Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 9795–9802, November 1993
Colloquium Paper

This paper was presented at a colloquium entitled "Images of Science: Science of Images," organized by Albert V. Crewe, held January 13 and 14, 1992, at the National Academy of Sciences, Washington, DC.

Shape and motion from image streams: A factorization method

CARLO TOMASI AND TAKEO KANADE

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

ABSTRACT **Inferring scene geometry and camera motion from a stream of images is possible in principle, but it is an ill-conditioned problem when the objects are distant with respect to their size. We have developed a *factorization method* that can overcome this difficulty by recovering shape and motion without computing depth as an intermediate step. An image stream can be represented by the $2F \times P$ measurement matrix of the image coordinates of P points tracked through F frames. Under orthographic projection this matrix is of rank 3. Using this observation, the factorization method uses the singular value decomposition technique to factor the measurement matrix into two matrices, which represent object shape and camera motion, respectively. The method can also handle and obtain a full solution from a partially filled-in measurement matrix, which occurs when features appear and disappear in the image sequence due to occlusions or tracking failures. The method gives accurate results and does not introduce smoothing in either shape or motion. We demonstrate this with a series of experiments on laboratory and outdoor image streams, with and without occlusions.**

Section 1. Introduction

The structure from motion problem—recovering scene geometry and camera motion from a sequence of images—has attracted much of the attention of the vision community over the last decade. Yet it is common knowledge that existing solutions work well for perfect images but are very sensitive to noise. We present a method that we have developed called the factorization method, which can robustly recover shape and motion from a sequence of images without assuming a model of motion, such as constant translation or rotation.

More specifically, an image sequence can be represented as a $2F \times P$ measurement matrix W, which is made up of the horizontal and vertical coordinates of P points tracked through F frames. If image coordinates are measured with respect to their centroid, we prove the *rank theorem*: under orthography, the measurement matrix is of rank 3. As a consequence of this theorem, we show that the measurement matrix can be factored into the product of two matrices R and S. Here, R is a $2F \times 3$ matrix that represents camera rotation, and S is a $3 \times P$ matrix that represents shape in a coordinate system attached to the object centroid. The two components of the camera translation along the image plane are computed as averages of the rows of W. When features appear and disappear in the image sequence due to occlusions or tracking failures, the resultant measurement matrix W is only partially filled in. The factorization method can handle this situation by growing a partial solution obtained from an initial full submatrix into a full solution with an iterative procedure.

The rank theorem precisely captures the nature of the redundancy that exists in an image sequence and permits a large number of points and frames to be processed in a conceptually simple and computationally efficient way to reduce the effects of noise. The resulting algorithm is based on the singular value decomposition, which is numerically well-behaved and stable. The robustness of the recovery algorithm in turn enables us to use an image sequence with a very short interval between frames (an *image stream*), which makes feature tracking relatively easy.

We have demonstrated the accuracy and robustness of the factorization method in a series of experiments on laboratory and outdoor sequences, with and without occlusions.

Section 2. Relation to Previous Work

In Ullman's (1) original proof of existence of a solution for the structure from motion problem under orthography, as well as in the perspective formulation in ref. 2, the coordinates of feature points in the world are expressed in a world-centered system of reference. Since then, however, this choice has been replaced by most computer vision researchers with a camera-centered representation of shape (3–12, †, ‡). With this representation, the position of feature points is specified by their image coordinates and by their depths, defined as the distances between the camera center and the feature points, measured along the optical axis. Unfortunately, although a camera-centered representation simplifies the equations for perspective projection, it makes shape estimation difficult, unstable, and noise sensitive.

There are two fundamental reasons for this. First, when camera motion is small, effects of camera rotation and translation can be confused with each other; for example, small rotation about the vertical axis and small translation along the horizontal axis both generate a very similar change in an image. Any attempt to recover or differentiate between these two motions, though doable mathematically, is naturally noise sensitive. Second, the computation of shape as relative depth—for example, the height of a building as the difference of depths between the top and the bottom—is very sensitive to noise, since it is a small difference between large values. These difficulties are especially magnified when the objects are distant from the camera relative to their sizes, which is usually the case for interesting applications such as site modeling.

The factorization method we present in this paper takes advantage of the fact that both difficulties disappear when the problem is reformulated in world-centered coordinates, unlike the conventional camera-centered formulation. This new (old—in a sense) formulation links object-centered shape to image motion directly, without using retinotopic depth as an intermediate quantity, and leads to a simple and well-behaved solution. Furthermore, the mutual independence of shape and motion in world-centered coordinates makes it possible to cast the structure-from-motion problem as a factorization

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

[†] Heel, J., Proceedings of the DARPA Image Understanding Workshop, May 23–26, 1989, Palo Alto, CA, pp. 702–713.

[‡] Spetsakis, M. E. & Aloimonos, J., Proceedings of the IEEE Workshop on Visual Motion, March 1989, Irvine, CA, pp. 229–237.

problem, in which a matrix representing image measurements is decomposed directly into camera motion and object shape.

We have previously introduced this factorization method,[§][¶] where we treated the case of single-scan line images in a flat, two-dimensional world. In ref. 13 we presented the theory for the case of arbitrary camera motion in three dimensions and full two-dimensional images. This paper extends the factorization method for dealing with feature occlusions as well as presenting more experimental results with real-world images. Debrunner and Ahuja (14, [||]) have pursued an approach related to ours but using a different formalism. Assuming that motion is constant over a period, they provide both closed-form expressions for shape and motion and an incremental solution (one image at a time) for multiple motions by taking advantage of the redundancy of measurements. Boult and Brown[**] have investigated the factorization method for multiple motions, in which they count and segment separate motions in the field of view of the camera.

Section 3. The Factorization Method

Given an image stream, suppose that we have tracked P feature points over F frames. We then obtain trajectories of image coordinates $\{(u_{fp}, v_{fp}) | f = 1, \ldots, F, p = 1, \ldots, P\}$. We write the horizontal feature coordinates u_{fp} into an $F \times P$ matrix U: we use one row per frame and one column per feature point. Similarly, an $F \times P$ matrix V is built from the vertical coordinates v_{fp}. The combined matrix of size $2F \times P$

$$W = \begin{bmatrix} U \\ \hline V \end{bmatrix}$$

is called the *measurement matrix*. The rows of the matrices U and V are then registered by subtracting from each entry the mean of the entries in the same row:

$$\tilde{u}_{fp} = u_{fp} - a_f \qquad \tilde{v}_{fp} = v_{fp} - b_f, \qquad [3.1]$$

where $a_f = \frac{1}{P} \sum_{p=1}^{P} u_{fp}$ and $b_f = \frac{1}{P} \sum_{p=1}^{P} v_{fp}$. This produces two new $F \times P$ matrices $\tilde{U} = [\tilde{u}_{fp}]$ and $\tilde{V} = [\tilde{v}_{fp}]$. The matrix

$$\tilde{W} = \begin{bmatrix} \tilde{W} \\ \hline \tilde{V} \end{bmatrix}$$

is called the *registered measurement matrix*. This is the input to our factorization method.

3.1. The Rank Theorem. We now analyze the relation between camera motion, shape, and the entries of the registered measurement matrix $\tilde{W}$. This analysis leads to the key result that $\tilde{W}$ is highly rank-deficient.

Referring to Fig. 1, suppose we place the origin of the world reference system x–y–z at the centroid of the P points $\{\mathbf{s}_p = (x_p, y_p, z_p)^T, p = 1, \ldots, P\}$, in space that corresponds to the P feature points tracked in the image system. The orientation of the camera reference system corresponding to frame number f is determined by a pair of unit vectors, $\mathbf{i}_f$ and $\mathbf{j}_f$, pointing along the scan lines and the columns of the image, respectively, and defined with respect to the world reference system. Under orthography, all projection rays are then parallel to the cross product of $\mathbf{i}_f$ and $\mathbf{j}_f$:

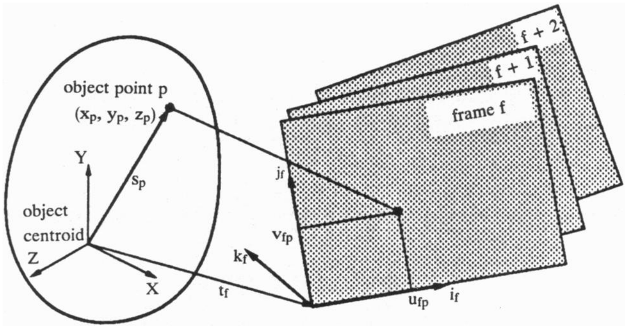

FIG. 1. Systems of reference used in our problem formulation.

$$\mathbf{k}_f = \mathbf{i}_f \times \mathbf{j}_f.$$

From Fig. 1, we see that the projection (u_{fp}, v_{fp})—i.e., the image feature position—of point $\mathbf{s}_p = (x_p, y_p, z_p)^T$ onto frame f is given by the equations

$$u_{fp} = \mathbf{i}_f^T(\mathbf{s}_p - \mathbf{t}_f) \qquad v_{fp} = \mathbf{j}_f^T(\mathbf{s}_p - \mathbf{t}_f),$$

where $\mathbf{t}_f = (a_f, b_f, c_f)^T$ is the vector from the world origin to the origin of image frame f. Here note that since the origin of the world coordinates is placed at the centroid of object points, $\frac{1}{P} \sum_{p=1}^{P} \mathbf{s}_p = 0$.

We can now write expressions for the entries $\tilde{u}_{fp}$ and $\tilde{v}_{fp}$ defined in Eqs. **3.1** of the registered measurement matrix. For the registered horizontal image projection we have

$$\tilde{u}_{fp} = u_{fp} - a_f = \mathbf{i}_f^T(\mathbf{s}_p - \mathbf{t}_f) - \frac{1}{P} \sum_{q=1}^{P} \mathbf{i}_f^T(\mathbf{s}_q - \mathbf{t}_f)$$

$$= \mathbf{i}_f^T\left(\mathbf{s}_p - \frac{1}{P} \sum_{q=1}^{P} \mathbf{s}_q\right) = \mathbf{i}_f^T \mathbf{s}_p. \qquad [3.2]$$

We can write a similar equation for $\tilde{v}_{fp}$. To summarize,

$$\tilde{u}_{fp} = \mathbf{i}_f^T \mathbf{s}_p \qquad \tilde{v}_{fp} = \mathbf{j}_f^T \mathbf{s}_p. \qquad [3.3]$$

Because of the two sets of $F \times P$ equations (3.3), the registered measurement matrix $\tilde{W}$ can be expressed in a matrix form:

$$\tilde{W} = RS, \qquad [3.4]$$

where

$$R = \begin{bmatrix} \mathbf{i}_1^T \\ \vdots \\ \mathbf{i}_F^T \\ \mathbf{j}_1^T \\ \vdots \\ \mathbf{j}_F^T \end{bmatrix} \qquad [3.5]$$

represents the camera rotation, and

$$S = [\mathbf{s}_1 \cdots \mathbf{s}_P] \qquad [3.6]$$

is the shape matrix. The rows of R represent the orientations of the horizontal and vertical camera reference axes throughout the stream, while the columns of S are the coordinates of the P feature points with respect to their centroid.

[§] Tomasi, C. & Kanade, T., Proceedings of the Third International Conference in Computer Vision (ICCV), December 1990, Osaka.

[¶] Tomasi, C. & Kanade, T., Proceedings of the DARPA Image Understanding Workshop, September 1990, Pittsburgh, pp. 258–270.

[||] Debrunner, Christian, H. & Ahuja, N., Proceedings of the 10th International Conference on Pattern Recognition, June 1990, Atlantic City, NJ, pp. 384–389.

[**] Boult, T. E. & Brown, L. G. Proceedings of the IEEE Workshop on Visual Motion, October 1991, pp. 179–186.

Since $\boldsymbol{R}$ is $2F \times 3$ and $\boldsymbol{S}$ is $3 \times P$, Eq. **3.4** implies the following.

RANK THEOREM. *Without noise, the registered measurement matrix* $\tilde{\mathbf{W}}$ *is at most of rank 3.*

The rank theorem expresses the fact that the $2F \times P$ image measurements are highly redundant. Indeed, they could all be described concisely by giving F frame reference systems and P point coordinate vectors, if only these were known.

From the first and the last line of Eq. **3.2**, the original unregistered matrix $\boldsymbol{W}$ can be written as

$$\boldsymbol{W} = \boldsymbol{RS} + \mathbf{t}\mathbf{e}_P^T, \qquad [3.7]$$

where $\mathbf{t} = (a_1, \ldots, a_F, b_1, \ldots, b_F)^T$ is a $2F$-dimensional vector that collects the projections of camera translation along the image plane (see Eq. **3.2**), and $\mathbf{e}_P^T = (1, \ldots, 1)$ is a vector of P ones. In scalar form,

$$u_{fp} = \mathbf{i}_f^T\mathbf{s}_p + a_f \qquad v_{fp} = \mathbf{j}_f^T\mathbf{s}_p + b_f. \qquad [3.8]$$

Comparing with Eqs. **3.1**, we see that the two components of camera translation along the image plane are simply the averages of the rows of $\boldsymbol{W}$.

In the equations above, $\mathbf{i}_f$ and $\mathbf{j}_f$ are mutually orthogonal unit vectors, so they must satisfy the constraints

$$|\mathbf{i}_f| = |\mathbf{j}_f| = 1 \quad \text{and} \quad \mathbf{i}_f^T\mathbf{j}_f = 0. \qquad [3.9]$$

Also, the rotation matrix $\boldsymbol{R}$ is unique if the system of reference for the solution is aligned, say, with that of the first camera position, so that

$$\mathbf{i}_1 = (1, 0, 0)^T \quad \text{and} \quad \mathbf{j}_1 = (0, 1, 0)^T. \qquad [3.10]$$

The registered measurement matrix $\tilde{\boldsymbol{W}}$ must be at most of rank 3 without noise. When noise corrupts the images, however, $\tilde{\boldsymbol{W}}$ will not be exactly of rank 3. However, the rank theorem can be extended to the case of noisy measurements in a well-defined manner. The next section introduces the notion of approximate rank, using the concept of singular value decomposition (15).

3.2. Approximate Rank. Assuming[††] that $2F \geq P$, the matrix $\tilde{\boldsymbol{W}}$ can be decomposed (15) into a $2F \times P$ matrix $\boldsymbol{O}_1$, a diagonal $P \times P$ matrix Σ, and a $P \times P$ matrix $\boldsymbol{O}_2$,

$$\tilde{\boldsymbol{W}} = \boldsymbol{O}_1\Sigma\boldsymbol{O}_2, \qquad [3.11]$$

such that $\boldsymbol{O}_1^T\boldsymbol{O}_1 = \boldsymbol{O}_2^T\boldsymbol{O}_2 = \boldsymbol{O}_2\boldsymbol{O}_2^T = \mathcal{I}$, where $\mathcal{I}$ is the $P \times P$ identity matrix. Σ is a diagonal matrix whose diagonal entries are the *singular values* $\sigma_1 \geq \ldots \geq \sigma_P$ sorted in nondecreasing order. This is the *singular value decomposition* of the matrix $\tilde{\boldsymbol{W}}$.

If we pay attention only to the first three columns of $\boldsymbol{O}_1$, the first 3×3 submatrix of Σ, and the first three rows of $\boldsymbol{O}_2$. If we partition the matrices $\boldsymbol{O}_1$, Σ, and $\boldsymbol{O}_2$ as follows

$$\boldsymbol{O}_1 = [\underbrace{\boldsymbol{O}_1'}_{3}|\underbrace{\boldsymbol{O}_1''}_{P-3}]\ \}2F \qquad \Sigma = \underbrace{\left[\begin{array}{c|c} \Sigma' & 0 \\ \hline 0 & \Sigma'' \end{array}\right]}_{3\quad P-3} \begin{array}{l} \}3 \\ \}P-3 \end{array} \qquad [3.12]$$

$$\boldsymbol{O}_2 = \underbrace{\left[\begin{array}{c} \boldsymbol{O}_2' \\ \hline \boldsymbol{O}_2'' \end{array}\right]}_{P} \begin{array}{l} \}3 \\ \}P-3 \end{array},$$

we have $\boldsymbol{O}_1\Sigma\boldsymbol{O}_2 = \boldsymbol{O}_1'\Sigma'\boldsymbol{O}_2' + \boldsymbol{O}_1''\Sigma''\boldsymbol{O}_2''$.

Let $\tilde{\boldsymbol{W}}^*$ be the ideal registered measurement matrix—that is, the matrix we would obtain in the absence of noise. Because of the rank theorem, $\tilde{\boldsymbol{W}}^*$ has at most three nonzero singular values. Since the singular values in Σ are sorted in nonincreasing order, Σ' must contain all the singular values of $\tilde{\boldsymbol{W}}^*$ that exceed the noise level. As a consequence, the term $\boldsymbol{O}_1''\Sigma''\boldsymbol{O}_2''$ must be due entirely to noise, and the best possible rank-3 approximation to the ideal registered measurement matrix $\tilde{\boldsymbol{W}}^*$ is the product $\hat{\boldsymbol{W}} = \boldsymbol{O}_1'\Sigma'\boldsymbol{O}_2'$. We can now restate our rank theorem for the case of noisy measurements.

RANK THEOREM FOR NOISY MEASUREMENTS. *All the shape and rotation information in* $\tilde{\mathbf{W}}$ *is contained in its three greatest singular values, together with the corresponding left and right eigenvectors.*

Now if we define $\hat{\boldsymbol{R}} = \boldsymbol{O}_1'[\Sigma']^{1/2}$ and $\hat{\boldsymbol{S}} = [\Sigma']^{1/2}\boldsymbol{O}_2'$, we can write

$$\hat{\boldsymbol{W}} = \hat{\boldsymbol{R}}\hat{\boldsymbol{S}}. \qquad [3.13]$$

The two matrices $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$ are of the same size as the desired rotation and shape matrices $\boldsymbol{R}$ and $\boldsymbol{S}$: $\hat{\boldsymbol{R}}$ is $2F \times 3$, and $\hat{\boldsymbol{S}}$ is $3 \times P$. However, the decomposition (Eq. **3.13**) is not unique. In fact, if $\boldsymbol{Q}$ is *any* invertible 3×3 matrix, the matrices $\hat{\boldsymbol{R}}\boldsymbol{Q}$ and $\boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}$ are also a valid decomposition of $\hat{\boldsymbol{W}}$, since

$$(\hat{\boldsymbol{R}}\boldsymbol{Q})(\boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}) = \hat{\boldsymbol{R}}(\boldsymbol{Q}\boldsymbol{Q}^{-1})\hat{\boldsymbol{S}} = \hat{\boldsymbol{R}}\hat{\boldsymbol{S}} = \hat{\boldsymbol{W}}.$$

Thus, $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$ are in general different from $\boldsymbol{R}$ and $\boldsymbol{S}$. A striking fact, however, is that except for noise the matrix $\hat{\boldsymbol{R}}$ is a linear transformation of the true rotation matrix $\boldsymbol{R}$, and the matrix $\hat{\boldsymbol{S}}$ is a linear transformation of the true shape matrix $\boldsymbol{S}$. Indeed, in the absence of noise, $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$ both span the column space of the registered measurement matrix $\tilde{\boldsymbol{W}} = \tilde{\boldsymbol{W}}^* = \tilde{\boldsymbol{W}}$. Since that column space is three-dimensional because of the rank theorem, $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$ are different bases for the same space, and there must be a linear transformation between them.

Whether the noise level is low enough that it can be ignored at this juncture depends also on the camera motion and on shape. Notice, however, that the singular value decomposition yields sufficient information to make this decision: the requirement is that the ratio between the third and the fourth largest singular values of $\tilde{\boldsymbol{W}}$ be sufficiently large.

3.3. The Metric Constraints. We have found that the matrix $\hat{\boldsymbol{R}}$ is a linear transformation of the true rotation matrix $\boldsymbol{R}$. Likewise, $\hat{\boldsymbol{S}}$ is a linear transformation of the true shape matrix $\boldsymbol{S}$. More specifically, there exists a 3×3 matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{R} = \hat{\boldsymbol{R}}\boldsymbol{Q} \qquad \boldsymbol{S} = \boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}. \qquad [3.14]$$

To find $\boldsymbol{Q}$ we observe that the rows of the true rotation matrix $\boldsymbol{R}$ are unit vectors and the first F are orthogonal to corresponding F in the second half of $\boldsymbol{R}$. These *metric constraints* yield the over-constrained, quadratic system

$$\hat{\mathbf{i}}_f^T\boldsymbol{Q}\boldsymbol{Q}^T\hat{\mathbf{i}}_f = 1 \qquad \hat{\mathbf{j}}_f^T\boldsymbol{Q}\boldsymbol{Q}^T\hat{\mathbf{j}}_f = 1 \qquad \hat{\mathbf{i}}_f^T\boldsymbol{Q}\boldsymbol{Q}^T\hat{\mathbf{j}}_f = 0 \qquad [3.15]$$

in the entries of $\boldsymbol{Q}$. This is a simple data-fitting problem which, though nonlinear, can be solved efficiently and reliably. Its solution is determined up to a rotation of the whole reference system, since the orientation of the world reference system was arbitrary. This arbitrariness can be removed by enforcing the constraints (Eq. **3.10**)—that is, selecting the x–y axes of the world reference system to be parallel with those of the first frame.

3.4. Outline of the Complete Algorithm. Based on the development in the previous sections, we now have a complete algorithm for the factorization of the registered measurement matrix $\tilde{\boldsymbol{W}}$ derived from a stream of images into shape $\boldsymbol{S}$ and rotation $\boldsymbol{R}$ as defined in Eqs. **3.4–3.6**:

[††] This assumption is not crucial: if $2F < P$, everything can be repeated for the transpose of $\tilde{\boldsymbol{W}}$.

(*i*) Compute the singular-value decomposition $\boldsymbol{W} = \boldsymbol{O}_1\Sigma\boldsymbol{O}_2$. (*ii*) Define $\hat{\boldsymbol{R}} = \boldsymbol{O}_1'(\Sigma')^{1/2}$ and $\hat{\boldsymbol{S}} = (\Sigma')^{1/2}\boldsymbol{O}_2'$, where the primes refer to the block partitioning defined in Eqs. **3.12**. (*iii*) Compute the matrix $\boldsymbol{Q}$ in Eqs. **3.14** by imposing the metric constraints (Eqs. **3.15**). (*iv*) Compute the rotation matrix $\boldsymbol{R}$ and the shape matrix $\boldsymbol{S}$ as $\boldsymbol{R} = \hat{\boldsymbol{R}}\boldsymbol{Q}$ and $\boldsymbol{S} = \boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}$. (*v*) If desired, align the first camera reference system with the world reference system by forming the products $\boldsymbol{R}\boldsymbol{R}_0$ and $\boldsymbol{R}_0^T\boldsymbol{S}$, where the orthonormal matrix $\boldsymbol{R}_0 = [\mathbf{i}_1\ \mathbf{j}_1\ \mathbf{k}_1]$ rotates the first camera reference system into the identity matrix.

Section 4. Experiment

We test the factorization method with two real streams of images: one taken in a controlled laboratory environment with ground-truth motion data and the other in an outdoor environment with a hand-held camcorder.

4.1. "Hotel" Image Stream in a Laboratory. Some frames in this stream are shown in Fig. 2 *Upper*. The images depict a small plastic model of a building. The camera is a Sony charged coupled device camera with a 200-mm lens and is moved by means of a high-precision positioning platform. Camera pitch, yaw, and roll around the model are all varied as shown by the dotted curves in Fig. 2 *Lower*. The translation of the camera is such as to keep the building within the field of view of the camera.

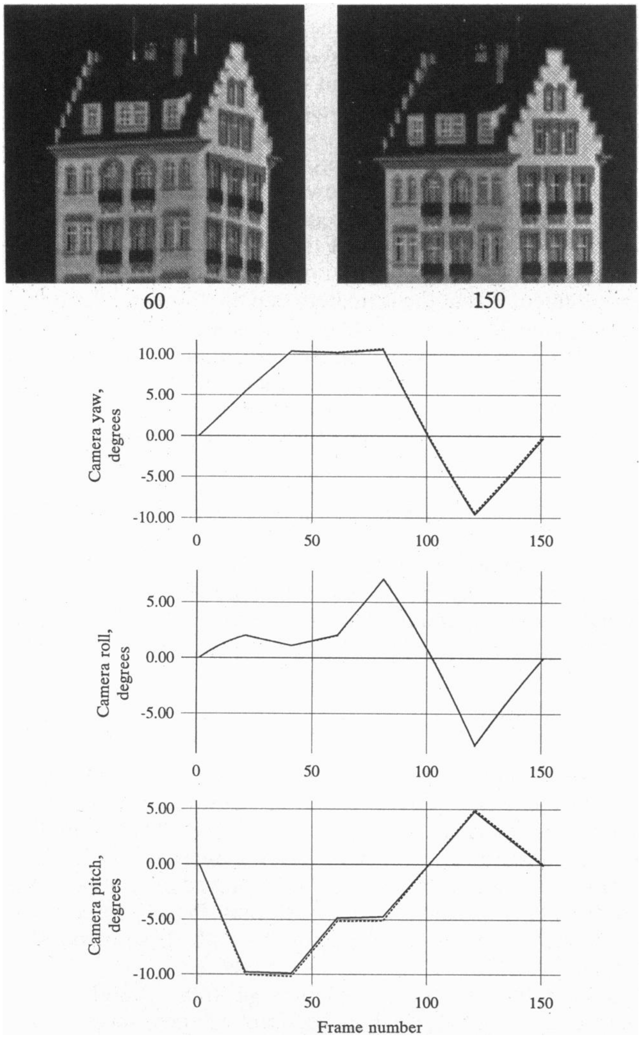

FIG. 2. (*Upper*) Two frames in the sequence. The whole sequence is 150 frames. (*Lower*) True (· · ·) and computed (—) camera yaw, roll, and pitch.

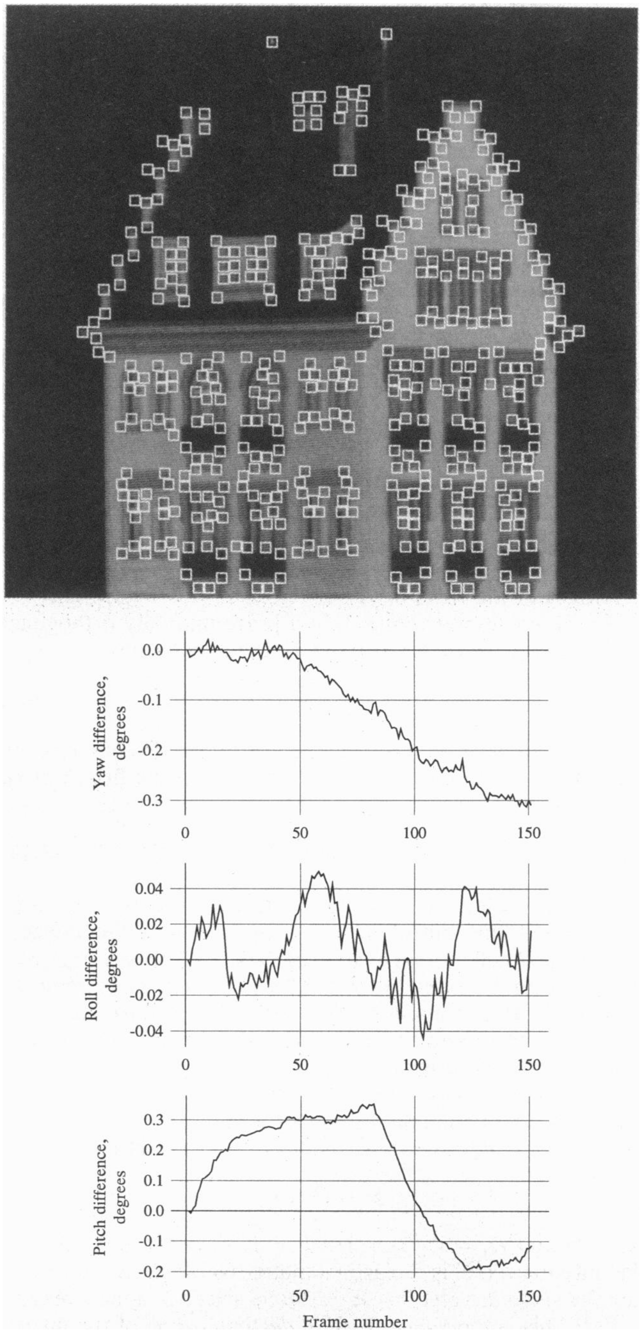

FIG. 3. (*Upper*) The 430 features selected by the automatic detection method. (*Lower*) Blowup of the errors in Fig. 2 *Lower*.

For feature tracking, we extended the Lucas–Kanade method[‡‡] to allow also for the automatic selection of image features. The Lucas–Kanade method of tracking obtains the displacement vector of the window around a feature as the solution of a linear 2 × 2 equation system. As good image features, we select those points for which the above equation systems are stable. The details are given in refs. 16 and 17.

The entire set of 430 features thus selected is displayed in Fig. 3 *Upper*, overlaid on the first frame of the stream. Of these features, 42 were abandoned during tracking because their appearance changed too much. The trajectories of the remaining 388 features are used as the measurement matrix for the computation of shape and motion.

[‡‡] Lucas, B. D. & Kanade, T., Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

The motion recovery is precise. The plots in Fig. 2 *Lower* compare the rotation components computed by the factorization method (solid curves) with the values measured mechanically from the mobile platform (dotted curves). The differences are magnified in Fig. 3 *Lower*. The errors are everywhere less than 0.4° and on average 0.2°. The computed motion also follows closely rotations with curved profiles, such as the roll profile between frames 1 and 20 (second plot in Fig. 2 *Lower*) and faithfully preserves all discontinuities in the rotational velocities: the factorization method does not smooth the results.

Between frames 60 and 80, yaw and pitch are nearly constant, and the camera merely rotates about its optical axis—i.e., the motion is actually degenerate during this period, but still it has been correctly recovered. This demonstrates that the factorization method can deal without difficulty with streams that contain degenerate substreams, because the information in the stream is used *as a whole*.

The shape results are evaluated qualitatively in Fig. 4 *Upper*, which shows the computed shape viewed from above. The view in Fig. 4 *Upper* is similar to that in Fig. 4 *Lower*, included for visual comparison. The walls, the windows on the roof, and the chimneys are recovered in their correct positions.

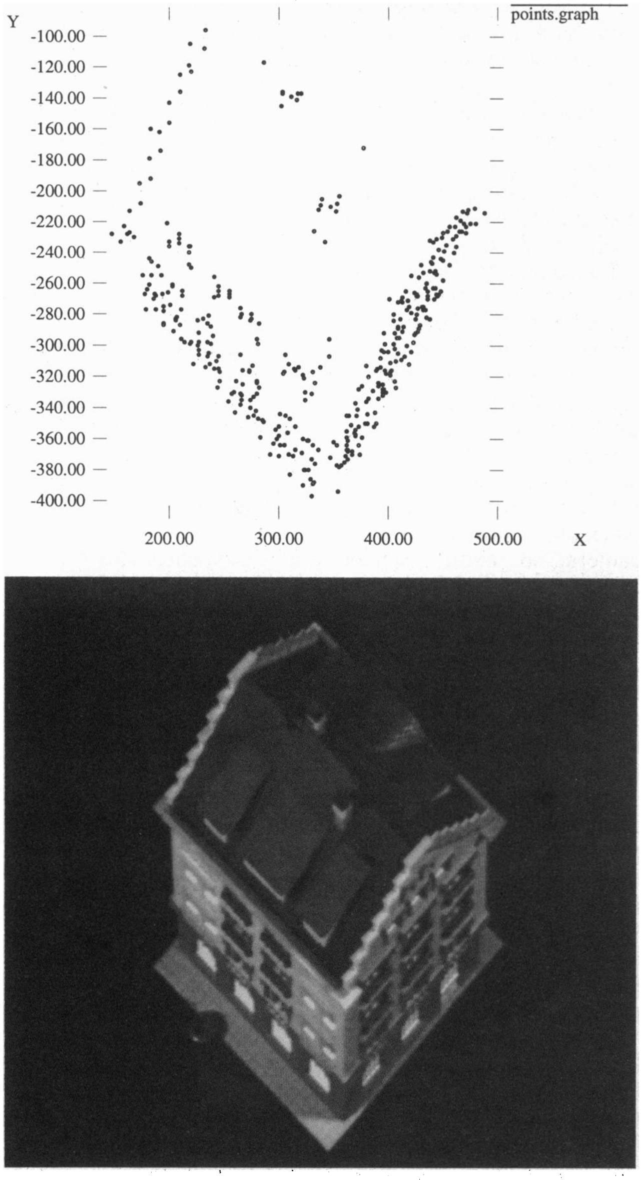

FIG. 4. (*Upper*) View of the computed shape from approximately above the building. (*Lower*) Real picture from above the building.

To evaluate the shape performance quantitatively, we measured some distances on the actual house model with a ruler and compared them with the distances computed from the point coordinates in the shape results. The measured distances between the steps along the right side of the roof (7.2 mm) were obtained by measuring five steps and dividing the total distance (36 mm) by five. The differences between computed and measured results are of the order of the resolution of our ruler measurements (1 mm).

Part of the error in the results is due to the use of orthography as the projection model. However, it tends to be fairly small for many realistic situations. In fact, it has been shown that errors due to the orthographic distortion are about the same percentage as the ratio of the object size in depth to the distance of the object from the camera (16).

4.2. Outdoor "House" Image Stream. The factorization method has been tested with an image stream of a real building, taken with a hand-held camera. Fig. 5 shows some of the 180 frames of the building stream. The overall motion covers a relatively small rotation angle, ≈15°. Outdoor images are harder to process than those produced in a controlled environment of the laboratory, because lighting changes less predictably and the motion of the camera is more difficult to control. As a consequence, features are harder to track: the images are unpredictably blurred by motion and corrupted by vibrations of the video recorder's head, both during recording and digitization. Furthermore, the camera's jumps and jerks produce a wide range of image disparities.

The features found by the selection algorithm in the first frame are shown in Fig. 6 *Upper*. There are many false features. The reflections in the window partially visible in the top left of the image move nonrigidly. More false features can be found in the lower left corner of the picture, where the vertical bars of the handrail intersect the horizontal edges of the bricks of the wall behind. We masked away these two parts of the image from the analysis.

In total, 376 features were found by the selection algorithm and tracked. Fig. 6 *Lower* plots the tracks of 60 of the features for illustration. Notice the very jagged trajectories due to the vibrating motion of the hand-held camera.

Fig. 7 shows a front (*Upper*) and a top (*Lower*) view of the building as reconstructed by the factorization method. To render these figures for display, we triangulated the computed three-dimensional points into a set of small surface patches and mapped the pixel values in the first frame onto the resulting surface. The structure of the visible part of the building's three walls has clearly been reconstructed. The left wall appears to bend somewhat on the right where it intersects the middle wall. This occurred because the feature selector found features along the shadow of the roof just on the right of the intersection of the two walls, rather than at the intersection itself. Thus, the appearance of a bending wall is an artifact of the triangulation done for rendering.

This experiment with an image stream taken outdoors with the jerky motion produced by a hand-held camera demon-

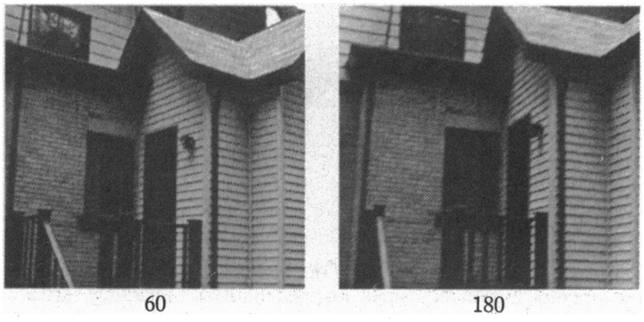

FIG. 5. Two of the 180 frames of the real house image stream.

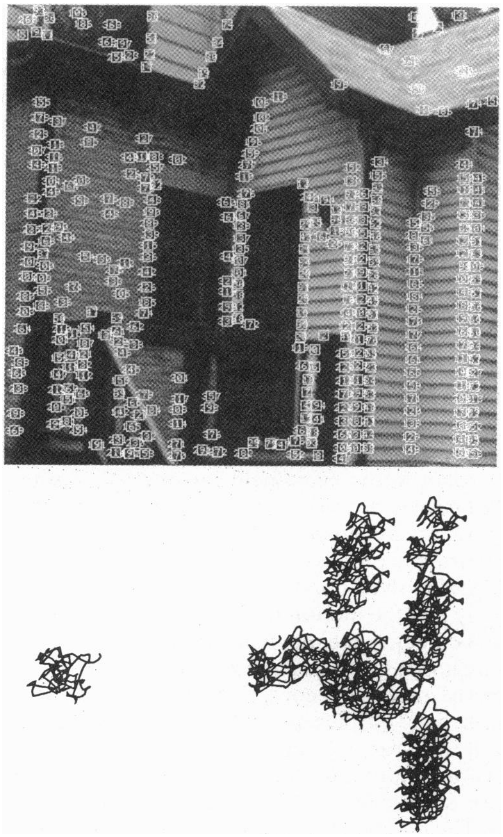

FIG. 6. (*Upper*) Features selected in the first frame of the real house stream (Fig. 5). (*Lower*) Tracks of 60 randomly selected features from the real house stream (Fig. 5).

strates that the factorization method does not require a smooth motion assumption. The identification of false features—that is, of features that do not move rigidly with respect to the environment—remains an open problem that must be solved for a fully autonomous system. An initial effort has been seen.**

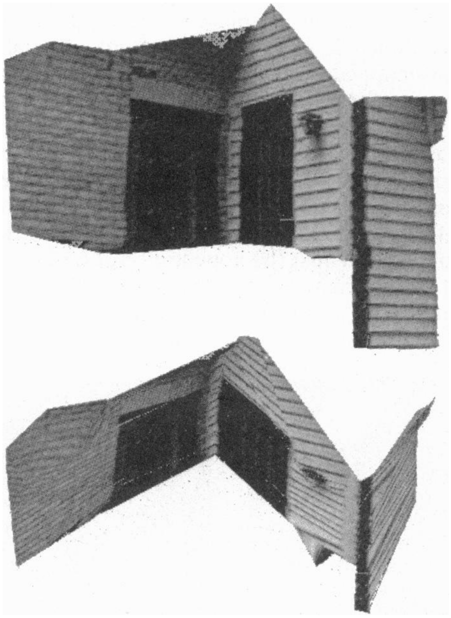

FIG. 7. (*Upper*) Front view of the three reconstructed walls, with the original image intensities mapped onto the resulting surface. (*Lower*) View from above of the three reconstructed walls, with image intensities mapped onto the surface.

Section 5. Occlusions

In reality, as the camera moves, features can appear and disappear from the image, because of occlusions. Also, a feature-tracking method will not always succeed in tracking features throughout the image system. These phenomena are frequent enough to make a shape and motion computation method unrealistic if it cannot deal with them.

Sequences with appearing and disappearing features result in a measurement matrix W that is only partially filled in. The factorization method introduced in *Section 3* cannot be applied directly. However, there is usually sufficient information in the stream to determine all the camera positions and all the three-dimensional feature point coordinates. If that is the case, not only can we solve the shape and motion recovery problem from the incomplete measurement matrix W, but we can even hallucinate the unknown entires of W by projecting the computed three-dimensional feature coordinates onto the computed camera positions.

Suppose that a feature point is not visible in a certain frame. If the same feature is seen often enough in other frames, its position in space should be recoverable. Moreover, if the frame in question includes enough other features, the corresponding camera position should be recoverable as well. Then with point and camera positions thus recovered, we should also be able to reconstruct the missing image measurement. Formally, we have the following sufficient condition.

Condition for Reconstruction: In the absence of noise, an unknown image measurement pair (u_{fp}, v_{fp}) in frame f can be reconstructed if point p is visible in at least three more frames f_1, f_2, f_3 and if there are at least three more points p_1, p_2, p_3 that are visible in all the four frames: the original f and the additional f_1, f_2, f_3.

Based on this, we have developed an algorithm to recover the three-dimensional shape of a scene that is partially occluded in the input image sequence. The details are presented in Tomasi and Kanade (18). The following are examples of processing results with image streams that include substantial occlusions. We first use an image stream taken in a laboratory. Then, we demonstrate the robustness of the factorization method with another stream taken with a hand-held amateur camera.

5.1. "Ping-Pong Ball" Image Stream. A ping-pong ball with black dots marked on its surface is rotated 450° in front of the camera, so features appear and disappear. The rotation between adjacent frames is 2°, so the stream is 226 frames long. Fig. 8 *Upper* shows the first frame of the stream, with the automatically selected features overlaid.

Every 30 frames (60°) of rotation, the feature tracker looks for new features. In this way, features that disappear on one side around the ball are replaced by new ones that appear on the other side. Fig. 8 *Lower* shows the tracks of 60 features, randomly chosen among the total 829 found by the selector.

If all measurements are collected into the noisy measurement matrix W, the U and V parts of W have the same fill pattern: if the x coordinate of a measurement is known, so is the y coordinate. Fig. 9 *Upper* shows this *fill matrix* for our experiment. This matrix has the same size as either U or V (that is $F \times P$). A column corresponds to a feature point, and a row corresponds to a frame. Shaded regions denote known entries. The fill matrix shown has $226 \times 829 = 187{,}354$ entries, of which 30,185 (about 16%) are known.

To start the motion and shape computation, the algorithm finds a large full submatrix by applying simple heuristics based on typical patterns of the fill matrix. The choice of the starting matrix is not critical, as long as it leads to a reliable initialization of the motion and shape matrices. The initial solution is then grown by repeatedly solving overconstrained versions of the linear system to add new rows and new columns. The rows and columns to add are selected so as to

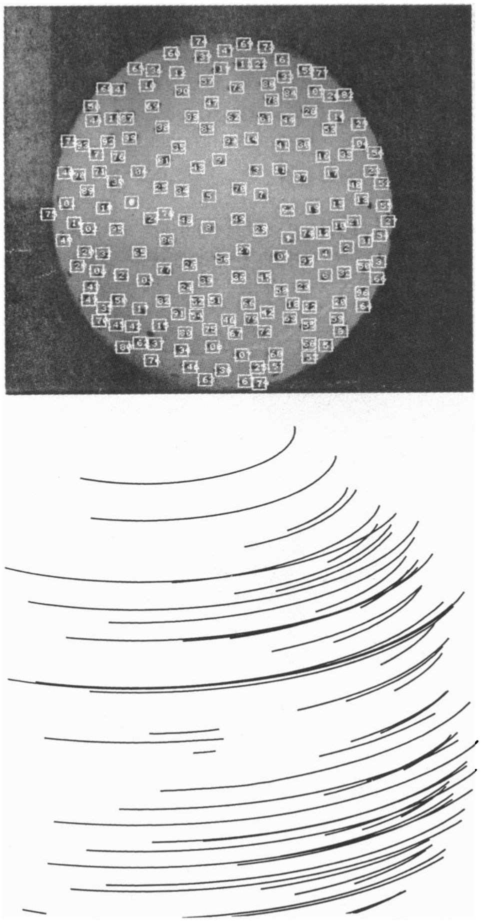

FIG. 8. (*Upper*) First frame of the ping-pong stream, with overlaid features. (*Lower*) Tracks of 60 randomly selected features from the stream in *Upper*.

maximize the redundancy of the linear systems. Eventually, all of the motion and shape values are determined. As a result, the unknown 84% of the measurement matrix can be hallucinated from the known 16%.

Fig. 9 *Lower* shows two views of the final shape results, taken from the top and from the side. The missing features at the bottom of the ball in the side view correspond to the part of the ball that remained always invisible, because it rested on the rotating platform.

To display the motion results, we look at the $\mathbf{i}_f$ and $\mathbf{j}_f$ vectors directly. We recall that these unit vectors point along the rows and columns of the image frames f in $1, \ldots, F$. Because the ping-pong ball rotates around a fixed axis, both $\mathbf{i}_f$ and $\mathbf{j}_f$ should sweep a cone in space, as shown in Fig. 10 *Upper*. The tips of $\mathbf{i}_f$ and $\mathbf{j}_f$ should describe two circles in space, centered along the axis of rotation. Fig. 10 *Lower* shows two views of these vector tips, from the top and from the side. Those trajectories indicate that the motion recovery was done correctly. Notice the double arc in the top part of Fig. 10 *Lower* corresponding to more than 360° of rotation. If the motion reconstruction were perfect, the two arcs would be indistinguishable.

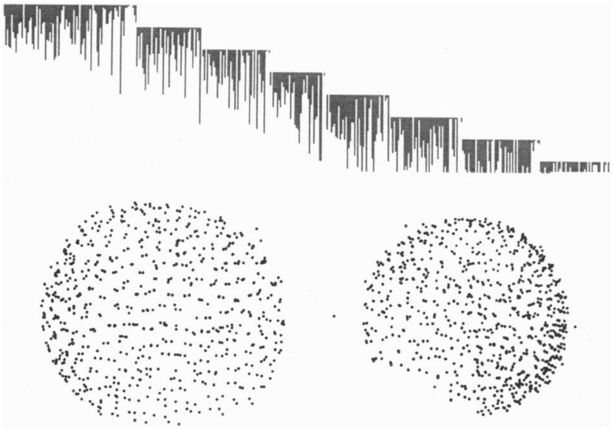

FIG. 9. (*Upper*) Fill matrix for the ping-pong ball experiment. Shaded entries are known. (*Lower*) Top and side views of the reconstructed ping-pong ball.

5.2. "Cup and Hand" Image Stream. We describe an experiment with a natural scene including occlusion as a dominant phenomenon. A hand holds a cup and rotates it by about 90° in front of the camera mounted on a fixed stand. Fig. 11 *Top* shows two of the 240 frames of the stream.

An additional need in this experiment is figure/ground segmentation. Since the camera was fixed, however, this problem is easily solved: features that do not move belong to the background. Also, the stream includes some nonrigid motion: as the hand turns, the configuration and relative position of the fingers changes slightly. This effect, however, was small and did not affect the results appreciably.

A total of 207 features were selected. Occlusions were marked by hand in this experiment. The fill matrix of Fig. 11 *Bottom* illustrates the occlusion pattern. Fig. 11 *Middle* shows the image trajectory of 60 randomly selected features.

Fig. 12 shows a front (*Upper*) and a top (*Lower*) view of the cup and the visible fingers as reconstructed by the propagation method. The shape of the cup was recovered, as well as the rough shape of the fingers. These renderings were obtained, as for the house image stream in *Section 4.1*, by triangulating the tracked feature points and mapping pixel values onto the resulting surface.

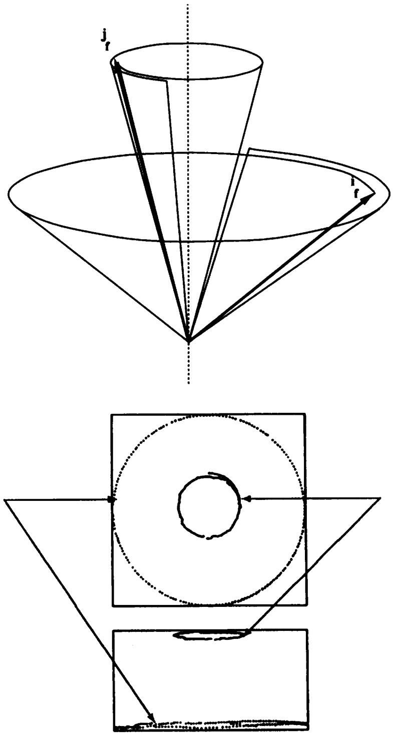

FIG. 10. (*Upper*) Rotational component of the camera motion for the ping-pong stream. Because rotation occurs around a fixed axis, the two mutually orthogonal unit vectors $\mathbf{i}_f$ and $\mathbf{j}_f$, pointing along rows and columns of the image sensor, sweep two 450° cones in space. (*Lower*) Top and side views of the $\mathbf{i}_f$ and $\mathbf{j}_f$ vectors identifying the camera rotation.

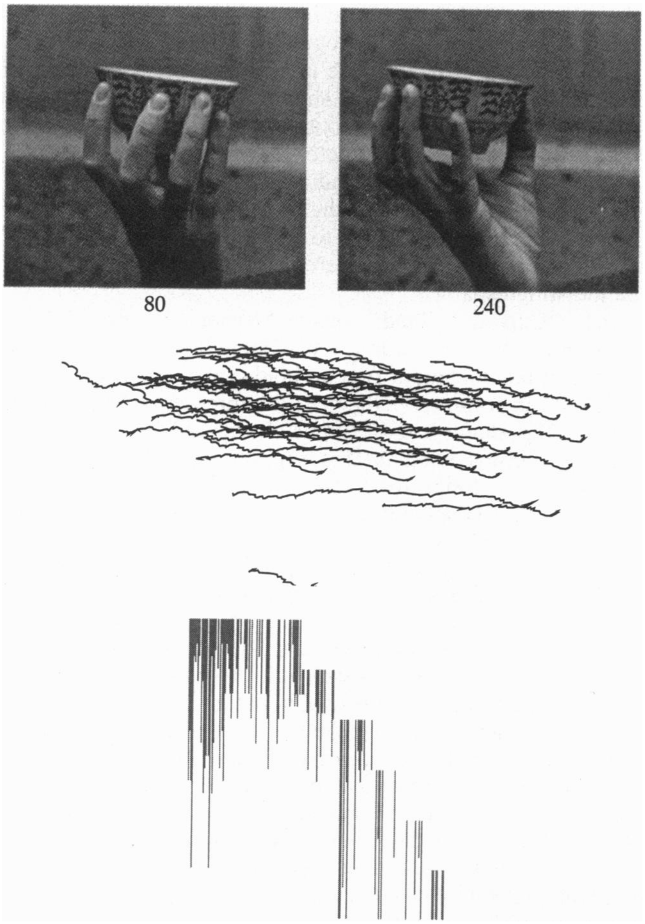

FIG. 11. (*Top*) Two of the 240 frames of the cup image stream. (*Middle*) Tracks of 60 randomly selected features from the cup stream. (*Bottom*) The 240 × 207 fill matrix for the cup stream (*Top*). Shaded entries are known.

Section 6. Conclusion

The rank theorem, which is the basis of the factorization method, is both surprising and powerful. It is surprising because it states that the correlation among measurements made in an image stream has a simple expression *no matter what the camera motion is and no matter what the shape of an object is*, thus making motion or surface assumptions (such as smooth, constant, linear, planar, and quadratic) fundamentally superfluous. It is powerful because the rank theorem leads to a factorization of the measurement matrix into shape and motion in a well-behaved and stable manner.

The factorization method exploits the redundancy of the measurement matrix to counter the noise sensitivity of structure from motion and allows very short interframe camera motion to simplify feature tracking. The structural insight into shape from motion afforded by the rank theorem led to a systematic procedure to solve the occlusion problem within the factorization method. The experiments in the lab demonstrate the high accuracy of the method, and the outdoor experiments show its robustness.

The rank theorem is strongly related to Ullman's 12-year-old result that three pictures of four points determine structure and motion under orthography. Thus, in a sense, the theoretical foundation of our result has been around for a long time. The factorization method evolves the applicability of that foundation from mathematical images to actual noisy image streams.

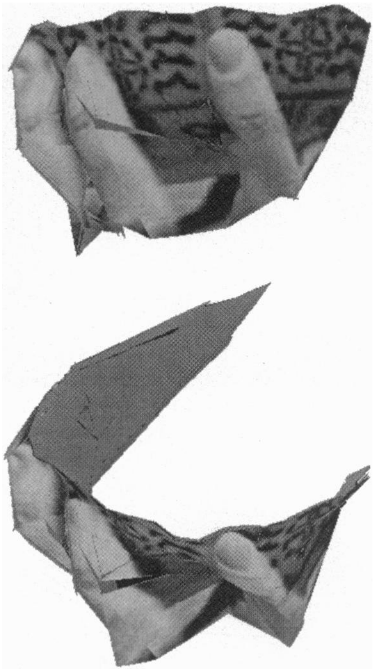

FIG. 12. (*Upper*) Front view of the cup and fingers, with the original image intensities mapped onto the resulting surface. (*Lower*) View from above of the cup and fingers with image intensities mapped onto the surface.

This research was sponsored by the Avionics Laboratory, Wright Research and Development Center, Aeronautical Systems Division (AFSC), U.S. Air Force, Wright-Patterson Air Force Base, Ohio, under Contract F33615-90-C-1465, Advanced Research Projects Agency Order 7597.

1. Ullman, S. (1979) *The Interpretation of Visual Motion* (Massachusetts Inst. of Technol. Press, Cambridge, MA).
2. Roach, J. W. & Aggarwal, J. K. (1979) *IEEE Trans. Pattern Anal. Machine Intelligence* **PAMI-1(2),** 127–135.
3. Adiv, G. (1985) *IEEE Pattern Anal. Machine Intelligence* **7,** 384–401.
4. Bolles, R. C., Baker, H. H. & Marimont, D. H. (1987) *Int. J. Comput. Vis.* **1(1),** 7–55.
5. Broida, T. J., Chandrashekhar, S. & Chellappa, R. (1990) *IEEETrans. Aerosp. Electron. Syst.* **26(4),** 639–656.
6. Bruss, A. R. & Horn, B. K. P. (1983) *Comput. Vis. Graphics Image Proc.* **21,** 3–20.
7. Heeger, D. J. & Jepson, A. (1989) *Visual Perception of Three-Dimensional Motion* (Massachusetts Inst. Technol. Media Laboratory, Cambridge, MA), Tech. Rep. No. 124.
8. Horn, B. K. P., Hilden, H. M. & Negahdaripour, S. (1988) *J. Opt. Soc. Am. A Opt. Image Sci.* **5(7),** 1125–1135.
9. Matthies, L., Kanade, T. & Szeliski, R. (1989) *Int. J. Comput. Vis.* **3(3),** 209–236.
10. Prazdny, K. (1980) *Biol. Cybernetics* **102,** 87–102.
11. Tsai, R. Y. & Huang, T. S. (1984) *IEEE Trans. Pattern Anal. Machine Intelligence* **PAMI-6(1),** 13–27.
12. Waxman, A. M. & Wohn, K. (1985) *Int. J. Robot. Res.* **4,** 95–108.
13. Tomasi, C. & Kanade, T. (1991) *Shape and Motion from Image Streams: A Factorization Method: 2. Point Features in 3d Motion* (Carnegie Mellon Univ., Pittsburgh), Tech. Rep. CMU-CS-91-105.
14. Debrunner, C. H. & Ahuja, N. (1991) *Motion and Structure Factorization and Segmentation of Long Multiple Motion Image Sequences* (Univ. of Illinois, Urbana–Champaign), Tech. Rep. UI-BI-CV-5-91.
15. Golub, G. H. & Reinsch, C. (1971) in *Singular Value Decomposition and Least Squares Solutions* (Springer, New York), Vol. 2, pp. 134–151.
16. Tomasi, C. (1991) PhD thesis (Carnegie Mellon University, Pittsburgh).
17. Tomasi, C. & Kanade, T. (1991) *Shape and Motion from Image Streams: A Factorization Method, Part 3: Detection and Tracking of Point Features* (Carnegie Mellon Univ., Pittsburgh), Tech. Rep. CMU-CS-91-132.
18. Tomasi, C. & Kanade, T. (1992) *Int. J. Computer Vision* **9,** 137–154.