pitch, and yaw in the recovered motion: The solid lines correspond to the sequential method, the dotted lines to the original method. The difference in motion errors between the original and sequential methods is quite small.

Both results are unstable for a short period at the beginning of the sequence. After that, they show two kinds of errors: random and structural. Random errors are due to Gaussian noise added to the feature positions. Structural errors are due to perspective distortion, and relate to the motion patterns. The structural errors show a negative peak at about the 60th frame and are almost constant between the 90th and 120th frames. Note that the pattern corresponds to the motion pattern shown in Fig. 1.

Of course, these intrinsic errors cannot be eliminated in the sequential method. The point to observe is that the differences between the two solutions are sufficiently smaller than the intrinsic errors.

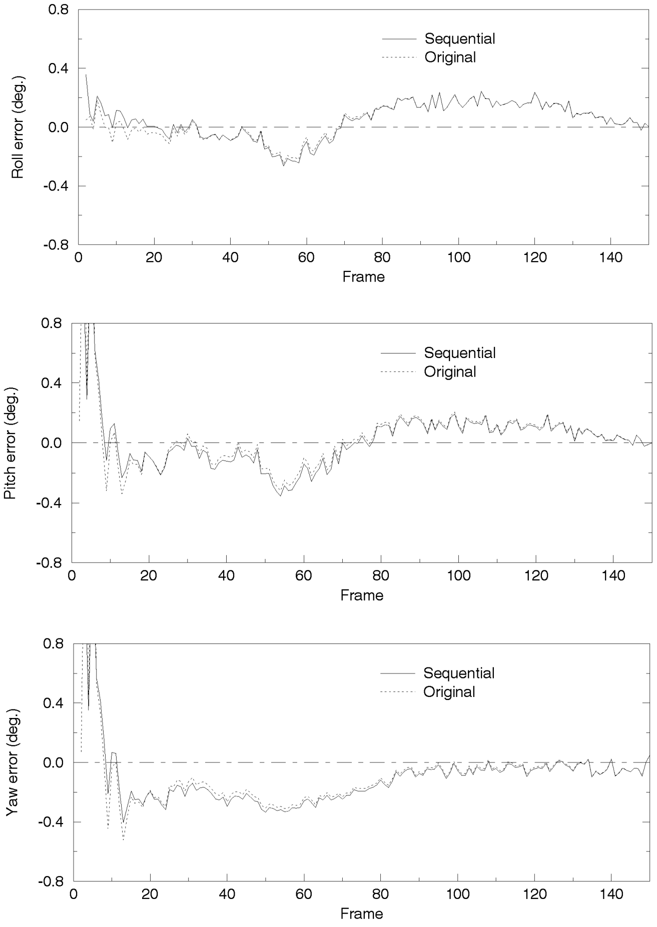


Fig. 5. Motion errors. Errors of recovered camera roll (top), pitch (middle), and yaw (bottom). The errors given by the sequential method are plotted with solid lines, while the errors given by the original method are plotted with dotted lines.

Shape errors, which are compared in Fig. 6, also indicate the same results. Again, the differences between the two methods are quite small compared to the intrinsic errors which the original method possesses. Note that no Gaussian noise appears in the shape errors, since they are averaged over all the feature points.

We conclude from these results that the sequential method is nearly as accurate as the original method except that some extra frames are required to converge.

4.2 Real Images

Experiments were performed on two sets of real images. The first set is an image sequence of a satellite rotating in space. The other experiment uses a long video recording (764 images) of a house taken with a hand-held camera. These experiments demonstrate the applicability of the sequential factorization method in real situations. In both experiments, features are selected and tracked using the method presented by Tomasi and Kanade [1].

4.2.1 *Satellite Images*

Fig. 7 shows an image of the satellite with selected features indicated by small squares. The image sequence was digitized from a video recording [4] actually taken by a space shuttle

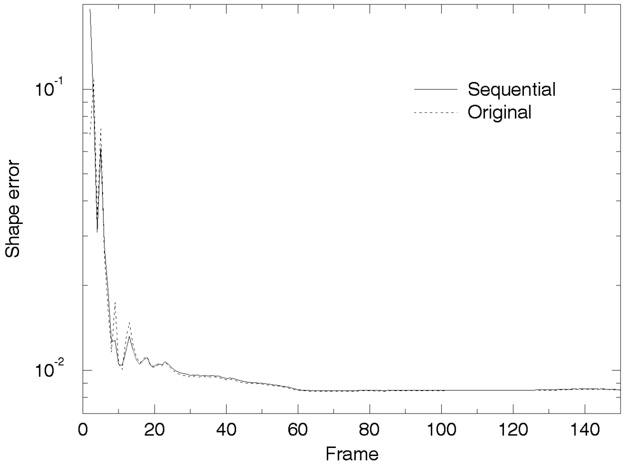


Fig. 6. Shape error. This figure compares the shape errors given by the two methods. The errors given by the sequential method are plotted with solid lines, while the errors given by the original method are plotted with dotted lines. The errors are computed as the root-mean-square errors of the recovered shape with respect to the true shape, at each frame.

astronaut. The feature tracker automatically selected and tracked 32 features throughout the sequence of 101 images. Of these, five features on the astronaut maneuvering around the satellite were manually eliminated because they had a different motion. Thus, the remaining 27 features were processed. Fig. 8 shows the recovered motion in terms of roll, pitch, and yaw. The side view of the recovered shape is displayed in Fig. 9, where the features on the solar panel are marked with opaque squares and others with filled squares. No ground-truth is available for the shape or the motion in this experiment. Yet, it appears that the solutions are satisfactory, since the features on the solar panel almost lie in a single line in the side view.

4.2.2 House Images

Fig. 10 shows the first image of the sequence used in the second experiment. Using a hand-held camera, one of the authors took this sequence while walking. It consists of 764 images which correspond to about 25 seconds. The feature tracker detected and tracked 62 features. The recovered motion and shape are shown in Figs. 11 and 12. It is clearly seen that the shape is qualitatively correct. It is also reasonable to observe that only the camera yaw is increasing, because the camera is moving parallel to the ground. In addition, note that the computed roll motion reveals the pace of the recorder's steps, which is about one step per second.

Further evaluation of accuracy in these experiments is difficult. However, this qualitative analysis of the results with real images, and quantitative analysis of the results with synthetic data essentially shows that the sequential method works as well with real images as the original batch method.

4.3 Computational Time

Finally, we compare the processing time of the sequential method with the original method. The computational complexity of the original method is dominated by the cost of the singular value decomposition, which needs $14FP^2 + 11P^3/3$ computations for a $2F \times P$ measurement

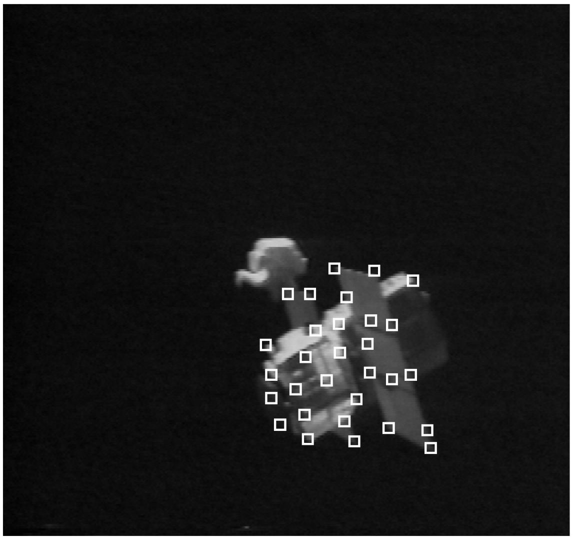


Fig. 7. An image of a satellite. The first frame of the satellite image sequence. The superimposed squares indicate the selected features.

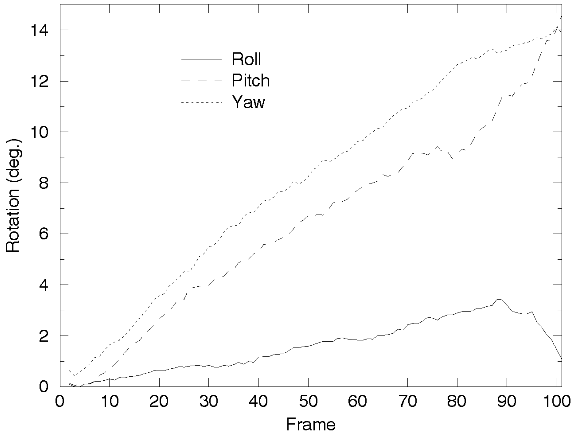


Fig. 8. Recovered motion of satellite. Recovered camera roll (solid line), pitch (dashed line), and yaw (dotted line) for the satellite image sequence.

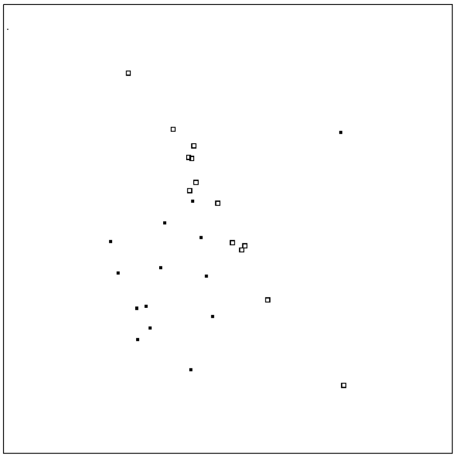


Fig. 9. Side view of the recovered shape. A side view of the recovered shape of the satellite. The features on the solar panel are shown with opaque squares and others with filled squares. Notice that the features on the solar panel correctly lie in a single plane.

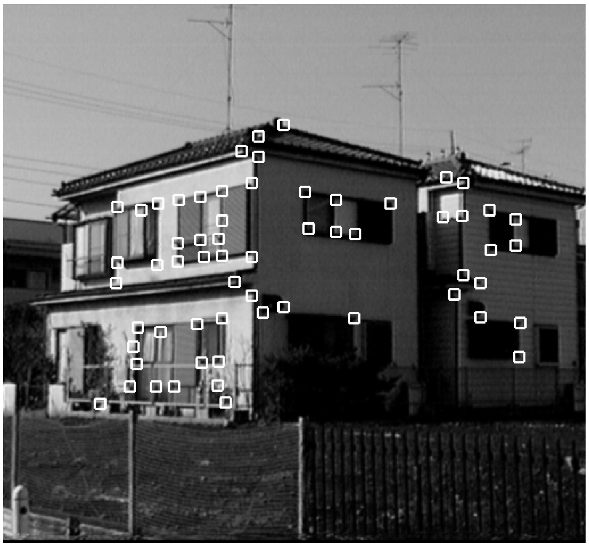


Fig. 10. An image of a house. The first frame of the house image sequence. The superimposed squares indicate the selected features.

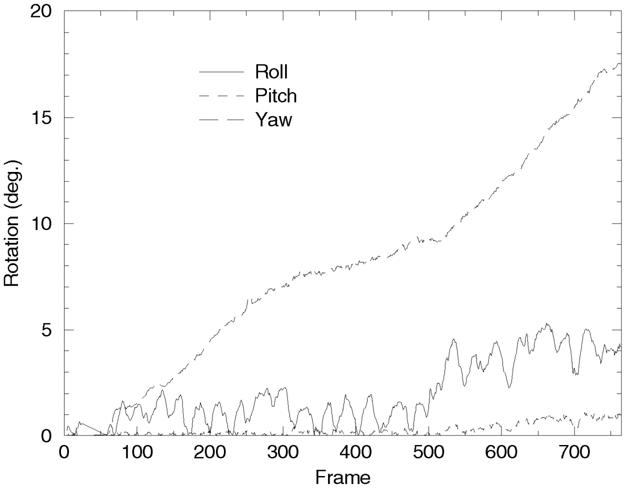


Fig. 11. Recovered motion of house. Recovered camera roll (solid line), pitch (dashed line), and yaw (dotted line) for the house image sequence.

matrix with $2F \geq P$ [5]. Note that F corresponds to the number of frames and P to the number of features. On the other hand, the complexity of the sequential method is $22P^2 + 54P$ for computing dominant eigenvectors, plus $4P^2$ for updating the Z matrix. Computing the solution for frame F, therefore, takes only $O(P^2)$ using the sequential method, while the original method would require $O(FP^2)$ operations.

Fig. 13 shows the actual processing time of the sequential method on a SparcStation-10 compared together with that of the original method. The number of features varied from 10 to 500, while the number of frames was fixed at 120. The processing time for selecting and tracking features was not included. The singular value decomposition of the original method is based on a routine found in [6]. The results sufficiently agree with our analysis above. In addition, when the number of features is less than 40, the sequential method is possible to run within 1/30 s, which means video-rate processing on a SparcStation-10.

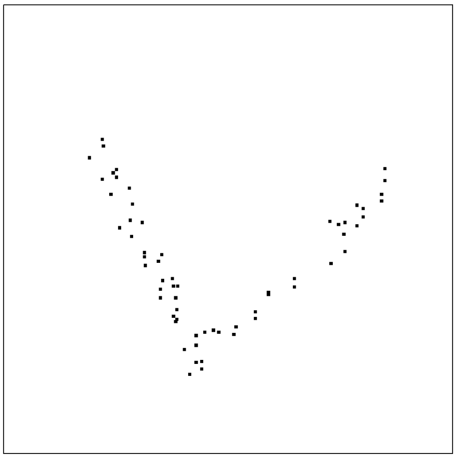


Fig. 12. Top view of the recovered shape. A view of the recovered shape of the house from above. The features on the two side walls are correctly recovered.

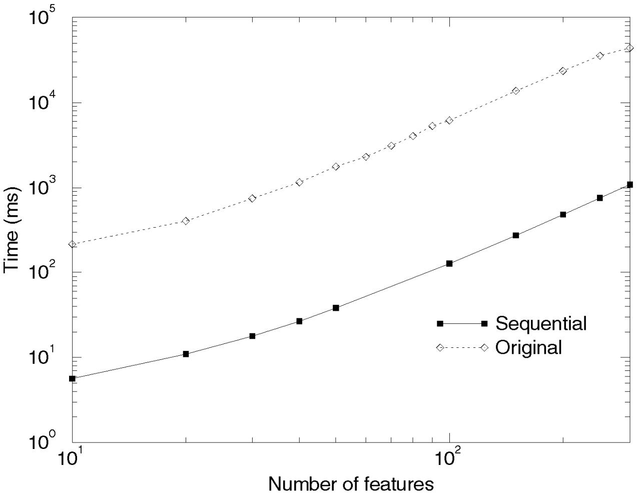


Fig. 13. Processing time. The processing time of the sequential method on a Sun4/10 (solid line) compared with that of the original method (dotted line), as a function of the number of features which is varied from 10 to 500. The number of frames is fixed at 120.

5 Conclusions

We have presented the sequential factorization method, which provides estimates of shape and motion at each frame from a sequence of images. The method produces as accurate and robust results as the original method, while significantly reducing the computational complexity. The reduction in complexity is important for applying the factorization method to real-time applications. Furthermore, the method does not require storing any growing matrices so that its implementation in VLSI or DSP is feasible.

Faster convergence in the *shape space* computation could be achieved using more sophisticated algorithms, such as the orthogonal iteration with Ritz acceleration [3] instead of the basic orthogonal iteration. Also, it is possible to use scaled orthographic projection or paraperspective projection [2] to improve the accuracy of the sequential factorization method.

ACKNOWLEDGMENTS

The authors wish to thank Conrad J. Poelman and Richard Madison for their helpful comments.

REFERENCES

[1] C. Tomasi and T. Kanade, “Shape and Motion from Image Streams: A Factorization Method,” Technical Report CMU-CS-91-172, Carnegie Mellon Univ., 1991. Later published as 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, Nov. 1992.

[2] C.J. Poelman and T. Kanade, “A Paraperspective Factorization Method for Shape and Motion Recovery,” Technical Report CMU-CS-92-208, Carnegie Mellon Univ., 1992. Revised and superceded as Technical Report CMU-CS-93-219, Carnegie Mellon Univ., 1993. Part of the latter technical report was presented at and included in the *Proc. Third European Conf. Computer Vision*, vol. 1, pp. 97-110, Stockholm, Sweden, May 1994.

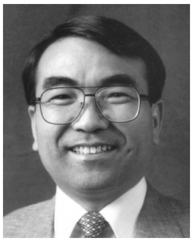
[3] G.H. Golub and C.F. Van Loan, *Matrix Computations*, second edition. The Johns Hopkins Univ. Press, 1989.

[4] “Satellite Rescue in Space: Highlights of Shuttle Flights 41C & 51A,” #V41, The Holiday Video Library, Finley-Holiday Film Corporation.

[5] P. Comon and G.H. Golub, “Tracking a Few Extreme Singular Values and Vectors in Signal Processing,” *Proc. IEEE*, vol. 78, no. 8, pp. 1,327-1,343, 1990.

[6] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge Univ. Press, 1988.

Toshihiko Morita received the BE degree in mechanical engineering from the University of Tokyo, Japan, in 1984. He has been with Fujitsu Laboratories Ltd., Kawasaki, Japan, since 1984. From 1992 to 1993, he was a visiting research scientist at the Robotics Institute, Carnegie Mellon University, Pittsburgh. His current research interests include pattern recognition, robot vision, and image processing systems.



Takeo Kanade received his PhD in electrical engineering from Kyoto University, Japan, in 1974. After holding a faculty position in the Department of Information Science at Kyoto University, he joined Carnegie Mellon University in 1980, where he is currently the U.A. Helen Whitaker professor of computer science and director of the Robotics Institute.

Dr. Kanade has worked in multiple areas of robotics: vision, manipulators, autonomous mobile robots, and sensors. He has written more that 150 technical papers and reports in these areas. He has been the principal investigator of several major vision and robotics projects at Carnegie Mellon. In the area of education, he was a founding chairperson of CMU’s robotics PhD program, probably the first of its kind.

Dr. Kanade has been elected to the National Academy of Engineering and is a fellow of the IEEE, a founding fellow of the American Association of Artificial Intelligence, and the founding editor of *International Journal of Computer Vision*. He has received several awards, including the Joseph Engelberger Award in 1995, and the Marr Prize Award in 1990.