

Estimating the Kinematics and Structure of a Rigid Object from a Sequence of Monocular Images

Ted J. Broida, *Member, IEEE*, and Rama Chellappa, *Senior Member, IEEE*

Abstract—The problem considered here involves the use of a sequence of noisy monocular images of a three-dimensional (3-D) moving object to estimate both its structure and kinematics. The object is assumed to be rigid, and its motion is assumed to be “smooth.” A set of object match points is assumed to be available, consisting of fixed features on the object, the image plane coordinates of which have been extracted from successive images in the sequence. Structure is defined as the 3-D positions of these object feature points, relative to each other. Rotational motion occurs about the origin of an object-centered coordinate system, while translational motion is that of the origin of this coordinate system.

In previous work [5]–[8] we have developed a model based approach for motion/structure estimation using a long sequence of monocular images. This approach provides a great deal of flexibility, by allowing the use of arbitrarily many image frames and feature points, and each model can easily be modified or extended for different problems. Our earlier work involved assumptions about object structure and/or motion, were primarily tested on simulated imagery, and did not address the issue of uniqueness of the model parameters.

In this paper, which is a continuation of the research started in [7], results of an experiment with real imagery are presented, involving estimation of 28 unknown translational, rotational, and structural parameters, based on 12 images with 7 feature points. Uniqueness results are summarized for the case of purely translational motion. A test based on a singular value decomposition is described that determines whether or not noise-free data from an image sequence uniquely determines the elements of any given parameter vector, and empirical support of this test is given.

Index Terms—Image sequence analysis, motion analysis, 3-D motion estimation, uniqueness.

I. INTRODUCTION

THE problem of estimating various 3-D parameters from a time sequence of 2-D projections (images) has been the focus of a significant amount of research during the past decade. This problem occurs in a wide variety of situations, which range from estimating rigid object structure and motion (e.g., [10], [22], [32]), to self-motion estimation and navigation (e.g., [9], [27]), to tracking certain 3-D parameters of “point” objects (e.g., [28]), and to many others. In all cases, there is either an implicit or explicit use of various models for structure, motion, and imaging, the parameters of which are

to be estimated, whether they are translational or rotational increments, rates, feature positions, etc. Most of the advantages of model-based methods discussed in [9] (for recursive estimation of self-motion) apply to model-based batch and hybrid batch/recursive estimation procedures. There are several aspects of this modeling that are of particular relevance in distinguishing among the various research endeavors: the number and type of the unknown parameters; the amount and type of data to be used; and the suitability of the model for its intended purpose.

The number of unknown parameters (the number of “degrees of freedom”) is directly related to estimation performance and data requirements. If many parameters are unknown, as is the case with the research presented in this paper, the estimation process can be delicate and difficult, and batch methods may be required to extract as much information as possible from the data. If fewer parameters are involved, such as in [13] and [27], the estimation process is more robust, as evidenced by the successful use of real- or near-real time recursive techniques in these three cases. Similarly, the additional information contained in stereo imagery allows the use of simpler models and solutions as in [17]; a robust linear solution for monocular vision has proved more elusive. For example, an approximate linear method for monocular vision discussed in [34] is used to generate initial guesses for a nonlinear least squares objective function similar to ours, with the intent being to reduce the number of unknown parameters in the iterative optimization. However, this method is tested only on very large objects for simulated imagery and fixed scenes (self-motion) for real imagery (both having very low noise levels as quantified by the measure discussed below), and appears to be limited to motion with small rotation.

Since the problem is nonlinear (central projection imaging and rotational motion), the particular values of the unknown parameters are also important. Pure translational motion and object (or scene) structure are more easily estimated than rotational motion, and the presence of significant rotational motion can lead to estimation problems, especially with regard to convergence, due to qualitative changes in the objective function (creation of multiple local minima). Also, self-motion estimation is more robust than object motion estimation, to the extent that self-motion estimation involves a very large “object” (a fixed scene). The use of long sequences (more than four or five images) gives a substantial improvement in achievable estimation accuracy, and accuracy improves as the object image gets larger. Since these theoretical lower bounds on variances of estimate [8] are independent of any particular

Manuscript received May 31, 1988; revised January 2, 1991. Recommended for acceptance by W. B. Thompson. This work was supported in part by the National Science Foundation under Grant IRI-87-13585 and by the Hughes Aircraft Company Doctoral Fellowship Program.

T. Broida is with Hughes Aircraft Company, Radar Systems Group, P.O. Box 92426, Los Angeles, CA 90009.

R. Chellappa is with the Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089.

IEEE Log Number 9100189.

estimation algorithm, and reflect the information content of the data, they apply to any estimation method using point features with similar parameterization.

Model selection is essentially the process of choosing a suitable coordinate system or parameterization with which to describe the quantities of interest. The choice of a suitable model can facilitate the estimation process, particularly in nonlinear problems. For example, a common model for monocular vision (e.g., [39], and most of the "two-view" solutions) assumes that the axis of "object" rotation passes through the origin of the camera coordinate system. The use of this model in an external object situation complicates the propagation of object motion in time, so that longer sequences of images cannot be easily used to reduce the effects of noise. More flexible models that represent rotational motion about an object-centered axis are used in [13], [26], [35]. The models used in [13], [26] are quite similar to those used in our research, but differ in certain ways: [13] assumes known structure, and [26] uses specific models for different situations (e.g., number of feature points, number of image frames). In [35], this rotational modeling is addressed by combining multiple solutions of the two-view problem using a local conservation of angular momentum (LCAM) model. However, the experiments reported there involve stereo imagery only. The smoothing of successive solutions to the two-view problem to deal with noise in monocular imaging was found to be inadequate in [37], as a result of the instability of the underlying two-view solutions.

Two measurement models are commonly used. "Feature-based" methods rely on the extraction of a set of discrete object features, such as points, lines, and patches of shadow, located in successive images in a sequence, the image coordinates of which are used as data to estimate object motion. An alternative is the use of "optical flow" methods to represent motion in the image plane as sampled, continuous velocity fields [3], [14]. Work such as [1], [14]–[16], [19], [30] demonstrates the usefulness of the optical flow approach, both in terms of the "low-level" problems of motion recognition and segmentation of scenes into their moving and stationary components, and in terms of estimating rigid object motion and structure. A comparison is given in [2]. It appears that both techniques have their uses, and will eventually be unified into a more general vision system. The research reported in this paper is based on the use of discrete features.

In forming estimates from image sequences, either batch or recursive solution methods can be used. A batch method repeatedly processes the available data, and iteratively adjusts parameter estimates until a minimum of an objective function (such as the sum of the squares of the errors, which is the least squares criterion) is reached. A recursive method starts with an initial guess, and refines this guess by considering new data sets (images) one at a time.

The present paper extends the research discussed in [6]–[8], which has been concerned with the simultaneous estimation of object structure and motion, when both the rotational and translational motion can be significant. In [7], a one dimensional (1-D) image of a two-dimensional object (2-D) undergoing 2-D motion was examined, to explore the prop-

erties of central projection imaging and the viability of the object/motion modeling approach. Some knowledge of object structure was assumed, and a recursive solution method was used on simulated data. The results presented there were extended to a 2-D image of a 3-D object, undergoing 3-D motion, and the various models were more fully developed. The accuracy of this model-based approach was addressed in [8], and the performance of the batch algorithm was compared with theoretical limits (Cramér–Rao lower bounds) for various numbers of image frames and feature points, based on simulated data. In [6], a nonlinear extension of the Kalman filter (iterated extended Kalman filter) was shown to be effective for recursive estimation of 3-D motion and structure parameters based on both simulated and real 2-D imagery. A batch algorithm was used for initializing the recursive procedure—the batch approach was shown to be more stable in the presence of inaccurate initial guesses and high noise levels.

The present paper represents a continuation of the research started in [7] by evaluating the performance of a batch algorithm applied to all the data from two sequences of real imagery. Although the batch approach is computationally more expensive than the recursive approach, it is both more accurate and more stable. In addition, the issue of the uniqueness of the estimates for the model-based approach is addressed theoretically (for translational motion) and with a numerical test based on the singular value decomposition (for motion involving rotation).

II. MODELS

The models used in this research are based on several assumptions. First, it is assumed that object motion is "smooth" in an inertial coordinate system. The constraint imposed on the motion is that some finite time derivative (say, the n th) of the variation in each kinematic attribute is constant, and that higher order derivatives are zero. Secondly, it is assumed that object motion can be decomposed into rotation about a point termed the center of rotation, and translation of that center of rotation. In the case of constant angular velocity rotation, the center of rotation is a point on the axis of rotation, while if higher order rotation is present, the center of rotation is the point remaining stationary in the object with respect to rotational motion. Fig. 1 illustrates the basic models for motion, structure, and the observation of the object.

A. Imaging Model

A central projection imaging model is used, defined by

$$h : S \mapsto P \quad (1)$$

where

$$s = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in S = \{(x, y, z)^T \in R^3, \text{ s.t. } z > 0\} \quad (2)$$

is a spatial point coordinate, and

$$p = \begin{pmatrix} X \\ Y \end{pmatrix} \in P = \{(X, Y)^T \in R^2 \text{ s.t. } -A \leq X \leq A;$$

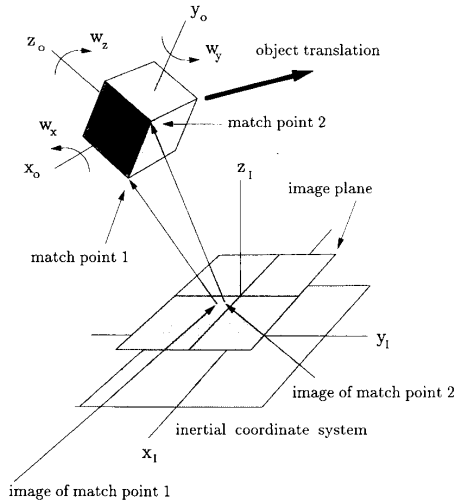


Fig. 1. Fundamental model for object motion and imaging.

$$-B \leq Y \leq B; A, B > 0\} \quad (3)$$

is an image plane point coordinate. The space P is nominally a finite rectangle, corresponding to the image plane of a camera; however, this fact is incidental and is not further discussed. Then,

$$X = f \cdot \frac{x}{z} + n_X, \quad Y = f \cdot \frac{y}{z} + n_Y \quad (4)$$

map spatial coordinates to noisy image coordinates, where the camera focal length f is set to unity without loss of generality for synthetic imagery. When real imagery is involved, the focal length must be measured or estimated, as will be discussed in Section IV. The terms n_X and n_Y are the image plane noise components, discussed below. Thus, the measurement model for a single point $s = (x, y, z)^T$ is

$$\mathbf{p} = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x/z \\ y/z \end{pmatrix} + \begin{pmatrix} n_X \\ n_Y \end{pmatrix} = \mathbf{h}[\mathbf{s}] + \mathbf{n}. \quad (5)$$

B. Noise Model

The measured image coordinates of the feature points are assumed to consist of the image coordinates of the true feature positions corrupted by additive independent zero mean Gaussian noise. The measure of noise power used here is a percentage of the object image size, where this size is taken as the projection in the image plane of the largest chord of the object in 3-D. The usual measure of noise in this type of problem is a percentage of image size; however, the effect of a given noise level of this type depends on object size, since estimation involving a small object image will clearly suffer more than that involving a larger object image for a given image plane noise level. As a result, the noise level increases when an object recedes from the camera, and decreases as it approaches the camera.

A different source of error is the incorrect labeling of a feature point, such as when two closely spaced features are erroneously switched. The impact of such an event would

vary, depending on noise level, number of points, number of frames, distance between feature points, etc. This type of error has received considerable attention in surveillance and tracking situations involving multiple moving objects, such as [4]. The effects of such errors are very difficult to quantify, and tend to be chaotic in nature, a single error often leading to an avalanche of additional errors. However, it has been observed that the chaotic effects tend to slowly disappear after a while, provided that the measurement accuracy, intervals, and point spacings are adequate to avoid frequent events of this type.

C. Object and Motion Model

An object-centered coordinate system is defined. The origin of this object-centered coordinate system is not observed, in general. Object structure is then defined as the coordinates of the object feature points in the object-centered coordinate frame. These positions are constant in time due to the rigidity assumption. Object translational kinematics are defined to be the position and motion of the origin of the object-centered coordinate frame with respect to the camera-centered (inertial) coordinate frame. Object rotational kinematics are defined to be the object angular position and motion about the origin of the object-centered frame. Object structure and translational kinematics can only be known to within a global scale factor, unless absolute *a priori* data is available about the object and/or its translational kinematics. Object rotational kinematics are not subject to this scale factor. The scale factor comes about because, as can be seen in the mapping \mathbf{h} defined in (5), any constant multiple of all spatial coordinates $(x, y, z)^T$ results in the same image.

The following model results: Let $\mathbf{s}_{iO} = (x_i, y_i, z_i)^T$ be the object-centered coordinates of feature point i . Let $\mathbf{s}_R(\underline{u}, t) = (x_R(\underline{u}, t), y_R(\underline{u}, t), z_R(\underline{u}, t))^T$ be the camera-centered (inertial) coordinates of the origin of the translating object reference frame (not observed directly). The vector \underline{u} contains the model parameters. Let $R(\underline{u}, t)$ be the 3×3 coordinate transformation matrix that rotates the object coordinate system from its orientation at time t_0 (aligned with the camera coordinate axes) to its orientation at time t . Let $\mathbf{s}_i(\underline{u}, t)$ be the spatial, camera-centered coordinates of feature point i at time t , with parameter vector \underline{u} .

The object motion model is then given by

$$\mathbf{s}_i(\underline{u}, t) = \mathbf{s}_R(\underline{u}, t) + R(\underline{u}, t)\mathbf{s}_{iO}, \quad (6)$$

or, at time of the k th image, t_k , as

$$\begin{aligned} \mathbf{s}_i(\underline{u}, t_k) &= \begin{pmatrix} x_i(\underline{u}, t_k) \\ y_i(\underline{u}, t_k) \\ z_i(\underline{u}, t_k) \end{pmatrix} \\ &= \begin{pmatrix} x_R(\underline{u}, t_k) \\ y_R(\underline{u}, t_k) \\ z_R(\underline{u}, t_k) \end{pmatrix} + \begin{pmatrix} R_x(\underline{u}, i, t_k) \\ R_y(\underline{u}, i, t_k) \\ R_z(\underline{u}, i, t_k) \end{pmatrix}. \end{aligned} \quad (7)$$

At time t_k the image plane measurements of the match points are, from (5),

$$\mathbf{p}_i(t_k) = \mathbf{h}[\mathbf{s}_i(\underline{u}, t_k)] + \mathbf{n}(t_k), \quad (8)$$

which can be written as

$$\mathbf{p}_i(t_k) = \begin{pmatrix} X_i(t_k) \\ Y_i(t_k) \end{pmatrix} = \begin{pmatrix} \frac{x_i(\underline{u}, t_k)}{z_i(\underline{u}, t_k)} \\ \frac{y_i(\underline{u}, t_k)}{z_i(\underline{u}, t_k)} \end{pmatrix} + \begin{pmatrix} n_X(t_k) \\ n_Y(t_k) \end{pmatrix}, \quad (9)$$

where $i = 1, \dots, M$, for M object match points, and $k = 1, \dots, N$, for N image frames.

1) *Translational Motion Model*: Since the spatial coordinates of the origin of the object-centered reference frame, $\mathbf{s}_R(\underline{u}, t)$, can be written in terms of an arbitrary number of nonzero derivatives, a variety of modeling options are available. Assuming it can be accurately modeled by a constant n th derivative,

$$\mathbf{s}_R(\underline{u}, t) = \mathbf{s}_R(\underline{u}, t_0) + \sum_{k=1}^n \frac{\partial^{(k)} \mathbf{s}_R(\underline{u}, t)}{\partial t^{(k)}} \bigg|_{t=t_0} \frac{(t-t_0)^k}{k!}. \quad (10)$$

For example, if the translational motion has constant velocity (the case examined in detail in this research),

$$\begin{aligned} \mathbf{s}_R(\underline{u}, t) &= \mathbf{s}_R(\underline{u}, t_0) + (t-t_0) \dot{\mathbf{s}}_R(\underline{u}, t_0) \\ &= \begin{pmatrix} x_R \\ y_R \\ z_R \end{pmatrix} + (t-t_0) \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} x_R(\underline{u}, t_k) \\ y_R(\underline{u}, t_k) \\ z_R(\underline{u}, t_k) \end{pmatrix}. \end{aligned} \quad (11)$$

Thus, the translational motion during the observation period is modeled by a finite number ($3n$) of parameters, which are simply the nonzero derivatives at a single point in time (t_0).

2) *Rotational Motion Model*: Quaternions, described for example in [11]–[13], [20], [36], can be used to propagate the rotation matrix $R(\underline{u}, t)$ in time, with the rotation of the object coordinate frame represented by the object rotation rates about inertial (x, y, z) axes at the reference time t_0 as $\underline{\omega}_t = (\omega_x, \omega_y, \omega_z)^T$. At this reference time, the object centered axes are aligned with the inertial axes, so an equivalent statement is that the angular motion occurs on an axis (possibly time-varying) through the origin of the object centered frame, measured about axes parallel to the inertial (camera centered) axes. With this approach, the rotation matrix $R(\underline{u}, t)$ can be written in terms of the unit quaternion $q(t) = (q_1(t), q_2(t), q_3(t), q_4(t))^T = \underline{q}(\underline{u}, t)$ as discussed in [12], [20], [36]. Suppressing the time dependency of the quaternion elements,

$$R(\underline{u}, t) = \begin{pmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 & 2(q_1 q_2 + q_3 q_4) & 2(q_1 q_3 - q_2 q_4) & 2(q_2 q_3 + q_1 q_4) \\ 2(q_1 q_2 - q_3 q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2(q_1 q_4 + q_2 q_3) & -2(q_1 q_3 + q_2 q_4) \\ 2(q_1 q_3 + q_2 q_4) & -2(q_1 q_4 - q_2 q_3) & q_1^2 - q_2^2 + q_3^2 + q_4^2 & 2(q_2 q_3 - q_1 q_4) \\ 2(q_2 q_3 - q_1 q_4) & 2(q_1 q_4 + q_2 q_3) & 2(q_1 q_3 - q_2 q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 \end{pmatrix}. \quad (12)$$

The rotational portion of the motion and object model can then be written

$$\begin{pmatrix} R_x(\underline{u}, i, t_k) \\ R_y(\underline{u}, i, t_k) \\ R_z(\underline{u}, i, t_k) \end{pmatrix} = R(t_k) \mathbf{s}_{iO} = R[q(t_k)] \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}. \quad (13)$$

The unit quaternion $\underline{q} = (q_1, q_2, q_3, q_4)^T$ is related to “standard” expressions of the angular relation between coordinate systems by

$$\underline{q} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = \begin{pmatrix} n_1 \sin \theta/2 \\ n_2 \sin \theta/2 \\ n_3 \sin \theta/2 \\ \cos \theta/2 \end{pmatrix} \quad (14)$$

where (n_1, n_2, n_3) are the direction cosines of an axis of rotation, and θ is the angle about that axis that describes the rotation of the object coordinate system from its initial orientation to its orientation at time t . In general, n_1, n_2, n_3 , and θ all change with time. It is termed a unit quaternion because $|q| = 1$. The quaternion \underline{q} propagates in time according to the differential equation [12], [36]

$$\dot{\underline{q}}(t) = \Omega(\underline{\omega}_t) \underline{q}(t), \quad \underline{q}(t_0) = \underline{q}_0 \quad (15)$$

where

$$\Omega(\underline{\omega}_t) = \frac{-1}{2} \begin{pmatrix} 0 & -\omega_z & \omega_y & -\omega_x \\ \omega_z & 0 & -\omega_x & -\omega_y \\ -\omega_y & \omega_x & 0 & -\omega_z \\ \omega_x & \omega_y & \omega_z & 0 \end{pmatrix}_t. \quad (16)$$

The solution to (15) when $\underline{\omega}$ is constant is simply

$$\underline{q}(t) = \exp[(t-t_0)\Omega] \underline{q}(t_0). \quad (17)$$

It might also be noted that the matrix $2i\Omega/|\underline{\omega}|$ is unitary, where $i = \sqrt{-1}$. As a result, the power series expansion for the matrix exponential can be reduced to [33]

$$\underline{q}(t) = \left[\cos(|\underline{\omega}|t/2) \mathbf{I}_4 + \frac{2}{|\underline{\omega}|} \sin(|\underline{\omega}|t/2) \Omega \right] \underline{q}(t_0) \quad (18)$$

which can be further reduced by assuming, without loss of generality, that the coordinate systems are aligned at t_0 , so that $\underline{q}(t_0) = (0, 0, 0, 1)^T$, and

$$\underline{q}(t) = \begin{pmatrix} \frac{\omega_x}{|\underline{\omega}|} \sin(|\underline{\omega}|t/2) \\ \frac{\omega_y}{|\underline{\omega}|} \sin(|\underline{\omega}|t/2) \\ \frac{\omega_z}{|\underline{\omega}|} \sin(|\underline{\omega}|t/2) \\ \cos(|\underline{\omega}|t/2) \end{pmatrix} \quad (19)$$

[compare to (14)]. This solution is valid only when $\underline{\omega}$ is constant, as it is for the test cases presented in this paper. When $\underline{\omega}(t)$ evolves in time according to a constant rate of precession, the quaternion can also be written in closed form [38]. The resulting closed form expression for the quaternion as a function of time is given in [38, Appendix A], when $\underline{\omega}$ moves with constant precession.

The initial quaternion $\underline{q}(t_0)$ can be written as above since there is no requirement for any particular initial orientation of the object centered coordinate system, with respect to the inertial coordinate system. The first parameters of interest are the changes in this relationship, as a function of time, which are the angular rates. The remaining unknown parameters, the coordinates of the feature points in the object centered reference frame, are important only in their relationship to each other, so the initial orientation of the object

centered coordinate system is in this sense also arbitrary. The above choice simply aligns the object centered system so that the x - and y -axes of the object-centered coordinate system are parallel to the X - and Y -axes of the image plane at time t_0 (which are also parallel to the inertial coordinate axes).

In general, the image of the match points at time t_k can be written, from (9) as

$$\begin{aligned} \mathbf{p}_i(t_k) &= \begin{pmatrix} x_i(\underline{u}, t_k) \\ y_i(\underline{u}, t_k) \\ z_i(\underline{u}, t_k) \end{pmatrix} + \begin{pmatrix} n_X(t_k) \\ n_Y(t_k) \end{pmatrix} \\ &= \begin{pmatrix} x_R(\underline{u}, t_k) + R_x(\underline{u}, i, t_k) \\ z_R(\underline{u}, t_k) + R_z(\underline{u}, i, t_k) \\ y_R(\underline{u}, t_k) + R_y(\underline{u}, i, t_k) \end{pmatrix} + \begin{pmatrix} n_X(t_k) \\ n_Y(t_k) \end{pmatrix}. \end{aligned} \quad (20)$$

When motion is more general than constant velocity, the translational component is straightforward, giving terms like $x_R + \dot{x}t_k + \ddot{x}(t_k^2/2)$, to the desired order. The rotational component is slightly more involved; however, as shown, there are closed form expressions for rotation involving constant angular velocity, and for rotations in which the axis of rotation changes at a constant precessional rate. Higher order rotation becomes more laborious, if it should be necessary, requiring numerical integration of (15), and the use of a Taylor series expansion of $\underline{\omega}(t)$ to compute the values of $\omega_x(t)$, $\omega_y(t)$, and $\omega_z(t)$ to be used in the matrix $\Omega(\underline{\omega}_i)$. The ordinary differential equation of (15) is well behaved, and in [36] an incremental approximation is discussed, that avoids the need for numerical integration. At each iteration of the attitude state equations, an average value of $\underline{\omega}(t)$ over the interval is used in (18), to give an approximate closed form single-step predictor. This amounts to a "zero-order" approximation, or a first order Euler quadrature formula with only a single step. Thus, the propagation of the state equations when the rotational motion is not constant velocity or constant precession is not overly difficult.

When the models discussed above are applied to the monocular vision problem, there is an unknown scale factor applied to all translational and structural states, as discussed above. In the research reported here, this has been taken into account by normalizing all affected states by the parameter z_R . Although any translational or structural parameter would suffice (\dot{z} , x_1 , etc.), this choice is appealing in that the normalized coordinates of the origin of the object-centered frame x_R/z_R and y_R/z_R are the actual image plane coordinates of these points (not directly observed in general). This gives rise to a normalized parameter vector \underline{u} , which will be used in the following discussions. That is,

$$\underline{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \mathbf{u}_4 \end{pmatrix}, \quad (21)$$

where

$$\mathbf{u}_1 = \begin{pmatrix} x_R/z_R \\ y_R/z_R \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} \dot{x}_R/z_R \\ \dot{y}_R/z_R \\ \dot{z}_R/z_R \\ \vdots \\ x^{(n)}/z_R \\ y^{(n)}/z_R \\ z^{(n)}/z_R \end{pmatrix}, \quad (22)$$

$$\mathbf{u}_3 = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \\ \dot{\omega}_x \\ \dot{\omega}_y \\ \dot{\omega}_z \\ \vdots \\ \omega_x^{(m-1)} \\ \omega_y^{(m-1)} \\ \omega_z^{(m-1)} \end{pmatrix}, \quad \text{and} \quad \mathbf{u}_4 = \begin{pmatrix} x_1/z_R \\ y_1/z_R \\ z_1/z_R \\ \vdots \\ x_M/z_R \\ y_M/z_R \\ z_M/z_R \end{pmatrix}. \quad (23)$$

The parameter vector \underline{u} has dimension $3M + 3n + 3m + 2$. When rotational motion has constant velocity, there is a floating degree of freedom with regard to the inertial coordinates of the center of rotation: one coordinate of one feature point is then selected to fully define the origin of the object-centered coordinate system along the axis of rotation. This reduces the dimensionality of the parameter vector to $3M + 3n + 3m + 1$. In a similar way, the complete absence of rotation not only eliminates the rotational states, but leaves the origin of the object centered coordinate system completely undefined. Then, one of the feature points can be selected as the origin, with the elimination of additional unknown parameters as discussed in [5].

In [29], [33], a geometric approach has been suggested for the structure from motion problem, under the assumptions of rigidity, constant velocity translation and constant angular velocity rotation for the orthographic projection case. Orthographic projection [33] enables more complicated rotation models to be considered. These methods are applicable to one of the real image sequences (bottle sequence) considered in this paper. But in general perspective model based motion analysis is much more general and harder than the one based on orthographic projection.

III. FORMULATION FOR BATCH SOLUTION

We present a batch approach for the estimation of the model parameters, when both the translation and rotational motion are of constant velocity. As discussed above, the parameters x_R , y_R , and z_R are defined to be the spatial coordinates of the origin of the object-centered coordinate system at time t_0 , the parameters \dot{x} , \dot{y} , and \dot{z} are the translational rates, and x_i , y_i , and z_i are the coordinates of the object feature point i in the object-centered coordinate system. However, there is no way to distinguish a large, distant object, moving quickly, from a small, nearby object, moving proportionately more slowly, in the absence of *a priori* information about the true value of one of these parameters. This scale factor has been taken into

account by dividing each of these parameters by z_R , which is equivalent to setting $z_R = 1$, as discussed in Section II.

Thus, if both the translational and rotational rates are constant during the image sequence, the vector \underline{u} of unknown parameters is

$$\underline{u} = \begin{pmatrix} x_R/z_R \\ y_R/z_R \\ \dot{x}/z_R \\ \dot{y}/z_R \\ \dot{z}/z_R \\ \omega_x \\ \omega_y \\ \omega_z \\ x_1/z_R \\ y_1/z_R \\ x_2/z_R \\ y_2/z_R \\ z_2/z_R \\ \vdots \\ x_M/z_R \\ y_M/z_R \\ z_M/z_R \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \\ u_{11} \\ u_{12} \\ u_{13} \\ \vdots \\ u_{3M+5} \\ u_{3M+6} \\ u_{3M+7} \end{pmatrix}. \quad (24)$$

It should be noted that the z -component of the first feature point is not included. This results from the fact that the origin of the object centered coordinate system is constrained by constant velocity rotation only to lie on an axis (the axis of rotation): the position of the origin along this axis is arbitrary. Thus, it is convenient simply to assign one coordinate of any feature point to an arbitrary value: in this case $z_1/z_R = 0$ was used. As long as the axis of rotation is not parallel or almost parallel to the x - y plane, this approach works. In cases where this is not appropriate, setting the x - and y -component to a fixed value would be better. For generality, a strategy of constraining the origin to lie in a plane orthogonal to the axis of rotation would be better, by including a constraint that forces

$$\omega_x x_1/z_R + \omega_y y_1/z_R + \omega_z z_1/z_R = 0 \quad (25)$$

as part of the objective function. The need to artificially constrain the origin of the object centered system also arises when motion is purely translational: in that case, the origin is entirely arbitrary, since all points move in 3-D along parallel translational velocity vectors. Then, a single point is arbitrarily picked as the origin, and all three components are removed from the parameter vector. If angular acceleration is of significant magnitude, only a single point (instead of an axis) remains invariant to rotational motion, which would eliminate the need for such a constraint. Such an implicit definition of the origin of this coordinate system occurs under the assumption stated above, namely that the axes of rotational acceleration and velocity intersect.

As discussed in Section II, the individual components of $\mathbf{s}_i(\underline{u}, t_k)$ of (6) are (setting $t_0 = 0$),

$$\mathbf{s}_i(\underline{u}, t_k) = \begin{pmatrix} x_i(\underline{u}, t_k) \\ y_i(\underline{u}, t_k) \\ z_i(\underline{u}, t_k) \end{pmatrix}$$

$$= \begin{pmatrix} x_R/z_R + t_k \dot{x}/z_R + R_x(\underline{u}, i, t_k) \\ y_R/z_R + t_k \dot{y}/z_R + R_y(\underline{u}, i, t_k) \\ 1 + t_k \dot{z}/z_R + R_z(\underline{u}, i, t_k) \end{pmatrix}. \quad (26)$$

The terms $R_x(\underline{u}, i, t_k)$, etc., refer to the x -, y -, and z -components of the camera (inertial) coordinates of the i th match point, which are expressed in the rotated (object-centered) coordinate system as $(x_i/z_R, y_i/z_R, z_i/z_R)$. The rotation matrix $R(\underline{u}, t_k)$ of (12) is a function of u_6, u_7, u_8 ($\omega_x, \omega_y, \omega_z$), and t_k . For example, denoting the rs component of $R(\underline{u}, t)$ as R_{rs} ,

$$R_x(\underline{u}, i, t_k) = R_{11}x_i/z_R + R_{12}y_i/z_R + R_{13}z_i/z_R \quad (27)$$

$$= (q_1^2 - q_2^2 - q_3^2 + q_4^2)x_i/z_R \quad (28)$$

$$+ 2(q_1q_2 + q_3q_4)y_i/z_R \\ + 2(q_1q_3 - q_2q_4)z_i/z_R \quad (29)$$

when the time argument of $q(t)$ has been suppressed. The 1 in the z -component of $\mathbf{s}_i(\underline{u}, t_k)$ is written as z_R/z_R , in an analogous manner to the terms $u_1 = x_R/z_R$ and $u_2 = y_R/z_R$, as discussed above. All components of the object structure and translational kinematics are then homogeneous in $1/z_R$, which accounts for the global scale factor.

Next, the components of (9) are written as

$$\begin{aligned} X_i(t_k) &= h_X[\mathbf{s}_i(\underline{u}, t_k)] + n_X(t_k) \\ &= \frac{x_i(\underline{u}, t_k)}{z_i(\underline{u}, t_k)} + n_X(t_k) \\ &= \frac{u_1 + t_k u_3 + R_x(\underline{u}, i, t_k)}{1 + t_k u_5 + R_z(\underline{u}, i, t_k)} + n_X(t_k), \end{aligned} \quad (30)$$

and

$$\begin{aligned} Y_i(t_k) &= h_Y[\mathbf{s}_i(\underline{u}, t_k)] + n_Y(t_k) \\ &= \frac{y_i(\underline{u}, t_k)}{z_i(\underline{u}, t_k)} + n_Y(t_k) \\ &= \frac{u_2 + t_k u_4 + R_y(\underline{u}, i, t_k)}{1 + t_k u_5 + R_z(\underline{u}, i, t_k)} + n_Y(t_k). \end{aligned} \quad (31)$$

Then, if the noise terms $n_X(t_k)$ and $n_Y(t_k)$ are assumed to be independent, identically distributed (IID), $\mathcal{N}(0, \sigma^2)$,

$$n_X(t_k) = X_i(t_k) - h_X[\mathbf{s}_i(\underline{u}, t_k)] \sim \mathcal{N}(0, \sigma^2), \quad (32)$$

$$n_Y(t_k) = Y_i(t_k) - h_Y[\mathbf{s}_i(\underline{u}, t_k)] \sim \mathcal{N}(0, \sigma^2). \quad (33)$$

This leads to the conditional density of the measurements given \underline{u} , \mathbf{s}_{iO} , and t_k ,

$$\begin{aligned} f(X_i(t_k), Y_i(t_k) | \underline{u}, \mathbf{s}_{iO}, t_k) &= \frac{1}{2\pi\sigma^2} \exp \\ &\cdot \left(-\frac{(X_i(t_k) - h_X[\mathbf{s}_i(\underline{u}, t_k)])^2 + (Y_i(t_k) - h_Y[\mathbf{s}_i(\underline{u}, t_k)])^2}{2\sigma^2} \right). \end{aligned} \quad (34)$$

Since the noise is assumed uncorrelated in time, and independent from feature point to feature point, the conditional density of the data $Z = \{X_{ik}, Y_{ik}, i = 1, \dots, M; k = 1, \dots, N\}$ given \underline{u} is the product

$$f(Z | \underline{u}) = \prod_{k=1}^N \prod_{i=1}^M f(X_i(t_k), Y_i(t_k) | \underline{u}, t_k). \quad (35)$$

Then, the maximum likelihood (ML) estimate $\hat{\underline{u}}$ of \underline{u} is found by maximizing the log-likelihood of (35), which reduces to minimizing the summed squared residuals with respect to \underline{u} ,

$$G(\underline{u}) = \sum_{k=1}^N \sum_{i=1}^M \left\{ (X_i(t_k) - h_X[s_i(\underline{u}, t_k)])^2 + (Y_i(t_k) - h_Y[s_i(\underline{u}, t_k)])^2 \right\}. \quad (36)$$

A variety of nonlinear search algorithms were tried, including the method of Steepest Descent, Powell's Sum of Squares, the Davidon-Fletcher-Powell method [24], and all of the routines in the IMSL package. Of these, the Conjugate Gradient descent technique from the IMSL package gave by far the best performance. The IMSL implementation of the Conjugate Gradient algorithm is based on a study by Powell in [21]. Explicit computation of the gradient is required, but no second derivatives are calculated. The Conjugate Gradient, as its name implies, is based on searching along a sequence of mutually orthogonal directions. The main feature of this implementation is that a method of periodic restarts of the search sequence is proposed, using as an additional term a partial single step in the direction of steepest descent. By using an adaptive test, based on the event of $g_k^T g_{k+1} / \|g_{k+1}\|^2 > c$, for some constant c , the restart is applied so as to avoid certain difficulties and to exploit any underlying quadratic or symmetric structure to accelerate convergence. Essentially, the test examines the degree of orthogonality present in two consecutive search directions. This technique has been used for all the results generated in this paper.

Recently, [31] an algorithm based on the Levenberg-Marquardt method has been suggested for minimizing (34). It is claimed that this algorithm converges faster than the conjugate gradient descent algorithm.

IV. EXPERIMENTAL RESULTS FROM REAL IMAGERY

The results of the batch algorithm discussed in Section III, as applied to real imagery, are summarized. The images used here were made with a standard 35 mm camera, set on a tripod for stability. Images were made one at a time, and the object was moved between images. Enlargements were made of the 35 mm negatives, and image plane coordinates were measured with a ruler, directly from the prints. The image plane coordinates must be measured with respect to an orthogonal coordinate system, however. To do this, two fixed reference points were located in the image sequence, that appear in every image. The distances from each reference point to each feature point were measured, and a simple trigonometric relation was applied to transform the data into coordinates along the baseline (the line connecting the two reference points) and perpendicular to this baseline. This data was input to a computer, and estimates were formed using the batch estimation model discussed in Section III, using the IMSL Conjugate Gradient routine.

The focal length f relates the actual image dimensions to both the estimates and ground truth. The distance z_R is used to account for the floating scale factor that arises in monocular

imagery using the central projection model. This distance can be estimated only when information about ground truth is available, as it is in the test cases presented in this section.

When working with real images, the camera focal length is not simply the focal length of the camera, but instead is the "effective" focal length of the entire imaging system. If measurements are made in pixels from a digitized image with a track-ball, the resolution of the image is part of the imaging system. If, as in this work, the measurements are made with a ruler on a print, the degree of enlargement is part of the imaging system. As a result, it is necessary to perform a calibration step, to determine the effective focal length of the imaging system.

Consider the following (simplified) objective function, where the sequences $\{x_n\}$, $\{y_n\}$, and $\{z_n\}$ indicate the time evolution of the inertial coordinates of a set of feature points,

$$G(\underline{u}) = \sum_{n=1}^{NM} \left(X_n - f \frac{x_n}{z_n} \right)^2 + \left(Y_n - f \frac{y_n}{z_n} \right)^2. \quad (37)$$

The summation is over all feature points (M) and images (N). The sequences $\{X_n\}$ and $\{Y_n\}$ are given as data; parameters \underline{u} that minimize G are desired, that are related to the sequences $\{x_n\}$, $\{y_n\}$, and $\{z_n\}$ according to the models presented earlier. The approach used here was to estimate f as a parameter, along with the other unknown terms in \underline{u} . The derivative of G with respect to f is

$$\frac{\partial G(\underline{u})}{\partial f} = -2 \sum_{n=1}^{NM} \left\{ \left(X_n - f \frac{x_n}{z_n} \right) \frac{x_n}{z_n} + \left(Y_n - f \frac{y_n}{z_n} \right) \frac{y_n}{z_n} \right\}, \quad (38)$$

which is appended to the gradient vector discussed in Section III. Since the Conjugate Gradient (CG) algorithm converges faster if the states to be estimated are of approximately the same magnitude [24], a parameter $\alpha f'$ was used instead, resulting in an additional coefficient of α multiplying the summation. When $\alpha = 10$ was used, the CG routine required approximately 4000 iterations in the Bottle experiment, with f initialized at about 0.92 inches, as opposed to about 5500 with $\alpha = 1$, and f' initialized at 9.2 inches. The initial estimate of 9.2 inches resulted from a simple scaling of the distance between two feature points from the sequence. Both iterations converged to the same estimate, $\hat{f} = \alpha \hat{f}' = 14.767$ inches. The "correct" answer is unknown. In addition, the parameter estimates $\hat{\underline{u}}$ were identical in both cases (to at least 5 decimal digits), and further were identical to the results given by the CG routine when f was given as an input parameter, although many fewer iterations were required in the latter case.

The parameter estimation procedure does not appear to be highly sensitive to the choice of f : when estimates were made with the guess of 9.2 inches, they were slightly poorer for some parameters (translation rates) but slightly better in others (rotation rates). In general, as the estimate of f is reduced, the translational and structure state estimates tend to get larger, so as to better fit the data. The rotational estimates stay more or less constant. The CG routine converges for fixed values of f as small as 4.6 inches, which was the lowest tried, although

the large biases in the state estimates prevent the objective function from getting very small.

In the Car experiment, the guess for $\alpha f'$ was 12 inches, and the estimate was 23.5 inches; the search was terminated when the maximum number of iterations (4000) were achieved. Out of curiosity, 20 000 iterations were allowed, with the result that search did not converge: instead, the estimate for f grew slowly but steadily, while the objective function was gradually reduced. The estimate of the translational motion in depth (orthogonal to the image plane) is approximately the same with $\hat{f} = 23.5$ and with $\hat{f} = 45.2$, while the estimates of the other translational components are scaled by a factor of about 2. The estimates of the z -components of the structure states also remain about the same, while the other components of these states are scaled by a factor of about two. It appears that allowing f to vary as a parameter results in an ambiguity in terms of object distortion in this case, in that the components of states orthogonal to the image plane remain constant, while those parallel to this plane change, while still remaining consistent with the data, as reflected in the objective function. The same effect has been observed in experiments involving pure translation, when the motion is confined to the z -direction. That such an ambiguity exists in the Car experiment and certain types of pure translation, but does not exist in the Bottle experiment, implies that the particular object structure and motion are involved. Thus, it appears that an examination of camera calibration techniques is required, to resolve the issue of focal length estimation and explain the phenomena observed here.

The second issue is the relationship between the estimates and ground truth. These are related by the global scale factor, taken here to be z_R . Having established the correspondences and processed the data, the results are estimates of *normalized* states, which are related to the ground truth by the factor z_R . In order to compute z_R , the 3-D normalized distance between each pair of feature points is computed; the ground truth distance between each pair of feature points is measured directly. By dividing the ground truth distances by the normalized estimates, an estimate is obtained of z_R . Discarding the largest and smallest estimates, an average of the remaining scale factors is computed, which is used to reconstruct the translational motion estimates and object feature coordinates. In the Bottle experiment, for example, the estimate of z_R is 49.8", and if the minimum and maximum are discarded, the sample standard deviation of this estimate is about 1.3". In the Car experiment, the estimate of z_R is 94.4", with a standard deviation of 41", reflecting the higher noise level and poorer observability in the latter experiment. The use of different estimates for the imaging system focal length results in different estimates for z_R ; the estimates of translational motion parameter values seem to scale very closely with the choice of f , consistent with the increase/decrease in the structure and translational motion state estimates.

A. Summary

The results of both experiments involving real imagery compare well to the ground truth. The data derived from the Car imagery is of significantly lower quality than that of the

Bottle imagery, with a noise level of about 3% compared with about 0.45%, based on "measurement" noise $\hat{\sigma}_n$ of 0.09 inches versus 0.02 inches, and object image sizes of 3 inches versus 4.5 inches. The "measurement" noise level is estimated from the summed squared residuals, and includes the effects of mismodeling and distortion as well as actual errors in the measured distances. The Car sequence involves more mismodeling than does the Bottle, and the focal length estimation is qualitatively different, since it diverges in the Car experiment, but converges nicely in the Bottle experiment.

The structure errors are larger for the Car than for the Bottle (errors up to 10 inches versus errors up to 0.6 inches). The rotational and translational motion estimates are surprisingly good in both experiments. The Bottle experiment resulted in errors of 5.4 degrees and 0.3 inches, while the Car experiment yielded total motion errors of 6.3 degrees total rotational and 1.5 inches total translational motion.

B. Rolling Bottle

This experiment involved a standard 5 gallon water bottle, rolling across a garage floor. The bottle was "instrumented" by affixing adhesive circles at various points to serve as feature points: 12 images were made of 7 feature points. The bottle was rolled 1 inch along the floor between images, and thus traveled 11 inches in all. There was no assumption made about object transparency: in no case was a feature point observed through the bottle. However, there was no occlusion either. For simplicity, the total rotation was limited so that all features were observed in all images. This was done more for convenience than for any other reason; however, it limited the allowable total motion, since the rotation was limited. The total rotation was 2.05 radians, or about 117.7 degrees in all, which was slightly less than 11 degrees between each pair of images.

As mentioned, the estimate of z_R was 48.8 inches, with a standard deviation (trimmed data) of 1.3 inches. This distance was measured when the data were collected; however, the measurement (about 40") differed from the above estimate.

The image plane trajectories of the feature points are illustrated in Fig. 2; the axes are measured in inches. It is easy to see that the features at the bottom of the figure, starting out at $X_{ik} \approx 1"$ to $2"$, are nearly in contact with the ground during the first four or five images. If a feature point is on the edge of a rolling object, such that it contacts a surface periodically, the image plane motion has a cusp every time it touches the ground; in the case of a cylinder, the motion generated is a projection of a cycloid. This event contradicts the notion that "smooth" motion in 3-D results in "smooth" motion in the image plane, an assumption made in [25], for example. It is true, however, that the motion in time of such a point is slowly changing, in that it decelerates and accelerates smoothly. Figs. 3 and 4 show four images from the Bottle sequence.

Table I gives the normalized state estimates, as output by the CG algorithm, expressed in normalized units (consistent with the estimates). The state estimates are related to actual 3-D coordinates by the factor z_R , except for the rotational rates which remain unchanged.

The estimate of the total rotation of the bottle during the sequence is 2.150 radians, compared to a measured rotation

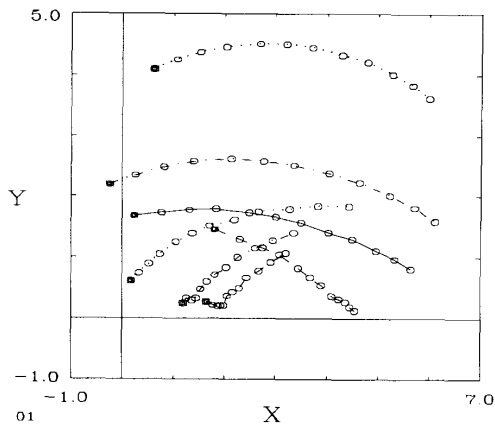


Fig. 2. Observed trajectories of feature points in image plane, Bottle. The darker rectangles indicate the initial feature point positions.

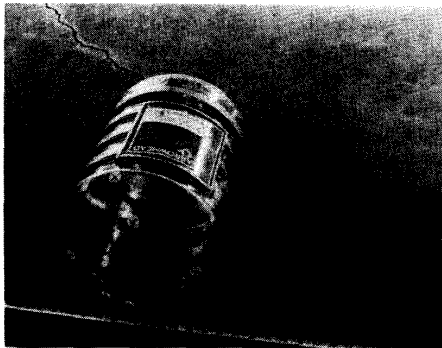


Fig. 3. Frames 1 and 4 from the Bottle sequence of 12 frames.

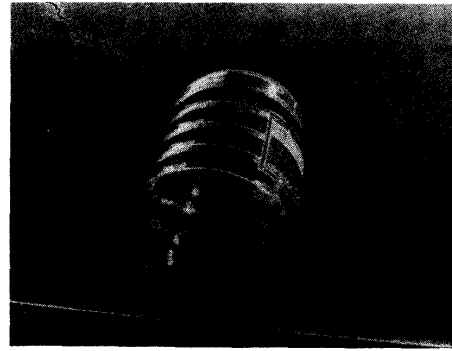


Fig. 4. Frames 7 and 11 from the Bottle sequence of 12 frames.

TABLE I
STATE ESTIMATES FOR THE BOTTLE EXPERIMENT. TWELVE
FRAMES WERE USED WITH SEVEN FEATURE POINTS IN EACH
FRAME. ALL VALUES ARE GIVEN IN NORMALIZED UNITS,
EXCEPT THE ROTATION RATES, WHICH ARE IN RADIANS/SECOND.

State	Estimate	State	Estimate
x_R/z_R	0.0627	y_3/z_R	-0.0741
y_R/z_R	0.1083	z_3/z_R	0.0316
\dot{x}/z_R	0.1997	x_4/z_R	-0.0759
\dot{y}/z_R	-0.0047	y_4/z_R	-0.0071
\dot{z}/z_R	0.0033	z_4/z_R	-0.0603
ω_x	0.4788	x_5/z_R	0.0173
ω_y	1.5196	y_5/z_R	-0.0109
ω_z	1.1325	z_5/z_R	-0.0924
x_1/z_R	0.0424	x_6/z_R	-0.1048
y_1/z_R	-0.0859	y_6/z_R	0.0253
x_2/z_R	0.0155	z_6/z_R	-0.0214
y_2/z_R	-0.0919	x_7/z_R	-0.0444
z_2/z_R	0.0217	y_7/z_R	0.1857
x_3/z_R	-0.0583	z_7/z_R	0.0986

of 2.055 radians, based on the circumference and distance rolled, so that the error is about 0.095 radians, or 5.4° . The measured circumference of the bottle, at the largest point, is 33.63 inches, and the bottle rolled 11 inches during the sequence.

The total translation is estimated by computing the magnitude of the normalized translation rate, 0.200 units/second,

times 1.1 seconds, yielding 0.220 units. The normalizing factor $z_R \approx 48.8''$, so the total distance estimate is 10.73 inches, as compared with an actual distance of 11 inches.

The structural state estimates have standard deviations between 0.12 and 0.28 inches, for each coordinate. The average error (averaged over all points) is 0.06 inches, with a standard deviation of 0.28 inches. It should be noted that these errors

TABLE II
3-D DISTANCES BETWEEN OBJECT FEATURE POINTS FOR THE BOTTLE
EXPERIMENT. NORMALIZED ESTIMATES, TRUE DISTANCES (INCHES),
SCALED ESTIMATES (INCHES), AND SCALE FACTORS. *INDICATES
SMALLEST AND LARGEST ESTIMATES, TRIMMED FROM DATA.

i	j	Estimate	True Distance	Scaled Estimate	\hat{z}_R
1	2	0.0351	1.625	1.71	*46.28
1	3	0.1062	5.000	5.18	47.08
1	4	0.1544	7.375	7.53	47.77
1	5	0.1216	5.813	5.93	47.80
1	6	0.1857	8.915	9.06	48.01
1	7	0.3017	15.330	14.72	50.81
2	3	0.0765	3.630	3.73	47.45
2	4	0.1492	7.183	7.28	48.14
2	5	0.1400	6.750	6.83	48.21
2	6	0.1734	8.368	8.46	48.26
2	7	0.2942	15.029	14.35	51.08
3	4	0.1150	5.438	5.61	47.29
3	5	0.1584	7.688	7.73	48.54
3	6	0.1219	5.970	5.95	48.97
3	7	0.2687	13.785	13.11	*51.30
4	5	0.0987	4.875	4.82	49.39
4	6	0.0583	2.785	2.85	47.77
4	7	0.2519	12.679	12.29	50.33
5	6	0.1458	7.130	7.11	48.90
5	7	0.2810	14.342	13.71	51.04
6	7	0.2092	10.500	10.21	50.19

are highly correlated. Table II lists the true and estimated interpoint distances.

Based on an objective function value of 0.0762 at the minimum, and given $N = 12$, $M = 7$, the image plane noise standard deviation is estimated to be about 0.021 inches. The sample variance of the residuals in the image plane is just the value of the objective function at the minimum divided by the number of residuals ($2NM$); this is simply because the estimates are least squares. The maximum chord in 3-D of the object is about 15 inches, and multiplied by $f/\hat{z}_R \approx 0.302$ (units of image plane inches to 3-D inches) gives approximately 4.54 inches in the image plane (a ruler gives 4.4 inches). The noise level, measured as a ratio of image plane noise level to maximum projected chord of the object, is $0.021/4.5$ or about 0.45%, which is fairly low, but far from being noise-free. This is consistent with the accuracy of the parameter estimates. Approximately 1600 iterations of the Conjugate Gradient algorithm were required.

The axis of rotation is actually perpendicular to the velocity vector \hat{d} , since the bottle is rolling; the computed angle between the two vectors, based on state estimates, is only 76 degrees, so there is some distortion in the estimates. The points p_6 and p_7 are located on a vector that is almost exactly parallel to the axis of rotation: the vector d_{67} connects them. The inner product of d_{67} and ω , normalized by their magnitudes, yields -0.895 , so the estimates of object structure and orientation of the axis of rotation are not truly parallel, but not too far off at -154 degrees. Similarly, d_{67} and \hat{d} form an angle of 104 degrees, again slightly off a right angle.

C. Car Sequence

The second experiment using real imagery involves randomly selected points on the side of the tire of a car approach-

ing the camera. Sixteen images were made, with eight feature points per frame. The feature points were again marked with adhesive dots, in order to facilitate the measurement process. The car was moved approximately 3 inches between each image, corresponding to a tire rotation of about 14.8 degrees. The direction of translational motion was towards the camera (motion in the negative z direction in inertial-camera centered-coordinates), with a component to the right (positive x direction). The camera was slightly higher than the tire, which resulted in a small component of translation in the negative y dimension. Obviously, the relationship between the rotation and translation is the same here as it is in the Bottle experiment, since both involve rolling motion. However, the experiments differ in several ways, as seen in the input data, plotted in Figs. 2 and 5. The object structure in the Bottle experiment involved feature points in a 3-D configuration; in this experiment, the points lie in a single plane, orthogonal to the axis of rotation, parallel to the direction of translation. The amount of motion is greater in this experiment, with a total translation of 45 inches, compared to 11 inches in the Bottle experiment, and a total rotation of 3.85 radians (about 220 degrees), compared to 117 degrees. As an aside, the estimates in the Car experiment are referenced to the end of the image sequence, while in the Bottle experiment the estimates were referenced to the beginning of the sequence; this issue is discussed in [5].

The object image size (size of the image of the tire) is about 2 inches at the start of the sequence, and about 3 inches at the end; the object image size of the bottle was about 4.5 inches during the entire sequence. The measure of noise level used here is the estimated image plane noise $\hat{\sigma}_n$ divided by the projection in the image plane of the largest chord of the object in 3-D. Thus, the noise level in the Car experiment is much higher than in the Bottle experiment: with a 3 inch image, with $\hat{\sigma}_n \approx 0.089$, the noise level is about 3% of the image size, or almost an order of magnitude greater than that of the Bottle experiment. This image plane noise σ is again computed from the summed squared residuals, since this is just the value of the objective function at the minimum. In this case, this value is 2.033; divided by $2NM$ this yields 0.089 inches, or almost a tenth of an inch. The actual measurement noise, which would result from errors in making the measurements, is estimated to be 0.03 to 0.04 inches. The remainder of the noise is due to modeling errors and distortion.

Mismodeling is present in this case, in several ways. First, the motion of the object is assumed to be uniform from image to image. This is not the case. There is an additional image in this sequence, that occurs before the start of the 16 image sequence, that was discarded. Using this additional image did not improve estimation accuracy; in fact, the estimated noise level increased from 0.089 to 0.099, since the motion in the additional interframe interval differed significantly from the others. Thus, there is some nonuniformity in object motion between images. Secondly, the rotational motion is assumed to be constant velocity; however, the wheel was turned slightly during the sequence. Finally, the imaging system focal length could not be unambiguously determined in this case, as discussed above, and the estimates are potentially distorted as a result.

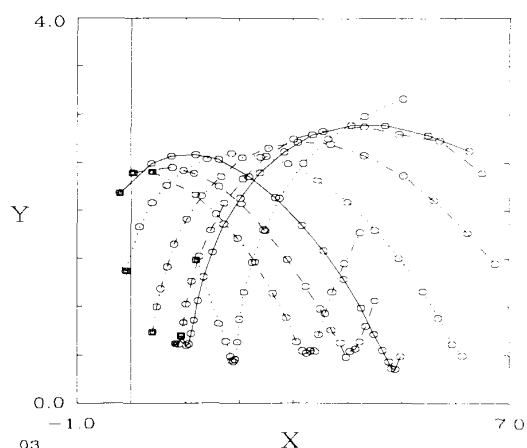


Fig. 5. Observed trajectories of feature points in image plane, for Car imagery. The darker rectangles indicate the initial feature point positions.

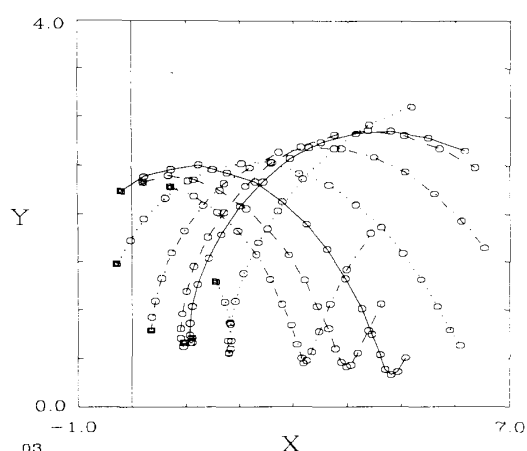


Fig. 6. Reconstructed trajectories of feature points in image plane from estimated parameters, Car. The darker rectangles indicate the initial feature point positions.

Figs. 5 and 6 show the measured and reconstructed image plane trajectories of the feature points. Four images taken from the Car sequence are in Figs. 7 and 8. Some of the unevenness of the original data can be seen by comparing the measurements at the cusps of the cycloids with the reconstructed trajectory data. The state estimates are given in Table III. As expected, the dominant source of rotational motion is about the axis orthogonal to the image plane; however, the estimates of ω_x and ω_y are significant, amounting to about 45 and 16 degrees of motion, respectively, during the observation interval. This apparent motion is probably due in part to the distortion discussed above.

An error of 0.05 in normalized coordinates is scaled to a 5 inch error in the true object dimensions by $z_R = 94.4$ inches, which accounts for most of the errors in the interpoint distances shown in Table IV, since the estimate of z_R is about 94.4 inches. The measured value of z_R in the original scene is about 108 inches, so the estimate of this quantity is better in the Car than in the Bottle (estimate of 49.8 inches, measure-

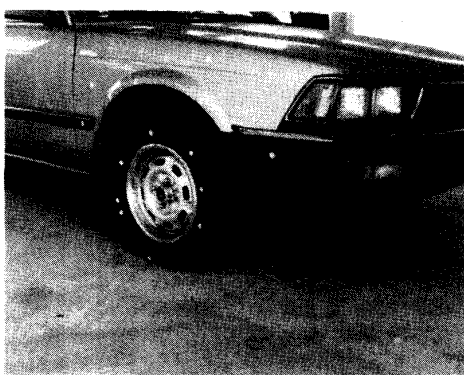


Fig. 7. Frames 1 and 4 from the Car sequence of 16 frames.

ment of 40 inches). The feature points lie in a single plane: it is believed that this is a factor in the ambiguity regarding the focal length.

The estimate of the translation rate corresponds to a translational distance estimate of 43.5 inches during the observation interval, as compared with the measured value of 45.0 inches. The rotational motion estimate is also fairly accurate, with a prediction of 3.96 radians total rotation compared with a measurement of about 3.85 radians yielding an error of about 6.3 degrees.

There are large errors in structure estimation. In this case, the errors appear to be due in part to distortion, which affects the interpoint distances directly. As seen in Table IV, the error of the interpoint distances tends to be the least for points adjacent to each other (e.g., points 1 and 2, points 1 and 8) and greatest for points nearly opposite each other (e.g., points 1 and 5, points 3 and 8). This consistent with a spatial distortion induced by an inaccurate focal length, since such a distortion would affect immediate neighbors the least.

The distortion becomes more evident when the angles between the various vectors are computed. Clearly, the axis of rotation is perpendicular to the axis of translation, since the rotation of the tire is directly related to the direction of travel. However, the estimated angle between these axes is about 150 degrees. When this issue is examined more closely, it is

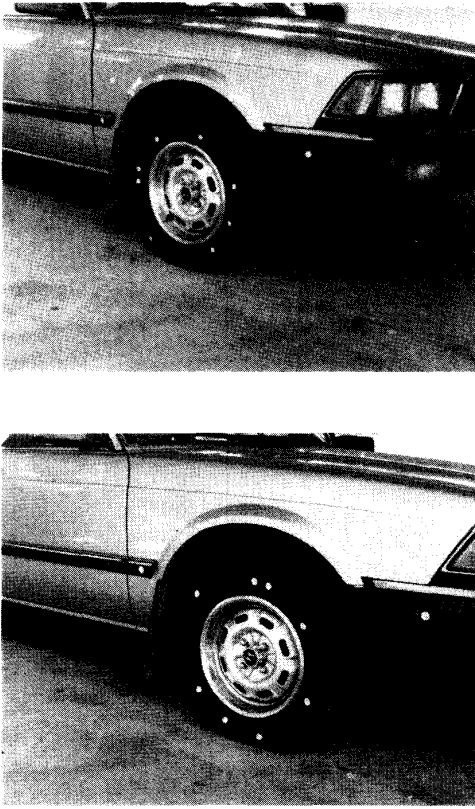


Fig. 8. Frames 7 and 13 from the Car sequence of 16 frames.

TABLE III
STATE ESTIMATES FOR CAR EXPERIMENT. SIXTEEN FRAMES
WERE USED WITH EIGHT FEATURE POINTS IN EACH FRAME.
ALL VALUES ARE GIVEN IN NORMALIZED UNITS, EXCEPT
THE ROTATION RATES, WHICH ARE IN RADIANS/SECOND.

State	Estimate	State	Estimate
x_R/z_R	0.2414	x_4/z_R	-0.0215
y_R/z_R	0.0743	y_4/z_R	-0.0576
x/z_R	0.2179	z_4/z_R	-0.1882
y/z_R	-0.0255	x_5/z_R	-0.0743
z/z_R	-0.5746	y_5/z_R	-0.0528
ω_x	0.7806	z_5/z_R	-0.2618
ω_y	0.3681	x_6/z_R	-0.0829
ω_z	5.2097	y_6/z_R	-0.0307
x_1/z_R	0.0485	z_6/z_R	-0.2517
y_1/z_R	0.0297	x_7/z_R	-0.0774
x_2/z_R	0.0546	y_7/z_R	-0.0027
y_2/z_R	0.0199	z_7/z_R	-0.2115
z_2/z_R	0.0005	x_8/z_R	0.0045
x_3/z_R	0.0400	y_8/z_R	0.0571
y_3/z_R	-0.0205	z_8/z_R	-0.0190
z_3/z_R	-0.0568		

found that the rotation axis is estimated to be about 10 degrees away from perpendicular to the image plane, while the axis of translation is estimated to be approximately 20 degrees away from perpendicular to the image plane, oriented in the opposite direction (translation vector is out of image, rotation vector is

TABLE IV
3-D DISTANCES BETWEEN OBJECT FEATURE POINTS, NORMALIZED ESTIMATES,
TRUE DISTANCES (INCHES), SCALED ESTIMATES (INCHES), AND SCALE FACTORS.
*INDICATES SMALLEST AND LARGEST ESTIMATES, TRIMMED FROM DATA.

i	j	Estimate	True Distance	Scaled Estimate	\hat{z}_R
1	2	0.0115	1.94	1.09	168.52
1	3	0.0763	7.63	7.20	100.00
1	4	0.2189	14.19	20.66	64.82
1	5	0.3007	18.13	28.38	60.28
1	6	0.2903	17.50	27.40	60.28
1	7	0.2482	16.13	23.42	64.97
1	8	0.0553	8.44	5.21	152.72
2	3	0.0716	5.94	6.76	82.96
2	4	0.2177	13.06	20.54	60.00
2	5	0.3012	17.81	28.42	59.14
2	6	0.2916	17.75	27.52	60.87
2	7	0.2507	16.94	23.66	67.56
2	8	0.0654	10.25	6.17	156.82
3	4	0.1497	8.13	14.13	54.28
3	5	0.2369	14.88	22.36	62.80
3	6	0.2306	16.44	21.76	71.28
3	7	0.1950	17.38	18.40	89.10
3	8	0.0933	14.69	8.81	157.41
4	5	0.0908	8.25	8.57	90.87
4	6	0.0924	11.75	16.44	71.28
4	7	0.0818	15.06	7.72	184.09
4	8	0.2060	18.25	19.44	88.59
5	6	0.0258	5.00	2.44	*193.72
5	7	0.0711	9.88	6.71	138.85
5	8	0.2779	18.38	26.23	66.12
6	7	0.0493	5.19	4.65	105.22
6	8	0.2636	15.50	24.88	58.80
7	8	0.2175	11.88	20.53	*54.60

into image). The resulting error of 60 degrees seems large; however, the Bottle experiment had an error in this same angle of 14 degrees, despite much higher accuracy overall. It is interesting to attempt to estimate the angles visually, by studying the images themselves; the guesses tend to be quite different than the algorithm's estimates, and the fact that the axis of rotation is perpendicular to the direction of translation is invoked almost immediately, information unavailable to the computer.

V. UNIQUENESS OF MODEL PARAMETERS

One of the important questions is to determine whether the motion as modeled in the present case, in a noise-free situation, using the central projection model, is uniquely determined from an image sequence. In the absence of *a priori* data about the object range, size or translational motion, there is an unknown global scale factor corresponding to the uncertainty in these attributes. The same data can be produced by a small object moving slowly at short range or a larger object, moving faster, at a longer range. As will be seen, this scale factor appears explicitly in the uniqueness proofs when the motion is pure translation. A second source of ambiguity can occur with respect to rotational motion. If the rotation of an object is observed at discrete intervals, rotation parameters that correspond to rates above the Nyquist sample rate are indistinguishable from the "correct" rotational parameters.

When motion is purely translational, the theorems presented here demonstrate the uniqueness of both the motion parameters

and the structural parameters, for both constant velocity and constant acceleration. The proofs of these theorems can be found in [5].

The proof of uniqueness of both motion and structure, for general motion and unknown structure, is more difficult. One approach is the use of the implicit function theorem (IFT), which can be used to show local uniqueness. That is, given a set of motion and structure parameters, the IFT can be used to show that a given parameter vector is uniquely determined by noise-free data, provided that a derivative matrix is of full rank. The results presented here indicate that in the cases examined, the use of three or more images in sequence is required for the parameters to be unique. A numerical procedure (singular value decomposition) is used to compute the rank of the matrix in question. When the matrix is not of full rank, the parameters are not uniquely determined by the data: this condition is observed when only two images are used to estimate the parameters. However, this does not address the global uniqueness problem, and as mentioned above, solutions are generally not globally unique. Further, since a numerical technique is used to determine the rank of the matrix, the arguments presented here do not constitute a formal proof. The use of (IFT) and the singular value decomposition to examine local uniqueness are discussed below.

A. Uniqueness for Pure Translation

Theorem 1: When an object is undergoing constant velocity translation, and no rotation, the noise-free central projections (images) of its feature points uniquely determine the corresponding motion and structure parameters, subject to the following conditions: 1) at least two feature points must not be parallel to the velocity vector, on a ray through the origin, 2) the object must be observed in at least three image frames, 3) the origin of the object coordinate system is not unique, but can be fixed by assigning an arbitrary feature point to be that origin, and 4) uniqueness is determined only up to a global scale factor $\beta > 0$, such that a parameter vector \underline{u} is indistinguishable from the parameter vector $\beta \underline{u}$.

The proof of this theorem is straightforward [5]. Thus, two points in three frames are required to uniquely determine motion and structure parameters, when the motion is constant velocity translation. The global scale factor is resolved if any information is known about the object dimensions, distance, or velocity. The structure parameters (coordinates of feature points in the object centered coordinate system) are unique in the sense that the vectors connecting the feature points are unique. Since the object coordinate system is not rotating, the choice of origin is arbitrary, and it suffices to use an arbitrary feature point as the origin. This feature point can be one of the two points required for uniqueness. The extension to constant acceleration is given in [5].

B. Uniqueness for General Motion

Even if one is willing to accept local uniqueness results, the use of the implicit function theorem requires finding the rank of a large matrix, of dimension equal to the number of unknown parameters. The validity of these results thus depends

on the assumption that the singular value decomposition (SVD) is a valid test for the rank of a matrix. As discussed in [18], the SVD can measure the “closeness” of a given matrix (represented to finite precision) to a singular matrix, in terms of the Frobenius norm, which for an $n \times m$ matrix A is given by

$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m \alpha_{ij}^2}. \quad (39)$$

This approximate approach is used to argue that in certain cases, because a matrix is very nearly singular, the parameters are not uniquely determined by the data, while in other cases a matrix is ill-conditioned but nonsingular, and the parameters are uniquely determined by the data.

The application of the implicit function theorem is that if f is a continuously differentiable mapping, $f(x, y) = 0$ can be solved uniquely for y in terms of x under certain conditions, as discussed by Rudin in [23]. Suppose x is a parameter vector, $h(\cdot)$ is the measurement function, and y is a vector of image coordinates. If $f(x, y) = h(x) - y = 0$ can be shown to uniquely determine x based on y , then the parameter vector x has been uniquely determined by the data y . The notation used here, as in Rudin, is that we write (\mathbf{x}, \mathbf{y}) for the point (or vector) $(x_1, \dots, x_n, y_1, \dots, y_m) \in R^{n+m}$. That is, if $\mathbf{a} = (a_1, \dots, a_n) \in R^n$ and $\mathbf{b} = (b_1, \dots, b_m) \in R^m$, we write the vector $(\mathbf{a}, \mathbf{b}) = (a_1, \dots, a_n, b_1, \dots, b_m) \in R^{n+m}$. Then, the relevant theorem is the following.

Implicit Function Theorem: Let \mathbf{f} be a C' -mapping of an open set $E \subset R^{n+m}$ into R^n , such that $\mathbf{f}(\mathbf{a}, \mathbf{b}) = 0$ for some point $(\mathbf{a}, \mathbf{b}) \in E$.

Put $A = \mathbf{f}'(\mathbf{a}, \mathbf{b})$ and assume that A_x (the derivative matrix of the vector-valued function \mathbf{f} with respect to its first argument $\mathbf{x} \in R^n$) is invertible.

Then there exist open sets $U \subset R^{n+m}$ and $W \subset R^m$, with $(\mathbf{a}, \mathbf{b}) \in U$ and $\mathbf{b} \in W$, having the following property:

To every $\mathbf{y} \in W$ there corresponds a unique \mathbf{x} such that

$$\mathbf{f}(\mathbf{g}(\mathbf{y}), \mathbf{y}) = \mathbf{0}, \quad (\mathbf{y} \in W), \quad (40)$$

$$\text{and } \mathbf{g}'(\mathbf{b}) = -(A_x)^{-1} A_y. \quad \square$$

As discussed previously, the mapping from parameter space to image coordinates in a noise-free case is given by

$$p_i(t_k) = \begin{pmatrix} h_X[s_i(\underline{u}, t_k)] \\ h_Y[s_i(\underline{u}, t_k)] \end{pmatrix}, \quad (41)$$

where

$$\begin{pmatrix} h_X[s_i(\underline{u}, t_k)] \\ h_Y[s_i(\underline{u}, t_k)] \end{pmatrix} = \begin{pmatrix} \frac{x_R/z_R + t_k \dot{x}/z_R + R_{11}x_{iO} + R_{12}y_{iO} + R_{13}z_{iO}}{1 + t_k \dot{z}/z_R + R_{31}x_{iO} + R_{32}y_{iO} + R_{33}z_{iO}} \\ \frac{y_R/z_R + t_k \dot{y}/z_R + R_{21}x_{iO} + R_{22}y_{iO} + R_{23}z_{iO}}{1 + t_k \dot{z}/z_R + R_{31}x_{iO} + R_{32}y_{iO} + R_{33}z_{iO}} \end{pmatrix}. \quad (42)$$

The substitutions $x_{iO} = x_i/z_R$, $y_{iO} = y_i/z_R$, and $z_{iO} = z_i/z_R$ have been made for brevity.

Denote the parameter space $\underline{u} \subset R^n$, (corresponding to the vector \mathbf{a}), where $n = 3M + 7$ in the case of constant translational and rotational rates with unknown structure, with

M feature points. The set W contains the data, which are the images of the M feature points, observed during N points in time. Thus, $\mathbf{h} = \{p_i(t_k), i = 1, \dots, M; k = 0, \dots, N-1\} = \{X_{ik}, Y_{ik}, i = 1, \dots, M; k = 0, \dots, N-1\} \in W$, has dimension $m = 2NM$. This corresponds to the vector \mathbf{b} in the statement of the theorem.

Consider the mapping from parameter space to data space at t_0 ,

$$\mathbf{h}(\underline{u}, t_0) = \begin{pmatrix} h_X(\underline{u}, t_0, i=1) \\ h_Y(\underline{u}, t_0, i=1) \\ h_X(\underline{u}, t_0, i=2) \\ h_Y(\underline{u}, t_0, i=2) \\ \vdots \\ h_X(\underline{u}, t_0, i=M) \\ h_Y(\underline{u}, t_0, i=M) \end{pmatrix}_{2M \times 1} \quad (43)$$

The mapping at times t_1, t_2 , etc., are analogous. We can concatenate these vectors for N image times as

$$\mathbf{h}(\underline{u}, t_0, \dots, t_{N-1}) = \begin{pmatrix} \mathbf{h}(\underline{u}, t_0) \\ \mathbf{h}(\underline{u}, t_1) \\ \vdots \\ \mathbf{h}(\underline{u}, t_{N-1}) \end{pmatrix}_{2NM \times 1} \quad (44)$$

This gives a mapping from parameter space to observation space. Then, $(\underline{u}, \mathbf{h}) \in U \subset R^{n+m}$ corresponds to (\mathbf{a}, \mathbf{b}) . The equation $\mathbf{f}(\underline{u}, \mathbf{h}) = \mathbf{0}$ has elements $h_X[s_i(\underline{u}, t_k)] - X_{ik}$ and $h_Y[s_i(\underline{u}, t_k)] - Y_{ik}$, giving

$$\mathbf{f}(\underline{u}, \mathbf{h}) = \begin{bmatrix} h_X[s_1(\underline{u}, t_0)] - X_{1,0} \\ h_Y[s_1(\underline{u}, t_0)] - Y_{1,0} \\ \vdots \\ h_X[s_M(\underline{u}, t_0)] - X_{M,0} \\ h_Y[s_M(\underline{u}, t_0)] - Y_{M,0} \\ \vdots \\ h_X[s_1(\underline{u}, t_{N-1})] - X_{1,N-1} \\ h_Y[s_1(\underline{u}, t_{N-1})] - Y_{1,N-1} \\ \vdots \\ h_X[s_M(\underline{u}, t_{N-1})] - X_{M,N-1} \\ h_Y[s_M(\underline{u}, t_{N-1})] - Y_{M,N-1} \end{bmatrix}_{2NM \times 1} = \mathbf{0}. \quad (45)$$

Next, consider the linear transformation $f'(\underline{u}, \mathbf{h}) = [A_{\underline{u}} I]$; this is simply the gradient of f with respect to the parameter vector \underline{u} , as the left submatrix, and the $m \times m$ identity matrix as the right submatrix. $A_{\underline{u}}$ is the Jacobian of the transformation from parameter space to image coordinates. It must be shown that $A_{\underline{u}}$ has rank of $n = 3M + 7$, which is required for h to be invertible. Here, $m = 2NM$. The IFT requires the number of rows of $A_{\underline{u}}$ to equal the $3M + 7$, so that the resulting linear transformation is consistent, and a solution exists. Since we assume that the data are generated by the model of the form presented here, and only the parameters are unknown, we are assured of a solution, and no existence proof is required for $m > 3M + 7$.

The difficulty with this approach to showing uniqueness is that geometric constraints on the object structure and/or

motion that insure full rank, hence uniqueness, are not easily obtained. Thus, we must rely on numerical methods to test for uniqueness, and must be satisfied at this point with evidence of uniqueness or nonuniqueness, as opposed to a proof.

In [8], it is shown that the Fisher Information matrix \mathbf{J} can be written as the sum of $2NM$ outer products of gradient vectors. This matrix can also be written as

$$\mathbf{J} = \frac{1}{\sigma_p^2} \mathbf{A}_{\underline{u}}^T \mathbf{A}_{\underline{u}}, \quad (46)$$

of dimension $n \times n$, where σ_p^2 is the image plane noise variance. If \mathbf{J} is of rank n , then $\mathbf{A}_{\underline{u}}$ is of rank n , and conversely, assuming that $m \geq n$, and the parameters are uniquely determined by the IFT. This is intuitively reasonable, since \mathbf{J}^{-1} produces lower bounds on estimation accuracy, so that if \mathbf{J} is singular, the inverse is undefined (one or more parameters have "infinite" variance).

Thus, the nonsingularity of \mathbf{J} indicates that $\mathbf{A}_{\underline{u}}$ is of full rank, which implies that the parameter vector is uniquely determined. This is the reason for the restriction of these results to local uniqueness. However, the numerical rank of \mathbf{J} is not equivalent to the true rank. In [18], SVD is suggested for testing the rank of a matrix, since it is easy to check how "close" a diagonal matrix described by finite word length is to a truly singular matrix. The SVD produces a diagonal matrix of singular values, in order of descending magnitude, resulting in the decomposition of a real matrix \mathbf{A} as

$$\mathbf{V}^T \mathbf{J} \mathbf{U} = \begin{pmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad (47)$$

where Σ is a diagonal $r \times r$ matrix of singular values σ_i , $i = 1, \dots, r$, such that $\sigma_i \geq \sigma_j$, $i < j \leq n$, and where \mathbf{V} and \mathbf{U} are unitary matrices. Because of finite computer word-length, the elements of matrix \mathbf{J} computed numerically will differ in general from those of the corresponding matrix \mathbf{J}' that would result if all computations were performed exactly (with infinite precision). This becomes particularly important in evaluating the rank of a matrix, since "random" roundoff errors perturb the computed singular values and blur the distinction between a truly singular matrix and a highly ill-conditioned (but nonsingular) matrix. The problem of estimating the rank of a matrix from computed singular values is addressed in [18], and the magnitudes of "spurious" computed singular values due to round-off errors are estimated.

For our purposes, the precise values of Σ and r , the rank of \mathbf{J} are not important beyond determining whether or not $r < n$ and the bounds derived in [18] strongly support the conclusion (but do not constitute a proof) that \mathbf{J} is truly singular for $N = 2$ and nonsingular for $N > 2$. That is, if the computed matrix \mathbf{J} is taken as a noisy approximation of the exact matrix \mathbf{J}' , approximate bounds are derived in [18] on the magnitude of the smallest "significant" singular value σ_r , consistent with the conclusion the \mathbf{J}' has rank $r < n$.

The singular values of \mathbf{J} have been computed for several cases, leading to three "types" of singular values. Four tables are presented, illustrating the various cases. First, Table V

TABLE V
SINGULAR VALUES OF $\mathbf{J} = \mathbf{A}_u^T \mathbf{A}_u / \sigma_p^2$; TWENTY
FRAMES ARE USED WITH FOUR FEATURE POINTS.

σ_1	0.3322D + 08	σ_2	0.2610D + 08	σ_3	0.9225D + 07
σ_4	0.5704D + 07	σ_5	0.4300D + 07	σ_6	0.3729D + 07
σ_7	0.3373D + 07	σ_8	0.3223D + 07	σ_9	0.2463D + 07
σ_{10}	0.2214D + 07	σ_{11}	0.1689D + 07	σ_{12}	0.1343D + 07
σ_{13}	0.7801D + 06	σ_{14}	0.3375D + 06	σ_{15}	0.3242D + 05
σ_{16}	0.2131D + 05	σ_{17}	0.8025D + 04	σ_{18}	0.5189D + 04
σ_{19}	0.1627D + 04				

TABLE VI
SINGULAR VALUES OF \mathbf{J} IN A RANK-DEFICIENT CASE.
TWO FRAMES ARE USED WITH SEVEN FEATURE POINTS.

σ_1	0.1556D + 07	σ_2	0.1467D + 07	σ_3	0.3831D + 06
σ_4	0.3031D + 06	σ_5	0.2488D + 06	σ_6	0.2336D + 06
σ_7	0.2195D + 06	σ_8	0.2147D + 06	σ_9	0.1900D + 06
σ_{10}	0.1818D + 06	σ_{11}	0.1715D + 06	σ_{12}	0.1544D + 06
σ_{13}	0.1490D + 06	σ_{14}	0.1427D + 06	σ_{15}	0.3956D + 04
σ_{16}	0.3206D + 04	σ_{17}	0.3016D + 04	σ_{18}	0.1764D + 04
σ_{19}	0.1477D + 04	σ_{20}	0.1391D + 04	σ_{21}	0.8593D + 03
σ_{22}	0.1076D + 03	σ_{23}	0.1162D + 02	σ_{24}	0.8975D + 01
σ_{25}	0.6384D + 01	σ_{26}	0.6446D - 03	σ_{27}	0.3789D - 11
σ_{28}	0.1466D - 11				

TABLE VII
SINGULAR VALUES OF \mathbf{J} FOR A VERY ILL-CONDITIONED
CASE. FOUR FRAMES ARE USED WITH TWO FEATURE POINTS.

σ_1	0.1350D + 07	σ_2	0.1220D + 07	σ_3	0.4578D + 06
σ_4	0.3746D + 06	σ_5	0.3420D + 05	σ_6	0.2966D + 04
σ_7	0.1506D + 04	σ_8	0.2804D + 03	σ_9	0.1587D + 02
σ_{10}	0.7387D + 00	σ_{11}	0.2322D + 00	σ_{12}	0.1549D - 01
σ_{13}	0.8885D - 03				

shows the singular values of $\mathbf{A}_u^T \mathbf{A}_u / \sigma_p^2 = \mathbf{J}$ when four feature points are observed in twenty frames. This is a very well-conditioned case, with the ratio of largest to smallest singular value being about 2×10^4 . This is compared to the singular values in Table VI, with the ratio being about 1×10^{18} . Ratios of this size are observed in all cases examined when only two images of data are used. The third case, tabulated in Table VII, has a ratio of 1×10^9 , while the final case, in Table VIII, has a ratio of 5×10^6 . The last two cases involve only two feature points, with 4 and 9 frames, respectively. Based on data of this type, for four cases, the conclusion is that r is chosen such that singular values on the order of 10^{-10} or less should be equated to zero. This suggests that all the cases examined that involve the use of only two image frames are underdetermined, and thus nonunique. This conclusion is in agreement with the IMSL numerical checking, from which the Fisher Information matrices involving two frames consistently elicited warning messages concerning algorithmic singularity. The remaining situations, consistent with the invertibility of \mathbf{J} , are of full rank, although sometimes ill-conditioned.

As an aside, if one simply counts measurements and unknowns, it is required that $2NM \geq 3M + 7$. This means that with 2 points, the minimum number of frames result in a feasible solution is 4. Thus, the third case (very ill-conditioned) has relatively little information in the data. The ratio of the largest singular value to the smallest increases

TABLE VIII
SINGULAR VALUES OF \mathbf{J} CORRESPONDING TO AN ILL-CONDITIONED
CASE. NINE FRAMES ARE USED WITH TWO FEATURE POINTS.

σ_1	0.3329D + 07	σ_2	0.3090D + 07	σ_3	0.1000D + 07
σ_4	0.8225D + 06	σ_5	0.4171D + 06	σ_6	0.5193D + 05
σ_7	0.2979D + 05	σ_8	0.7518D + 04	σ_9	0.5223D + 03
σ_{10}	0.1073D + 03	σ_{11}	0.3049D + 02	σ_{12}	0.7749D + 00
σ_{13}	0.6654D + 00				

monotonically as the number of frames is increased, with the fourth case (9 frames) picked as a typical example. If the counting argument is applied to the use of 2 image frames, the fewest number of points admitting a solution is 7. In the cases considered, a maximum number of points is 10. In all the cases evaluated that involve translation, rotation, and unknown structure, the singular values associated with the use of only 2 frames consistently produce ratios on the order of 10^{18} : this is taken as evidence that with the parameterization used in this research, 2 frames are not sufficient to estimate the model parameters. Since there are a number of methods that provide exact solutions based on 2 frames of noise-free data, it seems likely that the model parameterization used here differs in terms of the minimal amount of data needed.

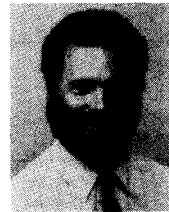
VI. CONCLUSIONS

The problem of estimating the structure and motion of a rigid body from a sequence of noisy monocular images is far from solved. The research reported here has addressed the issue of simultaneously estimating a large number of unknown parameters from a statistical, model based perspective, with no attempt made to form real-time estimates. The experimental and the uniqueness results presented here demonstrate the feasibility of this approach. The motivation for a nonreal-time batch solution is both as a theoretical tool and as a means of initializing the recursive algorithm [6]. This gives rise to a hybrid batch/recursive approach, [6] which differs from [9] in that our work addresses a small object (as opposed to a fixed scene) which leads to much noisier data in our work (by our measure of noise level), and simultaneous structure and motion estimation (as opposed to structure estimation first, followed by motion estimation).

REFERENCES

- [1] G. Adiv, "Determining three-dimensional motion and structure from optical flow generated by several moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, pp. 384-401, July 1985.
- [2] J. Aloimonos and C. M. Brown, "Perception of structure from motion: I: Optic flow vs. discrete displacements, II: Lower bound results," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 1986, pp. 510-517.
- [3] D. H. Ballard and O. A. Kimball, "Rigid body motion from depth and optical flow," *Comput. Vision, Graphics, Image Processing*, vol. 22, pp. 95-115, Apr. 1983.
- [4] S. S. Blackman, *Multiple-Target Tracking with Radar Applications*. Dedham, MA: Artech House, 1986.
- [5] T. J. Broida, "Estimating the kinematics and structure of a moving object from a sequence of images," Ph.D. dissertation, Univ. Southern California, 1987.
- [6] T. J. Broida, S. Chandrashekar, and R. Chellappa, "Recursive estimation of 3-D kinematics and structure from noisy monocular image sequences," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-26, pp. 639-656, Aug. 1990.

- [7] T. J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 90-99, Jan. 1986.
- [8] —, "Performance bounds for estimating three-dimensional motion parameters from a sequence of noisy images," *J. Opt. Soc. Amer. A*, vol. 6, pp. 879-889, June 1989.
- [9] E. D. Dickmanns, "An integrated approach to feature based dynamic vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Ann Arbor, MI, June 1988, pp. 820-825.
- [10] J.-Q. Fang and T. S. Huang, "Some experiments on estimating the 3-D motion parameters of a rigid body from two consecutive image frames," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 545-554, Sept. 1984.
- [11] O. D. Faugeras and M. Hebert, "A 3-D recognition and positioning algorithm using geometric matching between primitive surfaces," in *Proc. Int. Joint Conf. Artificial Intell.*, West Germany, Aug. 1983, pp. 996-1002.
- [12] B. Friedland, "Analysis of strapdown navigation using quaternions," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-14, pp. 764-768, Sept. 1978.
- [13] D. B. Gennery, "Tracking known 3-D objects," in *Proc. Nat. Conf. Artificial Intell.*, Aug. 1982, pp. 13-17.
- [14] E. C. Hildreth, "Computations underlying the measurement of visual motion," *Artificial Intell.*, vol. 23, pp. 309-354, Aug. 1984.
- [15] R. C. Jain, "Segmentation of frame sequences obtained by a moving observer," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 624-629, Sept. 1984.
- [16] J. K. Kearny, W. B. Thompson, and D. L. Boley, "Optical flow estimation: An error analysis of gradient-based methods with local optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 229-244, Mar. 1987.
- [17] Y. C. Kim and J. K. Aggarwal, "Determining object motion in a sequence of stereo images," *IEEE J. Robotics and Automat.*, vol. RA-3, pp. 599-614, Dec. 1987.
- [18] K. Konstantinides and K. Yao, "Statistical analysis of effective singular values in matrix rank determination," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 757-763, May 1988.
- [19] K. M. Mutch and W. B. Thompson, "Analysis of accretion and deletion at boundaries in dynamic scenes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, pp. 133-138, Mar. 1985.
- [20] G. D. Niva, "The use of quaternions with an all-attitude IMU," in *Proc. Annu. Rocky Mountain Guidance and Control Conf.*, Jan. 1982, pp. 269-283.
- [21] M. J. D. Powell, "Restart procedures for the conjugate gradient method," *Math. Program.*, vol. 12, pp. 241-254, Apr. 1977.
- [22] J. Roach and J. Aggarwal, "Determining the movement of objects from a sequence of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 554-562, Nov. 1980.
- [23] W. Rudin, *Principles of Mathematical Analysis*. New York: McGraw-Hill, 1976.
- [24] L. E. Scales, *Introduction to Nonlinear Optimization*. New York: Springer-Verlag, 1985.
- [25] I. K. Sethi and R. Jain, "Finding trajectories of feature points in a monocular image sequence," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 56-73, Jan. 1987.
- [26] H. Shariat and K. Price, "Motion estimation with more than two frames," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 417-434, May 1990.
- [27] B. Sridhar and A. V. Phatak, "Simulation and analysis of image-based navigation system of rotorcraft low altitude flight," in *Proc. Amer. Helicopter Soc. Meeting*, Atlanta, GA, Apr. 1988.
- [28] D. V. Stallard, "An angle-only tracking filter in modified spherical coordinates," in *Proc. AIAA Guidance, Navigation, and Control Conf.*, 1986.
- [29] N. Sugie and H. Inagaki, "Recovery of the 3-D structure of a moving object from orthographically projected optical flow," *Trans. Soc. Instrum. Contr. Eng.*, vol. 20, pp. 837-843, Sept. 1984.
- [30] W. B. Thompson, K. M. Mutch, and V. A. Bersins, "Dynamic occlusion analysis in optical flow fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, pp. 374-383, July 1985.
- [31] A. Tirumalai et al., "A non-linear optimization algorithm for the estimation of structure and motion parameters," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, San Diego, CA, June 1989, pp. 136-143.
- [32] R. Y. Tsai and T. S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 13-27, Jan. 1984.
- [33] J. A. Webb and J. K. Aggarwal, "Visually interpreting the motion of objects in space," *Computer*, vol. 14, pp. 40-46, Aug. 1981.
- [34] J. Weng, N. Ahuja, and T. S. Huang, "Closed form solution + maximum likelihood: A robust approach to motion and structure estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 1988.
- [35] J. Weng, T. S. Huang, and N. Ahuja, "3-D motion estimation, understanding, and prediction from noisy image sequence," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 370-389, May 1987.
- [36] J. R. Wertz, Ed., *Spacecraft Attitude Determination and Control*. Dordrecht, The Netherlands: D. Reidel, 1978.
- [37] Y. Yasumoto and G. Medioni, "Robust estimation of three-dimensional motion parameters from a sequence of image frames using regularization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 464-471, July 1986.
- [38] G. S. Young and R. Chellappa, "3-D motion estimation using a sequence of noisy stereo images: Models, estimation and uniqueness results," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 735-759, Aug. 1990.
- [39] X. Zhuang, T. S. Huang, and R. H. Haralick, "A simple procedure to solve motion and structure from three orthographic views," *IEEE J. Robotics Automat.*, vol. 4, pp. 236-239, Apr. 1988.



Ted J. Broida (S'82-M'86) received the B.A. degree in fine art from Antioch College, Yellow Springs, OH, in 1977, and the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles. His doctoral research was performed in the Signal and Image Processing Institute.

He has been employed by Hughes Aircraft Company since 1979, in the Advanced Programs Divisions of the Radar Systems Group, where he is a Senior Staff Engineer. His work includes design, analysis, and performance prediction of both single sensor and multiple sensor fusion systems for multiple target tracking, as well as real-time sensor and processor resource allocation problems. His research interests also include applications of estimation theory and nonlinear filtering theory. He is the author or coauthor of a number of papers in these areas, coauthored a chapter in *Multiple Target Tracking with Radar Applications* with S. S. Blackman, and recently presented at a UCLA short course on Tracking with Electro-Optical (Imaging) Sensors.

Dr. Broida was a recipient of the Hughes B.S. Scholarship, and M.S. and Ph.D. fellowships. In 1980 he was awarded the Certificate of Outstanding Academic Achievement at USC by the Metro LA section of the IEEE.



Rama Chellappa (S'79-M'87-SM'83) was born in Madras, India. He received the B.S. degree (honors) in electronics and communications engineering from the University of Madras in 1975, the M.S. degree (with distinction) in electrical communication engineering from the Indian Institute of Science in 1977, and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981, respectively.

During 1979-1981, he was a Faculty Research Assistant at the Computer Vision Laboratory, University of Maryland, College Park. Since 1986, he has been an Associate Professor in the Electrical Engineering—Systems, and as of September 1988, he became the Director of the Signal and Image Institute at the University of Southern California, Los Angeles. In August 1991, he will be joining the Department of Electrical Engineering at the University of Maryland, College Park, as a Professor. He will also be affiliated with the University of Maryland Institute for Advanced Computer Studies (UMIACS) and the Center for Automation Research. His current research interests are in signal and image processing, computer vision, and pattern recognition.

Dr. Chellappa is a member of Tau Beta Pi and Eta Kappa Nu. He coedited two volumes of selected papers on image analysis and processing, published in Fall 1985. He served as an Associate Editor for IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING and is currently a co-editor

of Computer Vision, Graphics and Image Processing: Graphic Models and Image Processing. He was a recipient of a National Scholarship from the Government of India during 1969–1975. He received the 1975 Jawaharlal Nehru Memorial Award from the Department of Education, Government of India, the 1985 Presidential Young Investigator Award, and the 1985 IBM Faculty Development Award. He also received the 1990 Excellence in

Teaching Award from the School of Engineering at the University of Southern California. He was the General Chairman of the 1989 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society Workshop on Artificial Intelligence for Computer Vision, and also Program Co-Chairman of the NSF-sponsored Workshop on Markov Random Fields for Image Processing, Analysis, and Computer Vision.