International Journal of Computer Vision, 9:2, 137–154 (1992)

Shape and Motion from Image Streams under Orthography: a Factorization Method

CARLO TOMASI
Department of Computer Science, Cornell University, Ithaca, NY 14850

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

Received

Abstract

Inferring scene geometry and camera motion from a stream of images is possible in principle, but 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 under orthography 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. We show that under orthographic projection this matrix is of rank 3.

Based on 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 rotation respectively. Two of the three translation components are computed in a preprocessing stage. The method can also handle and obtain a full solution from a partially filled-in measurement matrix that may result from 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.

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 new method called the *factorization method* which can robustly recover shape and motion from a sequence of images under orthographic projection. The effects of camera translation along the optical axis are not accounted for by orthography. Consequently, this component of motion cannot be recovered by our method and must be small relative to the scene distance. However, this restriction to shallow motion improves dramatically the quality of the computed shape and of the remaining five motion parameters. We demonstrate this with a series of experiments on laboratory and outdoor sequences, with and without occlusions.

In the factorization method, we represent an image sequence as a $2F \times P$ measurement matrix $\mathbf{W}$, 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 matrixes $\mathbf{R}$ and $\mathbf{S}$. Here, $\mathbf{R}$ is a $2F \times 3$ matrix that represents camera rotation, and $\mathbf{S}$ 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 $\mathbf{W}$. When features appear and disappear in the image sequence because of occlusions or tracking failures, the resulting measurement matrix $\mathbf{W}$ 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 complete solution with an iterative procedure.

The rank theorem captures precisely 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 simple and the assumption of orthography easier to approximate.

2 Relation to Previous Work

In Ullman's original proof of existence of a solution (Ullman 1979) for the structure-from-motion problem, the coordinates of feature points in the world are expressed in a world-centered system of reference and an orthographic projection model is assumed. Since then, however, most computer vision researchers opted for perspective projection and a camera-centered representation of shape (Prazdny 1980; Bruss & Horn 1983; Tsai & Huang 1984; Adiv 1985; Waxman & Wohn 1985; Bolles et al. 1987; Horn et al 1988; Heeger & Jepson 1989; Heel 1989; Matthies et al. 1989; Spetsakis & Aloimonos 1989; Broida et al. 1990). With this representation, the position of feature points is specified by their image coordinates and by their depths, defined as the distance 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, a small rotation about the vertical axis and a small translation along the horizontal axis can generate very similar changes in an image. Any attempt to recover or differentiate between these two motions, though possible 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 often the case for interesting applications such as site modeling.

The factorizaiton method we present here takes advantage of the fact that both difficulties disappear when the problem is reformulated in world-centered coordinates and under orthography. This new (and 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 together with the linearity of orthographic projection makes it possible to cast the structure-from-motion problem as a factorization problem, in which a matrix representing image measurements is decomposed directly into camera motion and object shape.

We first introduced this factorization method in (Tomasi & Kanade 1990), where we treated the case of single-scanline images in a flat, two-dimensional world. In (Tomasi & Kanade 1991a) we presented the theory for the case of shallow camera motion in three dimensions and full two-dimensional images. Here, we extend the factorization method for dealing with feature occlusions as well as present experimental results with real-world images. Debrunner and Ahuja have pursued an approach related to ours, but using a different formalism (Debrunner & Ahuja 1992). 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 (Boult & Brown 1991), in which they count and segment separate motions in the field of view of the camera.

3 The Factorization Method

Suppose that we have tracked P feature points over F frames in an image stream. We then obtain trajectories of image coordinates $\{(u_{fp}, v_{fp}) \mid f = 1, \ldots, F, p = 1, \ldots, P\}$. We write the horizontal feature coordinates u_{fp} into an $F \times P$ matrix U with 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 $2F \times P$

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

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:

$$\begin{aligned} \tilde{u}_{fp} &= u_{fp} - a_f \\ \tilde{v}_{fp} &= v_{fp} - b_f \end{aligned} \tag{1}$$

where

$$a_f = \frac{1}{P}\sum_{p=1}^{P} u_{fp}$$

$$b_f = \frac{1}{P}\sum_{p=1}^{P} v_{fp}$$

This produces two new $F\times P$ matrixes $\tilde{\mathbf{U}} = [\tilde{u}_{fp}]$ and $\tilde{\mathbf{V}} = [\tilde{v}_{fp}]$. The matrix

$$\tilde{W} = \begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix} \tag{2}$$

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 $\tilde{\mathbf{W}}$ under orthography. This analysis leads to the key result that $\tilde{\mathbf{W}}$ is highly rank-deficient.

Suppose that we place the origin of the world reference system at the centroid of the P points $s_p = (x_p, y_p, z_p)^T$, $p = 1, \ldots, P$ in space that correspond to the P feature points tracked in the image stream (figure 1). The orientation of the camera reference system corresponding to frame number f is determined by a pair of unit vectors $\mathbf{i}_f$, $\mathbf{j}_f$ pointing along the scanlines 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 $\mathbf{j}_f$:

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

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

$$\begin{aligned} u_{fp} &= \mathbf{i}_f^T (\mathbf{s}_p - \mathbf{t}_f) \\ v_{fp} &= \mathbf{j}_f^T (\mathbf{s}_p - \mathbf{t}_f) \end{aligned}$$

where $t_f = (a_f, b_f, c_f)^T$ is the vector from the world origin to the origin of image frame f.

Note that since the origin of the world coordinates is placed at the centroid of the object points, we have

$$\frac{1}{P}\sum_{p=1}^{P} \mathbf{s}_p = 0$$

We can now write expressions for the entries $\tilde{u}_{fp}$ and $\tilde{v}_{fp}$ defined in (1) of the registered measurement matrix. For the registered horizontal image projection we have

$$\begin{aligned} \tilde{u}_{fp} &= u_{fp} - a_f \\ &= \mathbf{i}_f^T (\mathbf{s}_p - \mathbf{t}_f) - \frac{1}{P}\sum_{q=1}^{P} \mathbf{i}_f^T (\mathbf{s}_q - \mathbf{t}_f) \\ &= \mathbf{i}_f^T \left(\mathbf{s}_p - \frac{1}{P}\sum_{q=1}^{P} \mathbf{s}_q \right) \\ &= \mathbf{i}_f^T \mathbf{s}_p \end{aligned} \tag{3}$$

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

$$\begin{aligned} \tilde{u}_{fp} &= \mathbf{i}_f^T \mathbf{s}_p \\ \tilde{v}_{fp} &= \mathbf{j}_f^T \mathbf{s}_p \end{aligned} \tag{4}$$

By collecting the two sets of $F\times P$ equations (4), the registered measurement matrix $\tilde{\mathbf{W}}$ (equation (2)) can be expressed in a matrix form:

$$\tilde{W} = RS \tag{5}$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{i}_1^T \\ \vdots \\ \mathbf{i}_F^T \\ \mathbf{j}_1^T \\ \vdots \\ \mathbf{j}_F^T \end{bmatrix} \tag{6}$$

represents the camera rotation and

$$\mathbf{S} = [\mathbf{s}_1 \cdots \mathbf{s}_P] \tag{7}$$

is the shape matrix. In fact, the rows of $\mathbf{R}$ represent the orientations of the horizontal and vertical camera reference axes throughout the stream, while the columns of $\mathbf{S}$ are the three-dimensional coordinates of the P feature points with respect to their centroid.

Since $\mathbf{R}$ is $2F\times 3$ and $\mathbf{S}$ is $3\times P$, equation (5) implies the following.

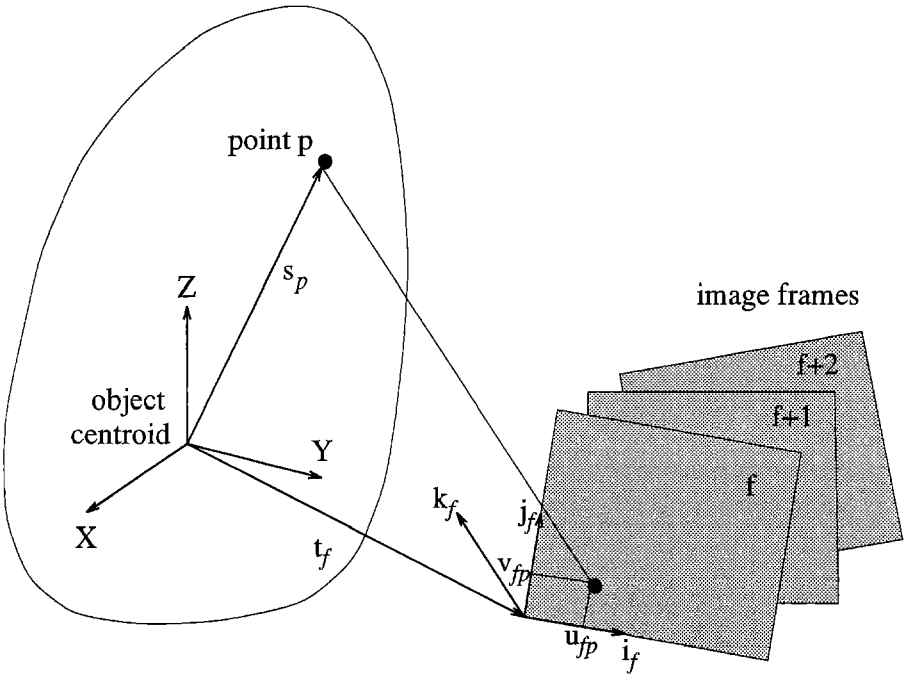


Fig. 1. The two systems of reference used in our problem formulation.

Rank Theorem. *Without noise, the registered measurement matrix $\tilde{W}$ is at most of rank three.*

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 equation (3), the original unregistered matrix $\mathbf{W}$ can be written as

$$\mathbf{W} = \mathbf{RS} + \mathbf{te}_P^T \tag{8}$$

where $\mathbf{t} = (a_1, \ldots, a_F, \ldots, b_F)^T$ is a $2F$-dimensional vector that collects the projections of camera translation along the image plane (see equation (3)), and $\mathbf{e}_P^T = (1, \ldots, 1)$ is a vector of P ones. In scalar form,

$$\begin{aligned} u_{fp} &= \mathbf{i}_f^T \mathbf{s}_p + a_f \\ v_{fp} &= \mathbf{j}_f^T \mathbf{s}_p + b_f \end{aligned} \tag{9}$$

Comparing with equations (1), we see that the two components of camera translation along the image plane are simply the averages of the rows of W.

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

$$|\mathbf{i}_f| = |\mathbf{j}_f| = 1 \quad \text{and} \quad \mathbf{i}_f^T \mathbf{j}_f = 0 \tag{10}$$

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

$$\mathbf{i}_1 = (1, 0, 0)^T \quad \text{and} \quad \mathbf{j}_1 = (0, 1, 0)^T \tag{11}$$

Without noise, the registered measurement matrix $\tilde{\mathbf{W}}$ must be at most of rank 3. When noise corrupts the images, however, $\tilde{\mathbf{W}}$ will not be exactly of rank 3. Fortunately, the rank theorem can be extended to the case of noisy measurements in a well-defined manner. The next subsection introduces the notion of approximate rank, using the concept of singular value decomposition (Golub & Reinsch 1971).

3.2 Approximate Rank

Assuming that $2F \geq P$, the matrix $\tilde{\mathbf{W}}$ can be decomposed (Golub & Reinsch 1971) into a $2F \times P$ matrix $\mathbf{O}_1$, a diagonal $P \times P$ matrix Σ, and a $P \times P$ matrix $\mathbf{O}_2$,

$$\tilde{W} = O_1 \, \Sigma \, O_2 \tag{12}$$

such that $O_1^T O_1 = O_2^T O_2 = O_2 O_2^T = I$, where I is the $P \times P$ identity matrix. The assumption $2F \geq P$ is not crucial: if $2F < P$, everything can be repeated for the transpose of $\tilde{\mathbf{W}}$. Σ is a diagonal matrix whose

diagonal entries are the *singular values* $\sigma_1 \geq \ldots \geq \sigma_P$ sorted in nonincreasing order. This is the *singular-value decomposition* (SVD) of the matrix $\tilde{\mathbf{W}}$.

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

$$\boldsymbol{O}_1 = \underbrace{[\ \boldsymbol{O}_1' \mid \boldsymbol{O}_1''\ ]}_{3 \quad P-3}\ \}\ 2F$$

$$\boldsymbol{\Sigma} = \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}' & 0 \\ \hline 0 & \boldsymbol{\Sigma}'' \end{bmatrix}}_{3 \quad P-3} \begin{matrix} \}\ 3 \\ \}\ P-3 \end{matrix} \qquad (13)$$

$$\boldsymbol{O}_2 = \underbrace{\begin{bmatrix} \boldsymbol{O}_2' \\ \hline \boldsymbol{O}_2'' \end{bmatrix}}_{P} \begin{matrix} \}\ 3 \\ \}\ P-3 \end{matrix}$$

we have

$$\boldsymbol{O}_1\, \boldsymbol{\Sigma}\, \boldsymbol{O}_2 = \boldsymbol{O}_1'\, \boldsymbol{\Sigma}'\, \boldsymbol{O}_2' + \boldsymbol{O}_1''\, \boldsymbol{\Sigma}''\, \boldsymbol{O}_2''$$

Let $\tilde{\mathbf{W}}^*$ be the ideal registered measurement matrix, that is, the matrix we would obtain in the absence of noise. Because of the rank theorem, $\tilde{\mathbf{W}}^*$ has at most three nonzero singular values. Since the singular values in $\boldsymbol{\Sigma}$ are sorted in nonincreasing order, $\boldsymbol{\Sigma}'$ must contain all the singular values of $\tilde{\mathbf{W}}^*$ that exceed the noise level. Furthermore, it can be shown (Golub & Van Loan 1989) that the best possible rank-3 approximation to the ideal registered measurement matrix $\tilde{\mathbf{W}}^*$ is the product

$$\hat{W} = \boldsymbol{O}_1'\, \boldsymbol{\Sigma}'\, \boldsymbol{O}_2'$$

We can now restate our rank theorem for the case of noisy measurements.

Rank Theorem for Noisy Measurements. *The best possible shape and rotation estimate is obtained by considering only the three greatest singular values of $\tilde{W}$, together with the corresponding left and right eigenvectors.*

Thus, $\hat{\mathbf{W}}$ is the best estimate of $\tilde{\mathbf{W}}^*$. Now if we define

$$\hat{\mathbf{R}} = \boldsymbol{O}_1'\, [\boldsymbol{\Sigma}']^{1/2}$$

$$\hat{\mathbf{S}} = [\boldsymbol{\Sigma}']^{1/2}\, \boldsymbol{O}_2'$$

we can write

$$\hat{W} = \hat{R}\, \hat{S} \qquad (14)$$

The two matrices $\hat{\mathbf{R}}$ and $\hat{\mathbf{S}}$ are of the same size as the desired rotation and shape matrices $\mathbf{R}$ and $\mathbf{S}$: $\hat{\mathbf{R}}$ is $2F\times3$, and $\hat{\mathbf{S}}$ is $3\times P$. However, the decomposition (14) is not unique. In fact, if $\mathbf{Q}$ is *any* invertible 3×3 matrix, the matrices $\hat{\mathbf{R}}\mathbf{Q}$ and $\mathbf{Q}^{-1}\hat{\mathbf{S}}$ are also a valid decomposition of $\hat{W}$, since

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

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

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

3.3 The Metric Constraints

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

$$\begin{aligned} \mathbf{R} &= \hat{\mathbf{R}}\mathbf{Q} \\ \mathbf{S} &= \mathbf{Q}^{-1}\, \hat{\mathbf{S}} \end{aligned} \qquad (15)$$

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

$$\begin{aligned} \hat{\mathbf{i}}_f^T\, \boldsymbol{Q}\boldsymbol{Q}^T\, \hat{\mathbf{i}}_f &= 1 \\ \hat{\mathbf{j}}_f^T\, \boldsymbol{Q}\boldsymbol{Q}^T\, \hat{\mathbf{j}}_f &= 1 \\ \hat{\mathbf{i}}_f^T\, \boldsymbol{Q}\boldsymbol{Q}^T\, \hat{\mathbf{j}}_f &= 0 \end{aligned} \qquad (16)$$

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 is arbitrary. This arbitrariness can be removed by enforcing the constraints (11), that is, by selecting the 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 $\tilde{\mathbf{W}}$ derived from a stream of images into shape $\mathbf{S}$ and rotation $\mathbf{R}$ as defined in equations (5)–(7).

1. Compute the singular-value decomposition $\tilde{\mathbf{W}} = O_1 \Sigma O_2$.
2. Define $\hat{\mathbf{R}} = O_1'(\Sigma')^{1/2}$ and $\hat{\mathbf{S}} = (\Sigma')^{1/2} O_2'$, , where the primes refer to the block partitioning defined in (13).
3. Compute the matrix $\mathbf{Q}$ in equations (15) by imposing the metric constraints (equations (16)).
4. Compute the rotation matrix $\mathbf{R}$ and the shape matrix $\mathbf{S}$ as $\mathbf{R} = \hat{\mathbf{R}}\mathbf{Q}$ and $\mathbf{S} = \mathbf{Q}^{-1}\hat{\mathbf{S}}$.
5. If desired, align the first camera reference system with the world reference system by forming the products $\mathbf{R}\mathbf{R}_0$ and $\mathbf{R}_0^T\mathbf{S}$, where the orthonormal matrix $\mathbf{R}_0 = [\mathbf{i}_1\ \mathbf{j}_1\ \mathbf{k}_1]$ rotates the first camera reference system into the identity matrix.

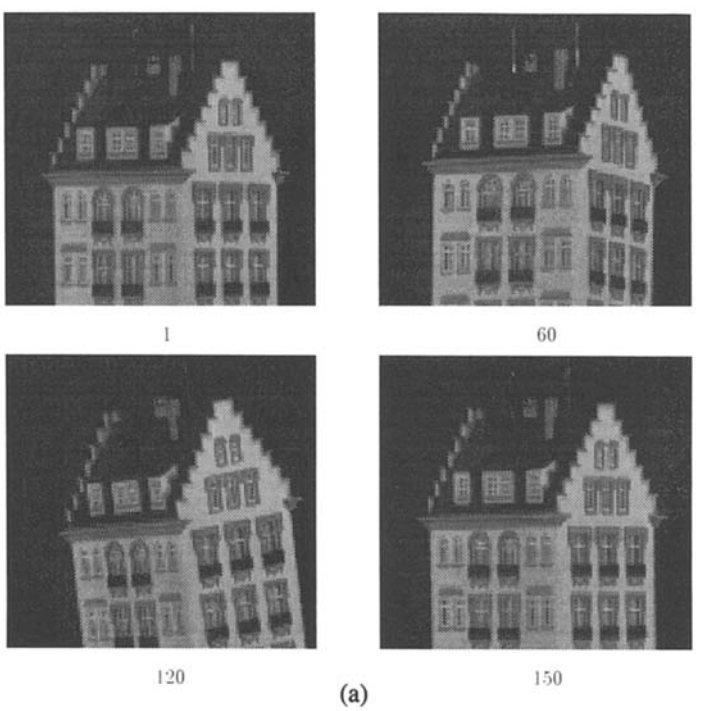
4 Experiments

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

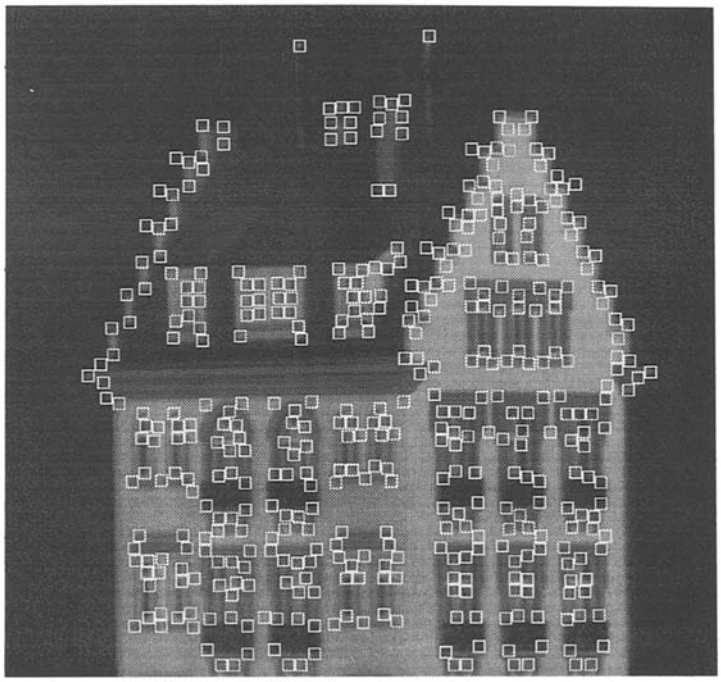
Some frames in this stream are shown in figure 2a. The images depict a small plastic model of a building. The camera is a Sony CCD camera with a 200 mm lens, and is moved by means of a high-precision positioning platform. Camera pitch, yaw, and roll around the model are all varied as shown by the dashed curves in figure 3a. The translation of the camera is such as to keep the building within the field of view of the camera.

For feature tracking, we extended the Lucas-Kanade method described in (Lucas & Kanade 1981) to allow also for the automatic selection of image features. This method obtains the displacement vector of the window around a feature as the solution of a linear 2×2 equation system. Good image features are automatically selected as those points for which the above equation systems are stable. The details are presented in (Tomasi & Kanade 1991b; Tomasi 1991).



(a)

Fig. 2a. The "Hotel" stream: four of the 150 frames.



(b)

Fig. 2b. The "Hotel" stream: the 430 features selected by the automatic detection method.

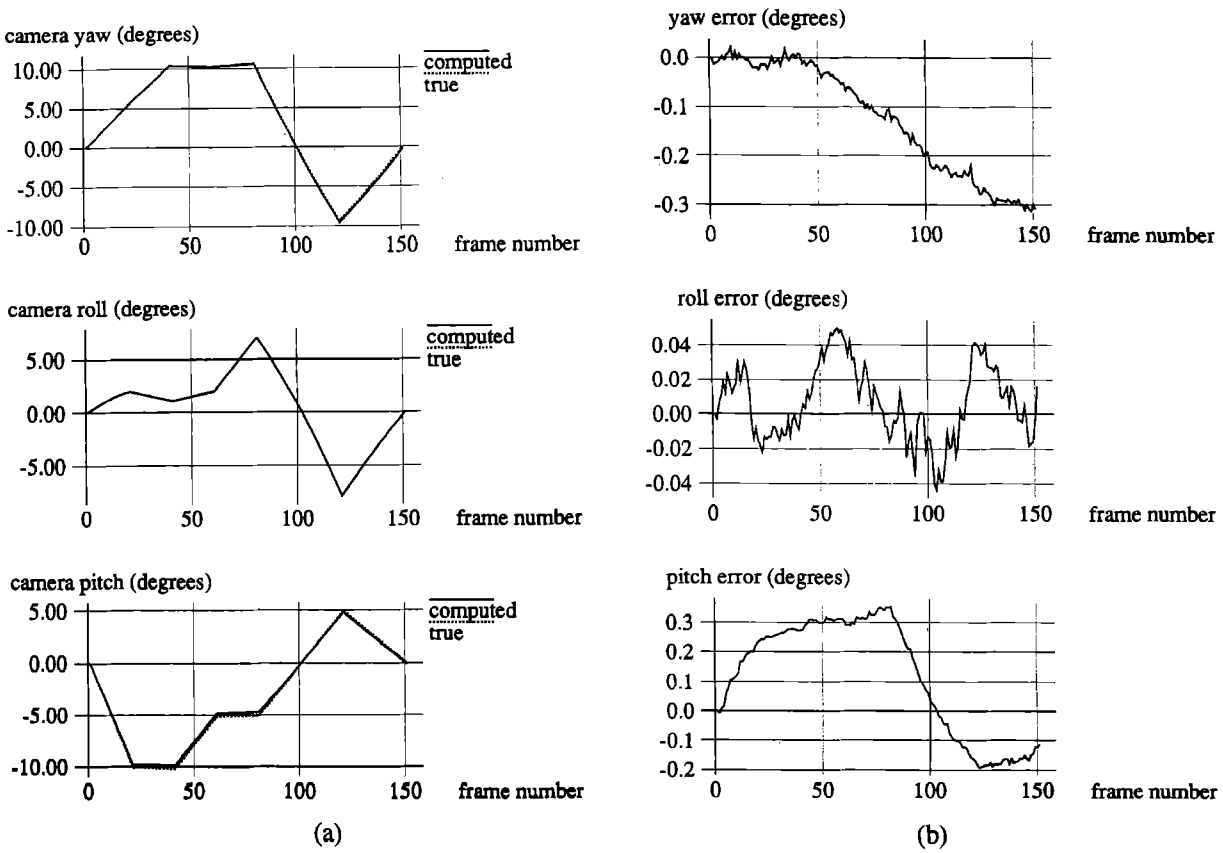


Fig. 3. Motion results for the "Hotel" stream: (a) true and computed camera rotation and (b) blow-up of the errors in (a).

The entire set of 430 features thus selected is displayed in figure 2b, 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.

The motion recovery is precise. The plots in figure 3a compare the rotation components computed by the factorization method (solid curves) with the values measured mechanically from the mobile platform (dashed curves). The differences are magnified in figure 3b. The errors are everywhere less than 0.4 degrees and on average 0.2 degrees. The computed motion follows closely also rotations with curved profiles, such as the roll profile between frames 1 and 20 (second plot in figure 3a), 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. That is, the motion is actually degenerate during this period, yet 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* in the method.

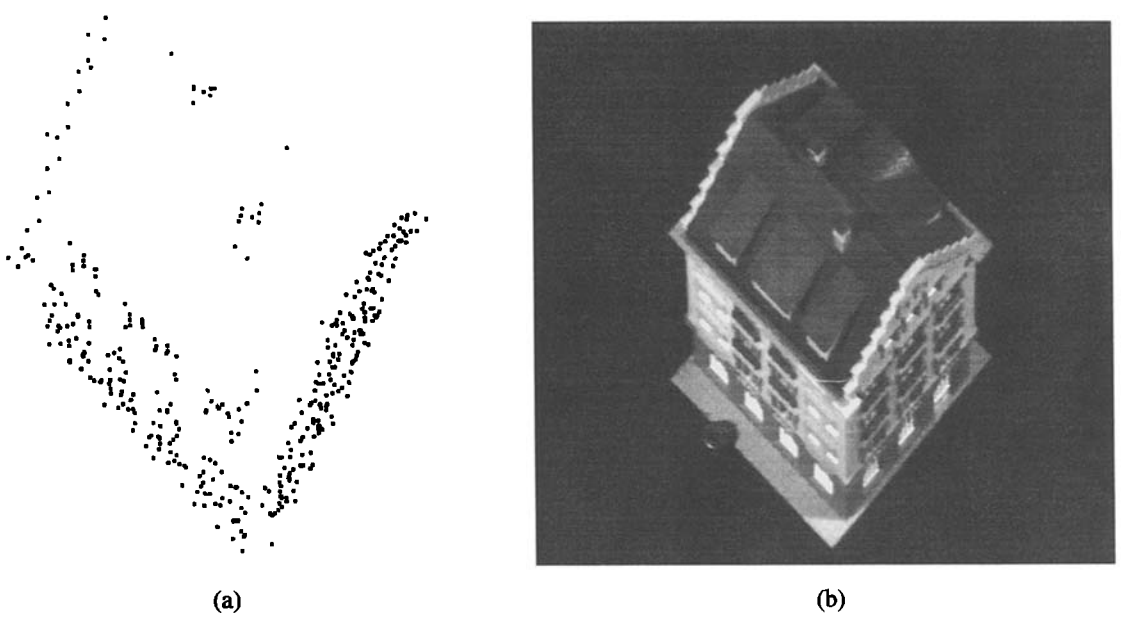
The shape results are evaluated qualitatively in figure 4, which compares the computed shape viewed from above (a) with the actual shape (b). Notice that the walls, the windows on the roof, and the chimneys are recovered in their correct positions.

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. Figure 5a shows the selected features. The diagram in figure 5b compares measured and computed distances. 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 (one millimeter).

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.

Figure 6a shows some of the 180 frames of the "House" stream. The overall motion covers a relatively small rotation angle, approximately 15 degrees. Outdoor images are harder to process than those produced



(a) (b)

Fig. 4. Qualitative shape results for the "Hotel" stream: top view of the (a) computed and (b) actual shape.

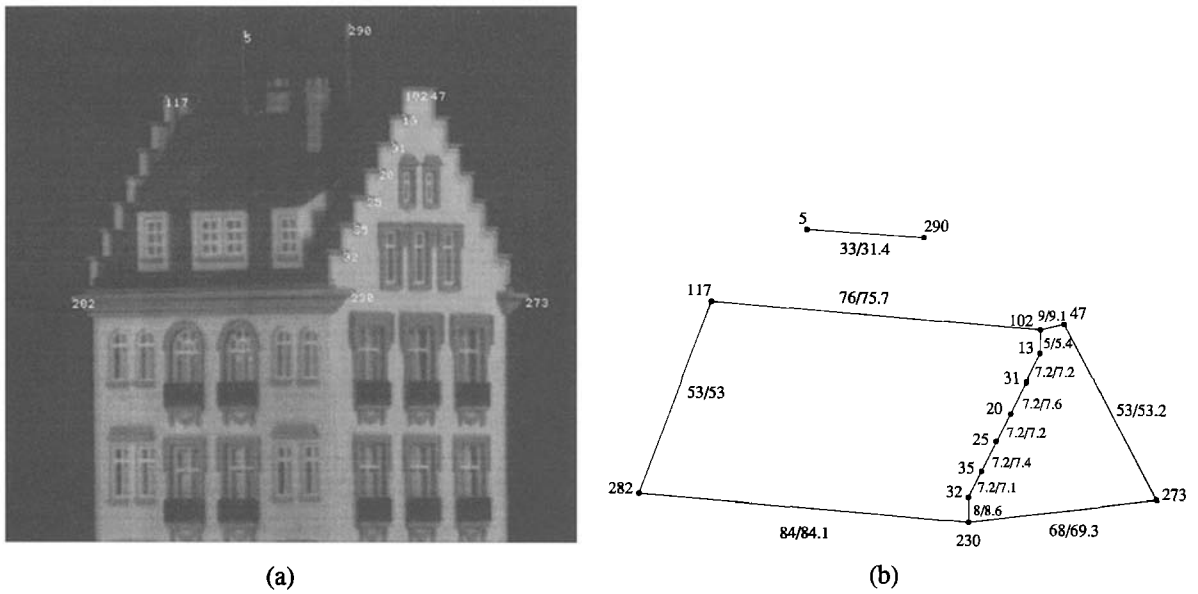


Fig. 5. Quantitative shape results for the "Hotel" stream: the features in (a) were measured with a ruler on the building model, and are compared in (b) with the computed distances (measured/computed, in mm). The scale factor was computed from the distance between features 117 and 282.

in the controlled environment of a 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 during both 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 figure 6b. 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 these two parts of the image from the analysis.

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

Figure 7 shows a front and a top view of the building as reconstructed by the factorization method. To render these figures for display, we triangulated the computed 3D 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. In these figures, 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 demonstrates 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 of the environment, remains an open problem that must be solved for a fully autonomous system. An initial effort has been seen in (Boult & Brown 1991).

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 stream. These phenomena are frequent enough to make a shape and motion computation method unrealistic if it cannot deal with them.

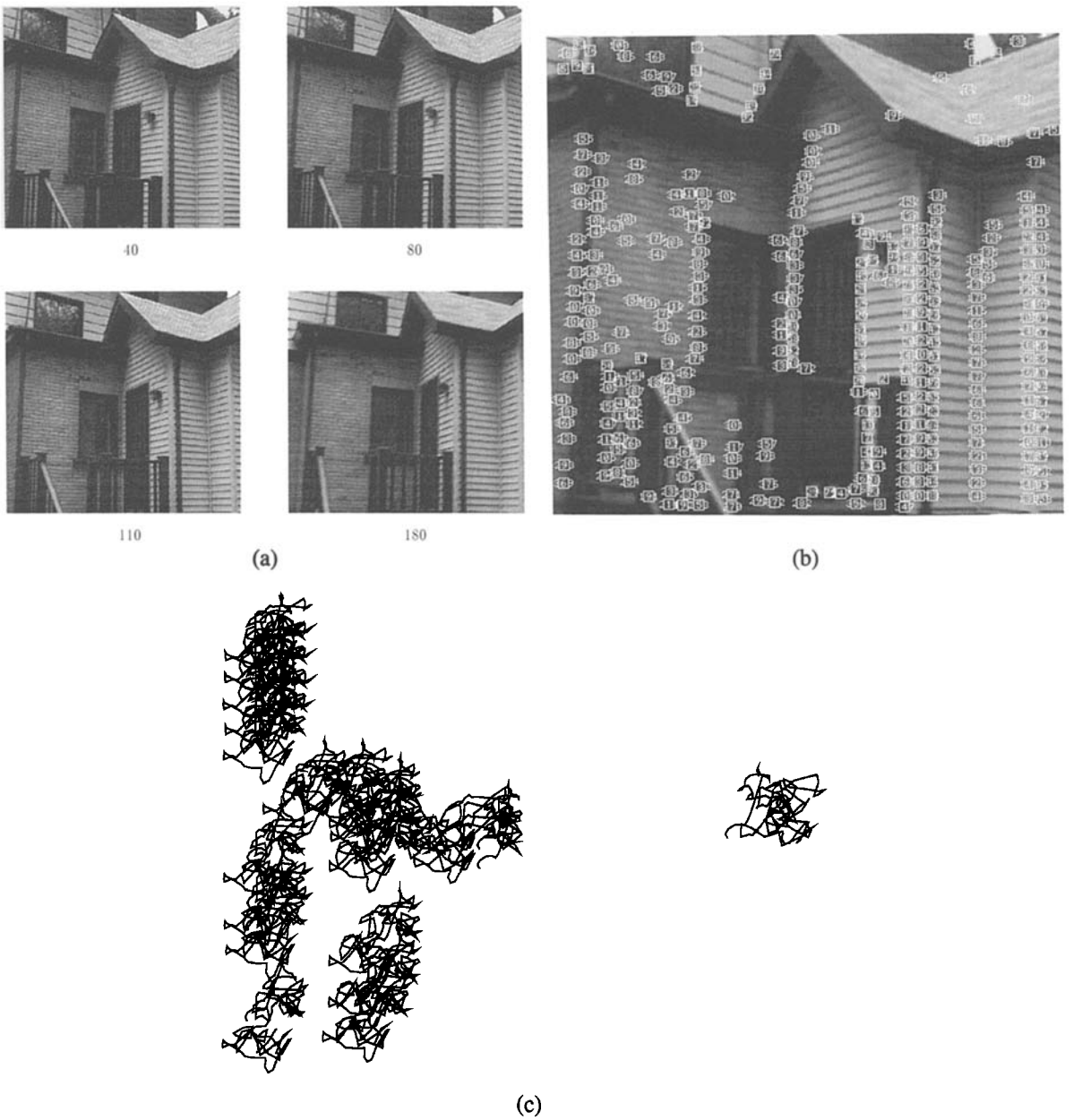


Fig. 6. The "House" stream: (a) four of the 180 frames, (b) the features automatically selected in the first frame, and (c) tracks of 60 features.

Sequences with appearing and disappearing features result in a measurement matrix **W** which 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, we cannot only solve the shape and motion recovery problem from the incomplete measurement matrix **W**, but we can even hallucinate the unknown entries of **W** by projecting the computed three-dimensional feature coordinates onto the computed camera positions, as shown in the following.

5.1 Solution for Noise-Free Images

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

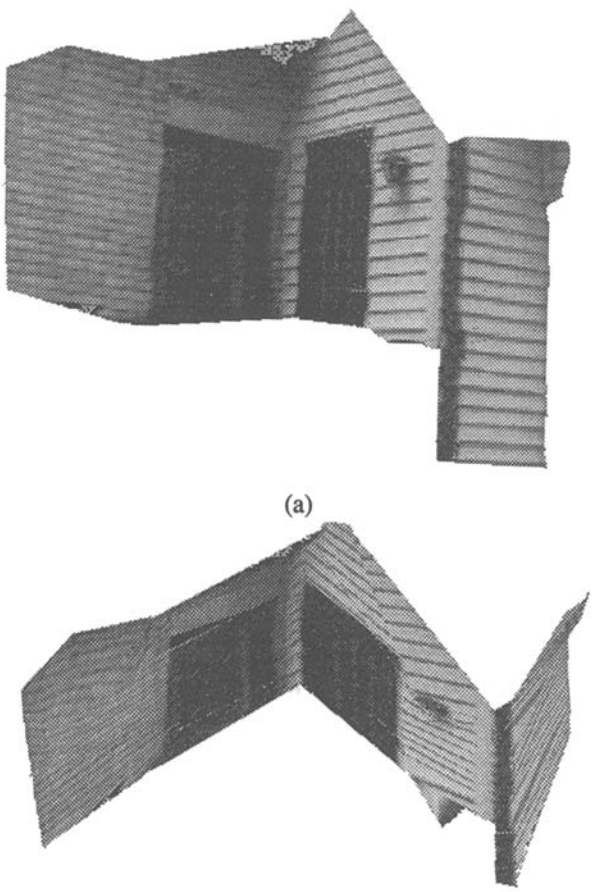


Fig. 7. Shape results for the "House" stream: (a) front and (b) top view of the three walls with image intensities mapped onto the reconstructed surface.

other features, the corresponding camera position should be recoverable as well. Then from point and camera positions thus recovered, we should also be able to reconstruct the missing image measurement. In fact, 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 f_1, f_2, f_3, f.

In figure 8, this means that the dotted entries must be known to reconstruct the question marks. This is equivalent to Ullman's result (Ullman 1979) that three views of four points determine structure and motion.

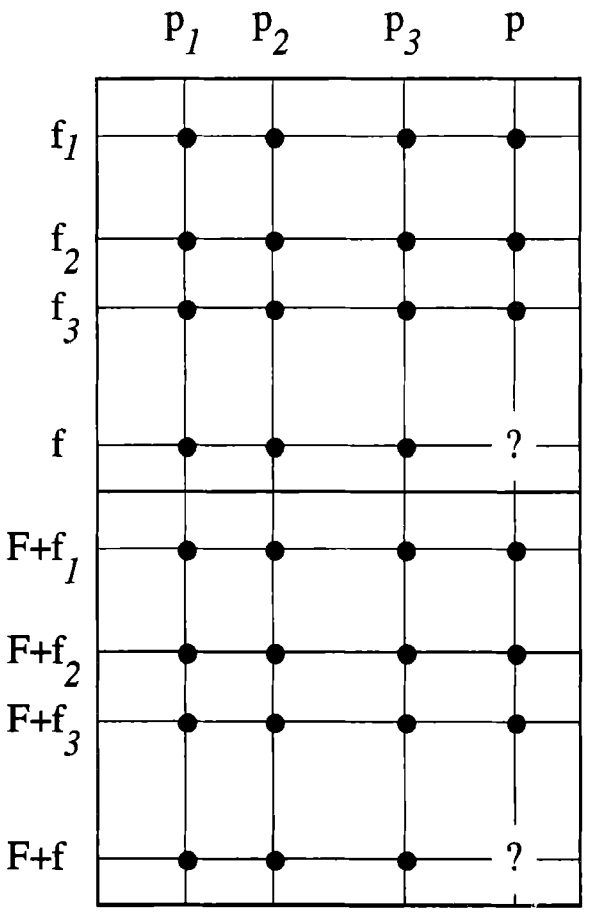


Fig. 8. The reconstruction condition. If the dotted entries of the measurement matrix are known, the two unknown ones (question marks) can be hallucinated.

In this subsection, we provide the reconstruction condition in our formalism and develop the reconstruction procedure. To this end, we notice that the rows and columns of the noise-free measurement matrix **W** can always be permuted so that $f_1 = p_1 = 1$, $f_2 = p_2 = 2$, $f_3 = p_3 = 3$, $f = p = 4$. We can therefore suppose that u_{44} and v_{44} are the only two unknown entries in the 8×4 matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \hline \mathbf{V} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & ? \\ v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & ? \end{bmatrix}$$

Then, the factorization method can be applied to the first three rows of **U** and **V**, that is, to the 6×4 submatrix

$$\mathbf{W}_{6\times4} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \end{bmatrix}$$

to produce the partial translation and rotation submatrices

$$\mathbf{t}_{6\times1} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \text{and} \quad \mathbf{R}_{6\times3} = \begin{bmatrix} \mathbf{i}_1^T \\ \mathbf{i}_2^T \\ \mathbf{i}_3^T \\ \mathbf{j}_1^T \\ \mathbf{j}_2^T \\ \mathbf{j}_3^T \end{bmatrix} \tag{18}$$

and the full-shape matrix

$$\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ \mathbf{s}_3 \ \mathbf{s}_4] \tag{19}$$

such that

$$\mathbf{W}_{6\times4} = \mathbf{R}_{6\times3}\,\mathbf{S} + \mathbf{t}_{6\times1}\,\mathbf{e}_4^T$$

where $\mathbf{e}_4^T = (1, 1, 1, 1)$.

To complete the rotation solution, we need to compute the vectors $\mathbf{i}_4$ and $\mathbf{j}_4$. However, a registration problem must be solved first. In fact, only three points are visible in the fourth frame, while equation (19) yields all four points in space. Since the factorization method computes the space coordinates with respect to the centroid of the points, we have $\mathbf{s}_1 + \mathbf{s}_2 + \mathbf{s}_3 + \mathbf{s}_4 = 0$, while the image coordinates in the fourth frame are measured with respect to the image centroid of just three observed points (1, 2, 3). Thus, before we can compute $\mathbf{i}_4$ and $\mathbf{j}_4$ we must make the two origins coincide by referring all coordinates to the centroid

$$\mathbf{c} = \frac{1}{3}(\mathbf{s}_1 + \mathbf{s}_2 + \mathbf{s}_3)$$

of the three points that are visible in all four frames. In the fourth frame, the projection of $\mathbf{c}$ has coordinates

$$a_4' = \frac{1}{3}(u_{41} + u_{42} + u_{43})$$

$$b_4' = \frac{1}{3}(v_{41} + v_{42} + v_{43})$$

so we can define the new coordinates

$$\mathbf{s}_p' = \mathbf{s}_p - \mathbf{c} \qquad \text{for} \qquad p = 1, 2, 3$$

in space and

$$\begin{aligned} u_{4p}' &= u_{4p} - a_4' \\ v_{4p}' &= v_{4p} - b_4' \end{aligned} \qquad \text{for} \qquad p = 1, 2, 3$$

in the fourth frame. Then, $\mathbf{i}_4$ and $\mathbf{j}_4$ are the solutions of the two 3×3 systems

$$\begin{aligned} [u_{41}' \ u_{42}' \ u_{43}'] &= \mathbf{i}_4^T\,[\mathbf{s}_1' \ \mathbf{s}_2' \ \mathbf{s}_3'] \\ [v_{41}' \ v_{42}' \ v_{43}'] &= \mathbf{j}_4^T\,[\mathbf{s}_1' \ \mathbf{s}_2' \ \mathbf{s}_3'] \end{aligned} \tag{20}$$

derived from equation (5). The second equation in (18) and the solution to (20) yield the entire rotation matrix $\mathbf{R}$, while shape is given by equation (19).

The components a_4 and b_4 of translation in the fourth frame with respect to the centroid of all four points can be computed by postmultiplying equation (8) by the vector $\eta_4 = (1, 1, 1, 0)^T$:

$$\mathbf{W}\,\eta_4 = RS\,\eta_4 + \mathbf{t}\mathbf{e}_4^T\,\eta_4$$

Since $\mathbf{e}_4^T\,\eta_4 = 3$, we obtain

$$\mathbf{t} = \frac{1}{3}(\mathbf{W} - \mathbf{RS})\,\eta_4 \tag{21}$$

In particular, rows 4 and 8 of this equation yield a_4 and b_4. Notice that the unknown entries u_{44} and v_{44} are multiplied by zeroes in equation (21).

Now that both motion and shape are known, the missing entries u_{44}, v_{44} of the measurement matrix $\mathbf{W}$ can be found by orthographic projection (equation (9)):

$$u_{44} = \mathbf{i}_4^T\,\mathbf{s}_4 + a_4$$

$$v_{44} = \mathbf{j}_4^T\,\mathbf{s}_4 + b_4$$

The procedure thus completed factors the full 6×4 submatrix of $\mathbf{W}$ and then reasons on the three points that are visible in all the frames to compute motion for the fourth frame.

Alternatively, one can start with the 8×3 submatrix

$$\mathbf{W}_{8\times3} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \\ v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \\ v_{41} & v_{42} & v_{43} \end{bmatrix} \tag{22}$$

In this case we first compute the full translation and rotation submatrices, and then from these we obtain the shape coordinates and the unknown entry of $\mathbf{W}$ for full reconstruction.

In summary, the full motion and shape solution can be found in either of the following ways:

1. *row-wise extension:* factor $\mathbf{W}_{6\times4}$ to find a partial motion and full shape solution, and propagate it to include motion for the remaining frame (equations (20)).
2. *column-wise extension:* factor $\mathbf{W}_{8\times3}$ to find a full motion and partial shape solution, and propagate it to include the remaining feature point.

5.2 Solution in the Presence of Noise

The solution-propagation method introduced in the previous subsection can be extended to $2F\times P$ measurement matrices with $F \geq 4$ and $P \geq 4$. In fact, the only difference is that the propagation equations (20) for row-wise extension and the analogous ones for column-wise extension become overconstrained. If the measurement matrix $\mathbf{W}$ is noisy, this redundancy is beneficial, since equations (20) can be solved in the least-square-error sense, and the effect of noise is reduced.

In the general case of a noisy $2F\times P$ matrix $\mathbf{W}$ the solution-propagation method can be summarized as follows. A possibly large, full subblock of $\mathbf{W}$ is first decomposed by factorization. Then, this initial solution is grown one row or one column at a time by solving systems analogous to those in (20) in the least-square-error sense.

However, because of noise, the order in which the rows and columns of $\mathbf{W}$ are incorporated into the solution can affect the exact values of the final motion and shape solution. Consequently, once the solution has been propagated to the entire measurement matrix $\mathbf{W}$, it may be necessary to refine the results with a steepest-descent minimization of the residue

$$\|\mathbf{W} - \mathbf{RS} - \mathbf{te}_P^T\|$$

(see equation (8)).

There remain the two problems of how to choose the initial full subblock to which factorization is applied and in what order to grow the solution. In fact, however, because of the final refinement step, neither choice is critical as long as the initial matrix is large enough to yield a good starting point. We illustrate this point in section 6.

6 More Experiments

We now test the propagation method with image streams which 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.

6.1 "Ball" Image Stream

A ping-pong ball with black dots marked on its surface is rotated 450 degrees in front of the camera, so features appear and disappear. The rotation between adjacent frames is 2 degrees, so the stream is 226 frames long. Figure 9a shows the first frame of the stream, with the automatically selected features overlaid.

The feature tracker looks for new features every 30 frames (60 degrees) of rotation. In this way, features that disappear on one side around the ball are replaced by new ones that appear on the other side. Figure 9b shows the tracks of 60 features, randomly chosen among the total of 829 found by the selector.

If all measurements are collected into the noisy measurement matrix $\mathbf{W}$, the $\mathbf{U}$ and $\mathbf{V}$ parts of $\mathbf{W}$ have the same fill pattern: if the x coordinate of a measurement is known, so is its y coordinate. Figure 9c shows this *fill matrix* for our experiment. This matrix has the same size as either $\mathbf{U}$ or $\mathbf{V}$, that is, $F\times P$. A column corresponds to a feature point, and a row to a frame. Shaded regions denote known entries. The fill matrix shown has $226 \times 829 = 187354$ entries, of which 30185 (about 16 percent) 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 a linear system similar to (20) to add new rows, and of the analogous system for the column-wise extension to add new columns. The rows and columns to add are selected so as to maximize the redundancy of the linear systems. Eventually, all of the motion and shape values are determined. As a result, the unknown 84 percent of the measurement matrix can be hallucinated from the known 16 percent.

Figure 10 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.

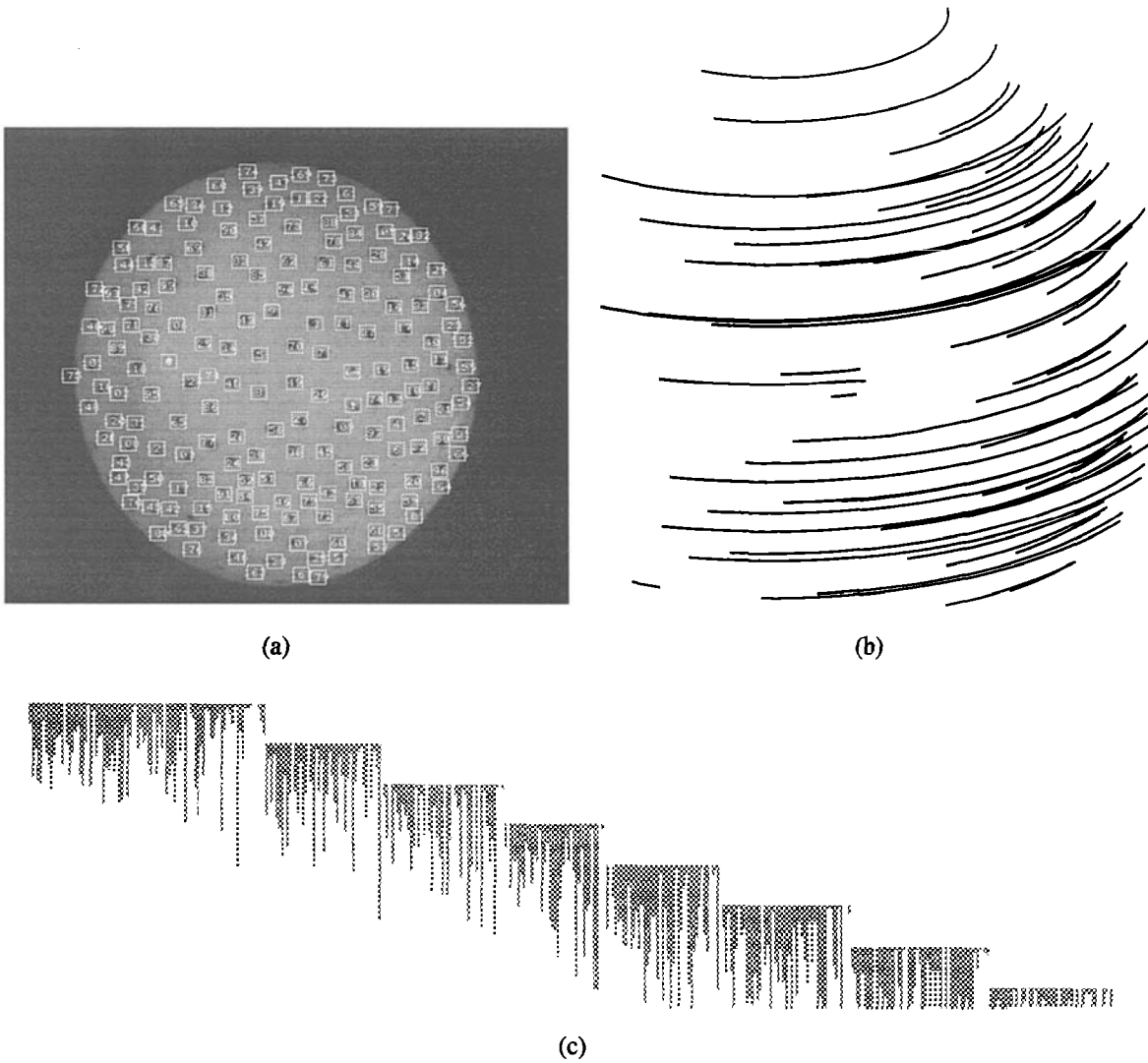


Fig. 9. The "Ball" stream : (a) the first frame, (b) tracks of 60 features, and (c) the fill matrix (shaded entries are known image coordinates).

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, ..., F. Because the ball rotates around a fixed axis, both $\mathbf{i}_f$ and $\mathbf{j}_f$ should sweep a cone in space, as shown in figure 11a. The tips of $\mathbf{i}_f$ and $\mathbf{j}_f$ should describe two circles in space, centered along the axis of rotation. Figure 11b 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 figure 11b corresponding to more than 360 degrees rotation. If the motion reconstruction were perfect, the two arcs would be indistinguishable.

6.2 *The "Hand" Image Stream*

In this subsection we describe an experiment with a natural scene including occlusion as a dominant phenomenon. A hand holds a cup and rotates it by about ninety degrees in front of the camera mounted on a fixed stand. Figure 12a shows four out 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

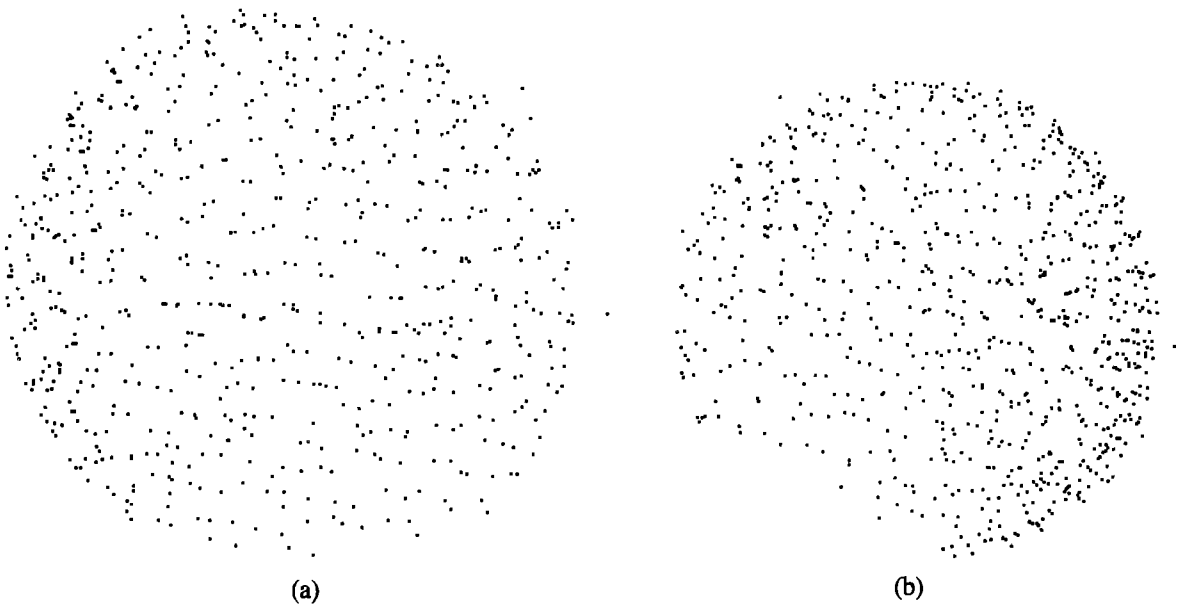


Fig. 10. Shape results for the "Ball" stream: (a) top and (b) side view.

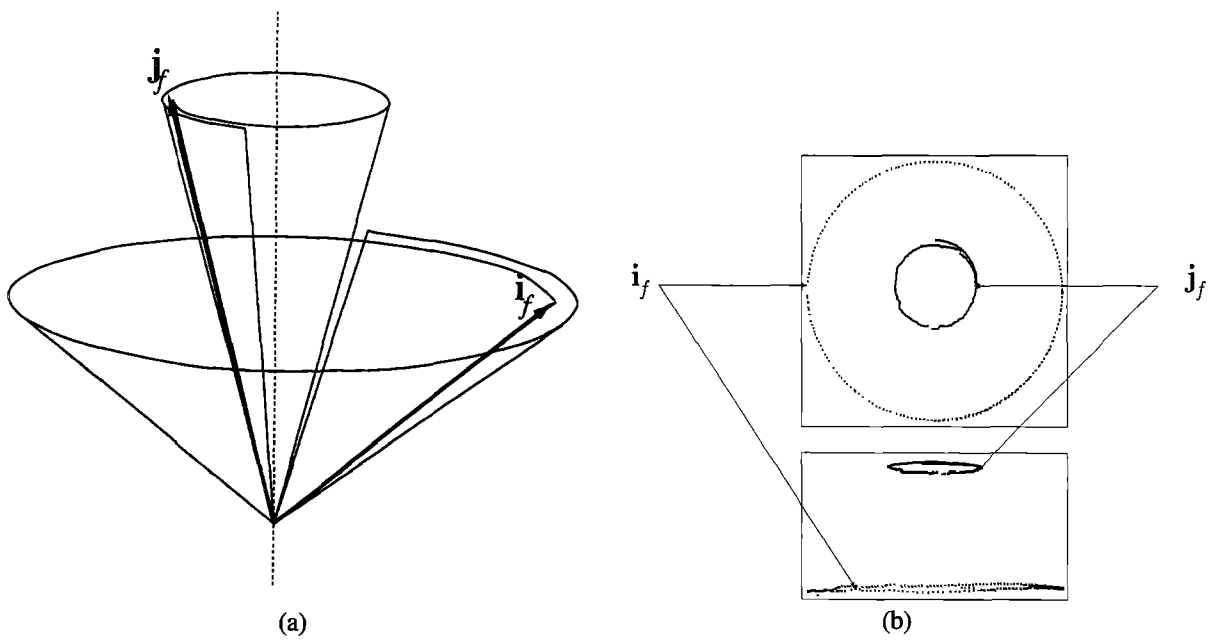


Fig. 11. Motion results for the "Ball" stream: (a) because the ball rotates around a fixed axis, the two orthogonal unit vectors $\mathbf{i}_f$ and $\mathbf{j}_f$ along rows and columns of the image sensor sweep two cones in space; (b) top and side views of the computed vectors $\mathbf{i}_f$ and $\mathbf{j}_f$.

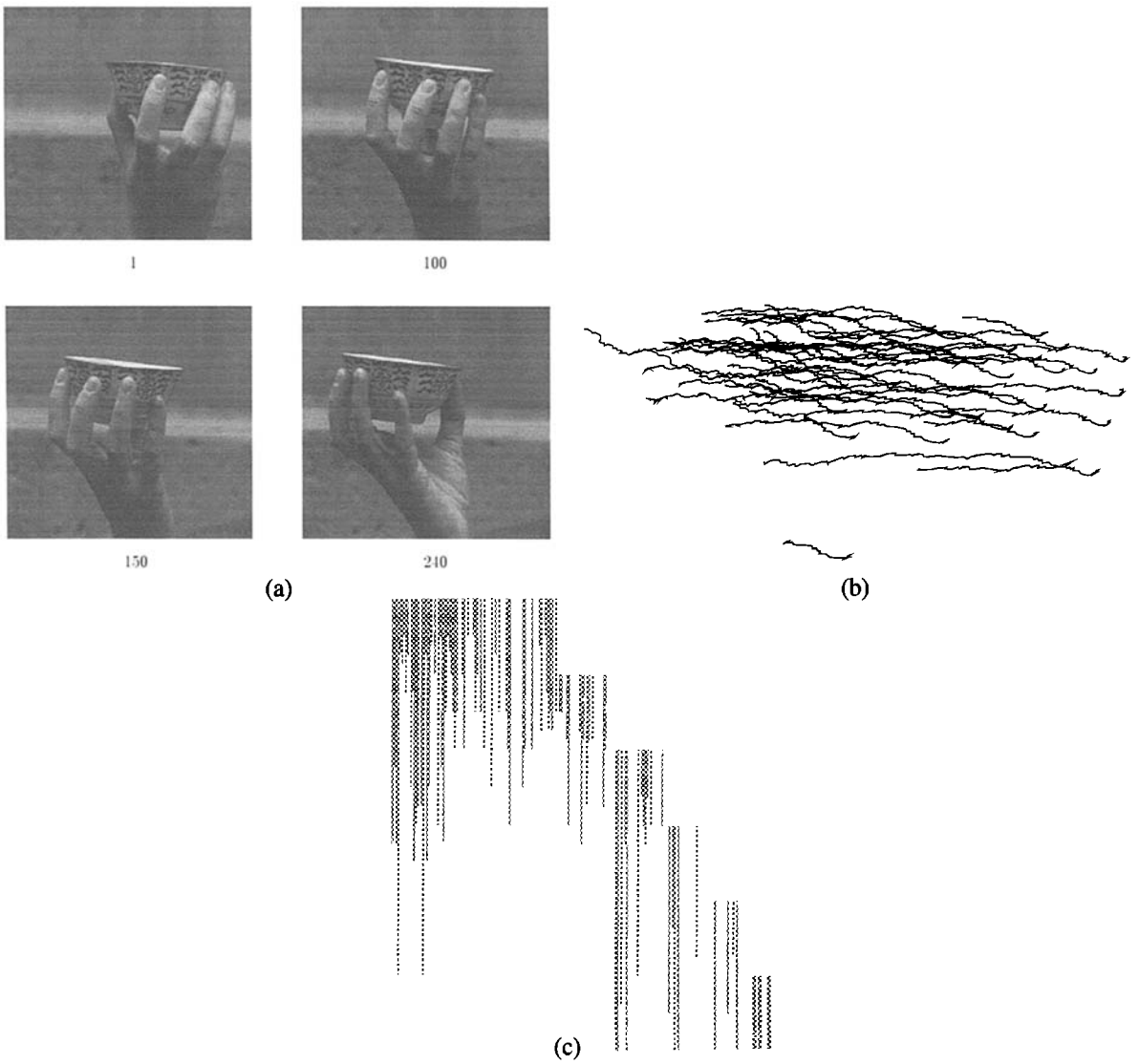


Fig. 12. The "Hand" stream: (a) four of the 240 frames, (b) tracks of 60 features, and (c) the fill matrix (shaded entries are known image coordinates).

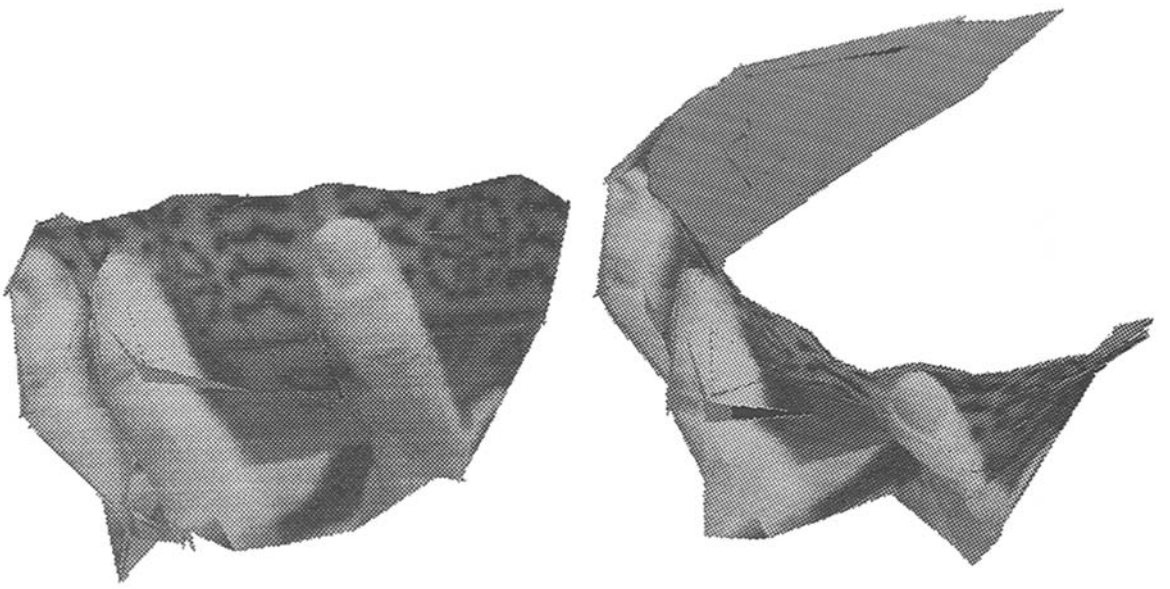
changes slightly. This effect, however, is small and did not affect the results appreciably.

A total of 207 features was selected. Figure 12b shows the image trajectory of 60 randomly selected features. Occlusions were marked by hand in this experiment. The fill matrix of figure 12c illustrates the occlusion pattern.

Figure 13 shows a front and a top 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 subsection 4.1, by triangulating the tracked feature points and mapping pixel values onto the resulting surface.

7 Conclusions

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 under orthography 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)



(a) (b)

Fig. 13. Shape results for the "Hand" stream: (a) front and (b) top view of the cup and fingers with image intensities mapped onto the reconstructed surface.

fundamentally superfluous. The theorem is powerful because the rank theorem leads to 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 using 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 twelve-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.

References

Adiv, G. 1985. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Patt. Anal. Mach. Intell.* 7:384–401.

Bolles, R.C., Baker, H.H., and Marimont, D.H. 1987. Epipolar-plane image analysis: An approach to determining structure from motion, *Intern. J. Comput. Vis.* 1(1):7–55.

Boult, T.E., and Brown, L.G. 1991. Factorization-based segmentation of motions, *Proc. IEEE Workshop on Visual Motion*, pp. 179–186.

Broida, T., Chandrashekhar, S., and Chellappa, R. 1990. Recursive 3D motion estimation from a monocular image sequence, *IEEE Trans. Aerospace Electroc. Syst.* 26(4):639–656.

Bruss, A.R., and Horn, B.K.P. 1983. Passive navigation. *Comput. Vis. Graph. Image Process.* 21:3–20.

Debrunner, C., and Ahuja, N. 1992. Motion and structure factorization and segmentation of long multiple motion image sequences. In Sandini, G., ed. *Europ. Conf. Comput. Vision*, 1992, pp. 217–221. Springer-Verlag: Berlin, Germany.

Golub, G.H., and Reinsch, C. 1971. Singular value decomposition and least squares solutions, In *Handbook for Automatic Computation*, vol. 2, ch. I/10, pp. 134–151. Springer Verlag: New York.

Golub, G.H., and Van Loan, C.F. 1989. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD.

Heeger, D.J., and Jepson, A. 1989. Visual perception of three-dimensional motion, Technical Report 124, MIT Media Laboratory, Cambridge, MA.

Heel, J. 1989. Dynamic motion vision. *Proc. DARPA Image Understanding Workshop*, Palo Alto, CA, pp. 702–713.

Horn, B.K.P., Hilden, H.M., and Negahdaripour, S. 1988. Closed-form solution of absolute orientation using orthonormal matrices. *J. Op. Soc. Amer. A*, 5(7):1127–1135.

Lucas, B.D., and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision, *Proc. 7th Intern. Joint Conf. Artif. Intell.*, Vancouver.

Matthies, L., Kanade, T., and Szeliski, R. 1989. Kalman filter-based algorithm for estimating depth from image sequences. *Intern. J. Comput. Vis.* 3(3):209–236.

Prazdny, K. 1980. Egomotion and relative depth from optical flow, *Biological Cybernetics* 102:87–102.

Spetsakis, M.E., and Aloimonos, J.Y. 1989. Optimal motion estimation. *Proc. IEEE Workshop on Visual Motion*, pp. 229–237. Irvine, CA.

Tomasi, C., and Kanade, T. 1990. Shape and motion without depth, *Proc. 3rd Intern. Conf. Comput. Vis.*, Osaka, Japan.

Tomasi, C., and Kanade, T. 1991a. Shape and motion from image streams: a factorization method—2. point features in 3D motion. Technical Report CMU-CS-91-105, Carnegie Mellon University, Pittsburgh, PA.

Tomasi, C., and Kanade, T. 1991b. Shape and motion from image streams: a factorization method—3. detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, PA.

Tomasi, C. 1991. Shape and motion from image streams: a factorization method. Ph.D. thesis, Carnegie Mellon University. Also appears as Technical Report CMU-CS-91-172.

Tsai, R.Y., and Huang, T.S. 1984. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *IEEE Trans. Patt. Anal. Mach. Intell.* 6(1):13–27.

Ullman, S. 1979. *The Interpretation of Visual Motion*. MIT Press: Cambridge, MA.

Waxman, A.M., and Wohn, K. 1985. Contour evolution, neighborhood deformation, and global image flow: planar surfaces in motion. *Intern. J. Robot. Res.* 4:95–108.