Reducing “Structure From Motion”: A General Framework for Dynamic Vision Part 1: Modeling

Stefano Soatto and Pietro Perona

Abstract—The literature on recursive estimation of structure and motion from monocular image sequences comprises a large number of apparently unrelated models and estimation techniques. We propose a framework that allows us to derive and compare all models by following the idea of dynamical system reduction. The “natural” dynamic model, derived from the rigidity constraint and the projection model, is first reduced by explicitly decoupling structure (depth) from motion. Then, implicit decoupling techniques are explored, which consist of imposing that some function of the unknown parameters is held constant. By appropriately choosing such a function, not only can we account for models seen so far in the literature, but we can also derive novel ones.

Index Terms—Visual motion estimation, epipolar geometry, motion decoupling, compensation, fixation, parallax, output stabilization, model reduction.

1 Introduction

Suppose that we are looking at a still scene through a moving camera. The problem of reconstructing both the motion of the camera and the structure of the scene, represented as the position of a number N of *point-features* in 3D space, is called “structure from motion” (SFM).

SFM has been the object of extensive study throughout the past two decades, and it is safe to say that the geometry of “N points in M views” is now fairly well understood. In other words, given the projection of N points onto M image-planes, and given knowledge of which point corresponds to which in different views, it is possible to reconstruct their position in space (possibly up to a change of projective or affine basis when the geometry of the imaging device is unknown). Luong and Faugeras provide a review of the state of the art in their forthcoming book [24].

Additionally, one may exploit the fact that images taken by a moving camera are adjacent in time, rather than being N snapshots taken from arbitrarily different vantage points. Such sequences of images can be analyzed either as a “batch” or in a recursive and causal fashion. In the latter case, the estimate at time t is obtained by processing the “past” measurements (up to time t), while in the former, the entire sequence is used. Recursive SFM can be traced back to the pioneering work of Dickmanns et al. [10] and Gennery [13]. The main motivations for such a choice lie in the fact that causal processing is necessary when the estimates are used to perform a closed-loop control action such as driving a car (we cannot use “future” images to turn the steering wheel “now”); moreover, recursive processing minimizes the storage and computational needs, a crucial issue in real-time operation.

In processing sequences of images there seems to be a fundamental tradeoff: If images are taken from very different vantage points (large baseline), the relative localization error is small, but features undergo large displacements on the image-plane, which makes the search for correspondences hard (global). If, on the other hand, images are sampled closely in time, so that the image-displacement is small, the correspondence problem is greatly simplified (local), but SFM becomes extremely sensitive to noise due to the small baseline [29]. The solution lies in *integrating information over time*, so that one can solve locally the correspondence problem for small displacements while effectively increasing the baseline and therefore making the estimates of SFM well behaved. The main obstacle in doing so is that individual feature-points may be visible only for very short time-spans or disappear due to occlusions and therefore we cannot just track them until the baseline is “long enough” and then apply any SFM algorithm. So far, only a few authors have addressed the issue of how to integrate structure and motion information over time even in the presence of a changeable set of feature-points (see for instance [19] for a batch approach and [27], [38] for recursive approaches). In the limit when only optical flow is available, so that the life-span of each feature point is two frames, the only algorithms that can integrate structure and motion information over time are [38], [37].

It is important to not confuse *time averaging* with *time integration*. We will illustrate this with an example: Suppose we take a sequence of noisy images closely sampled in time. Under these circumstances, due to the high noise level, an SFM algorithm applied to two subsequent views (time t and $t + 1$) is likely to return a flawed estimate. In

- *S. Soatto is with the Department of Electrical Engineering, Washington University, Campus Box 1127, One Brookings Dr., St. Louis, MO 63130. E-mail: soatto@ee.wustl.edu.*
- *P. Perona is with the Department of Electrical Engineering and Computation and Neural Systems, California Institute of Technology, Campus Box 136-93, Pasadena, CA 91125. E-mail: perona@vision.caltech.edu.*

Manuscript received 18 Dec. 1995; revised 2 July 1998. Recommended for acceptance by D.J. Kriegman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107115.

fact, the cost function associated to the problem of SFM has several local minima, as discussed in [44]. *Averaging* estimates corresponding to local minima will, in general, give a meaningless result. Instead, we would like to integrate information over time so as to effectively increase the baseline, even though individual features appear and disappear.

Despite the simplicity of the constraints that "define" the problem, the literature on recursive structure and motion estimation comprises a large number of quite diverse algorithms. We feel the need to understand the relationships between such algorithms and to assess the qualitative and quantitative properties of each one by comparing them on a common ground. In this paper, we will concentrate on recursive SFM algorithms and address the questions of whether there exist genuinely different approaches, whether there exists one that is "universally" better than the others or, if not, whether it is possible to choose the algorithm that is best suited to the application at hand.

1.1 Motion and Structure Estimation as an Optimization Problem

A variety of models have been proposed involving structure, motion, and images of feature points, for instance the coplanarity constraint [23], the subspace constraint [21], [15], [43], the so-called "plane plus parallax" constraint [4], [32], [34] and fixation constraints [12]. These have been exploited for estimating structure and / or motion from image sequences using a number of optimization schemes, either batch or recursive. Batch optimization techniques from two consecutive frames based upon the coplanarity constraint have been presented both in closed-form [23], [46], or iterative [17], [47]. The same holds for the subspace constraint [15]. Multiframe batch techniques have also been presented, both in closed-form under the orthographic and affine projection [31], [45], and iterative for the case of full perspective projection [1], [27], [30], [40], [41]. In this paper, we will be dealing with causal dynamical models for multiframe processing. In the companion paper [39], we will use such models for designing local observers, such as the Extended Kalman Filter (EKF) [20]. Schemes for recursive motion estimation also abound in the literature, see for instance [2], [6], [7], [9], [16], [27], [30], [37], [40]. Few of them, however, can account for a varying number of features, while only [37], [38] can *integrate* motion information in the limit where we measure optical flow (features survive for only two frames).

A simple counting of the dimensions will soon convince the reader that, regardless the estimation technique employed, the huge dimensionality of the problem and the highly nonlinear nature of the space of unknown parameters make the optimization so complicate that the issue of an appropriate *modeling* becomes crucial.

1.2 Decoupling as a Modeling Strategy

When facing a high-dimensional optimization problem, it is important to unravel the geometry of the space of unknown parameters, in order to see whether there are combinations of parameters that evolve independently in the cost objective. In that case one could decompose a high-dimensional optimization task into a number of smaller, simpler and better conditioned problems. In the case of structure and motion estimation, the work of Longuet-Higgins [23] (L-H) pioneered this approach, by decoupling structure from the motion parameters, which are encoded in a 3×3 matrix, called *Essential matrix*. Adiv [1] and Heeger and Jepson [15] (H-J) further decoupled the translational velocity from the rotational velocity.

We will rederive the constraints of H-J and L-H within a unified procedure. These lead, respectively, to the subspace constraint and the coplanarity constraint, interpreted as nonlinear implicit models with parameters on a manifold. Such a manifold is a five-dimensional space, called Essential manifold, in the discrete-time case of L-H and the two-dimensional sphere in the continuous-time case of H-J. This asymmetry between continuous and discrete time, which cannot be resolved in the context of the reduced-order observer, is what will motivate us towards alternative strategies for reducing the model.

1.3 "Explicit" Versus "Implicit" Decoupling

Although it is not always possible to decouple the unknown parameters in closed-form, it is possible to do so implicitly by imposing that some function of the parameters is held constant. We will see how this leads to a reduction of the model by constraining it onto a subspace of the parameter space. For instance, we may impose that the image of a point, a line, or a plane remains fixed. This procedure identifies slices of the parameter manifold where the model is constrained to evolve. For instance, these slices are four- and three-dimensional submanifolds of the Essential manifold, when a point or a line are fixated, and the two-dimensional sphere (also a submanifold of the Essential manifold), in the case in which a plane is fixated. Thus, we may interpret the compensation of the motion of a point, a line, or a plane, as a geometric stratification of the Essential manifold. By restricting the model to the appropriate slices, we derive four-, three-, and two-dimensional dynamic constraints, the latter being the discrete-time equivalent of the H-J constraint.

1.4 Relation to Previous Work

This paper starts with the standard rigid motion and perspective projection constraints, which are the essential ingredients of the problem and underlie all recursive schemes (for instance [2], [7], [16], [25], [27]), and derives the constraints of Longuet-Higgins [17], [23], [46], and Heeger and Jepson [15], in the context of the observer reduction.

An apparently unrelated line of work is motivated by the mechanics of the oculomotor system in primates. A number of studies have suggested that the task of estimating motion is made easier if some particular point on the scene is being fixated [12], [33], [42]. However, "made easier" cannot be directly quantified unless the different constraints are cast within the same framework and compared using the same optimization setup. We view such fixation constraints as instances of transformations of the input images that stabilize particular output functions such as the position of a point, a line or a plane in the image. This framework allows us to derive the point-fixation constraint [12], [33], [42], the plane-fixation underlying the so-called "plane-plus-parallax" representation [4], [32], [34], as well as intermediate constraints, for instance by fixating the motion of a point and a point on a line. All the constraints are imposed by considering slices of the parameter manifold, leaving the estimation technique untouched. This allows us to view all such models under the framework of epipolar geometry, and comparing them under equivalent conditions.

2 Recursive Estimation of Rigid Motion and Structure From Point-Features

2.1 The Basic Ingredients: Rigid Motion and Projection

We assume that the scene is described by a number N of *point-features* in 3D space, with coordinates $\mathbf{X}^i \in \mathbb{R}^3\ \forall i = \dots N$ relative to a reference frame centered in the optical center of the camera, which moves *rigidly* between successive time instants.

We call $\mathbf{X}^i = \left[X^i\ Y^i\ Z^i\right]^T \in \mathbb{R}^3$ the coordinates of a generic point $\mathbf{P}^i$ with respect to an orthonormal reference frame centered in the center of projection, with Z along the optical axis and X, Y parallel to the image plane and arranged as to form a right-handed frame. As the reference frame moves rigidly between time t and $t + 1$ (or equivalently, all points move rigidly relative to it), the coordinates of each point evolve according to

$$\mathbf{X}^i(t+1) = R(t)\mathbf{X}^i(t) + T(t) \qquad \forall i = 1 \dots N. \tag{1}$$

The matrix R belongs to the space of positive-determinant orthonormal 3×3 matrices, called $SO(3)$, and describes the change of orientation between the viewer’s reference at time t and that at time $t + 1$ relative to the object. $T \in \mathbb{R}^3$ describes the translation of the origin of the viewer’s reference frame. The instantaneous velocity of each featurepoint can be written as

$$\dot{\mathbf{X}}^i = \Omega \wedge \mathbf{X}^i + V \qquad \forall i = 1 \dots N \tag{2}$$

where, under the approximation that the velocity is constant between successive samples, the parameters (V, Ω) are related to (T, R) by the exponential map [28]. In particular, $R = e^{\Omega\wedge}$, where $\Omega\wedge$ belongs to the set of 3×3 skew-symmetric matrices, called $so(3)$, and describes the cross-product of Ω with a vector in $\mathbb{R}^3$. If we integrate (2) between time t_0 and the current time t, we end up with an equation of the form

$$\mathbf{X}^i(t) = {}^tR_{t_0}\mathbf{X}^i(t_0) + {}^tT_{t_0} \tag{3}$$

where ${}^tR_{t_0}$ and ${}^tT_{t_0}$ indicate the rotation and translation of the reference frame at time t relative to the one at the initial time.[1]

What we measure is the **perspective projection** π of the point features onto the image plane, which for simplicity we represent as the real projective plane $\mathbb{R}\mathrm{P}^2 \doteq \mathbb{R}^3 \backslash \mathbb{R}$. The projection map π associates to each $\mathbf{P}^i \neq 0$ its homogeneous coordinates:

$$\pi : \mathbb{R}^3 - \{0\} \to \mathbb{R}\mathrm{P}^2\ ;\ \mathbf{X} \mapsto \mathbf{x} \tag{4}$$

where $\mathbf{x} = \pi(\mathbf{X}) \doteq \left[\frac{X}{Z}\ \frac{Y}{Z}\ 1\right]^T$ with $Z \neq 0$. $\mathbf{x}$ is usually measured up to some error n, which we model as a white, zero-mean and normally distributed process with covariance R_n:

$$\mathbf{y}^i = \mathbf{x}^i + n^i \qquad n^i \in \mathcal{N}\left(0, R_n^i\right). \tag{5}$$

In practice, feature tracking and optical flow are subject to various sorts of errors:

1) pixel noise in the image,
2) erroneous correspondence, and
3) violations of the brightness constancy assumption [3].

Any algorithm for reconstructing 3D motion and / or structure in real-time must handle such errors in an automatic fashion. We will discuss a test for rejecting outliers in the companion paper [39].

2.2 Limitations of the Basic Model

The ensemble of equations (1)(5) or (2)(5) may be viewed as either a discrete-time or a continuous-time dynamical system that describes the evolution of point-features in space, depending upon a set of parameters that encode the rigid motion constraint. In the language of dynamical systems and control theory, (1) and (2) are called *state equations* (or model equations), and $\mathbf{X}^i$ are the *states*. Equation (5) is called *measurement equation*, or output equation. The motion parameters may be viewed either as the input to the model, or as unknown parameters in the model equation. Correspondingly, the task of estimating structure and motion may be seen as either a mixed state-estimation / model-inversion, or as a state-estimation / parameter-identification problem.

If the motion parameters (T, R) or (V, Ω) were known, then the position of the points in space could be recovered by estimating the state of the above linear dynamical systems (1)(5) or (2)(5) using an observer, for instance an EKF as in [25], [30]. Vice-versa, if the trajectory of the points in space was known, their motion parameters could be estimated by solving (2) as a linear system of algebraic equations. When neither the motion nor the structure of the scene are known, the problem becomes significantly more complicated, for we have to estimate both the state of the above models and identify their parameters.

Since we measure the output of such models over an interval of time, we may try to analyze the space[2] built of time-derivatives (or time-delays) of the output and see if it exhibits enough structure to allow reconstructing both the unknown states and the unknown parameters. The model that comes out of the basic constraints is “driftless,” in the sense that all of its dynamics depends upon the unknown input. This means that all constraints obtained from time-derivatives of the output couple the unknown states with the unknown input (parameters). Furthermore, it can be proven that only the first derivative produces independent constraints

[1] The parameters (T, R) that describe a rigid motion for a Lie group, called $SE(3)$ (Special Euclidean group acting on $\mathbb{R}^3$), and their instantaneous counterparts $(V, \Omega\wedge)$ are elements of the corresponding Lie algebra $so(3)$. For an introduction to the Lie groups $SO(3)$, $SE(3)$ and their corresponding Lie algebras $so(3)$, $se(3)$ see, for instance, [28].

[2] Such a space is called the “observability codistribution” [18], and is constructed by computing Lie derivatives of the output along the state vector field.

TABLE 1
GEOMETRIC STRATIFICATION OF THE PROBLEM OF ESTIMATING MOTION UNDER THE COMPENSATION OF THE IMAGE-MOTION OF A POINT, A POINT AND A LINE, AND A PLANE

Stabilized feature	Compensating 3D motion	Corresponding image deformation	Residual DOFs	State-space manifold
None	none	none	5	**E** Essential mfd
point	2D camera rotation	image center displacement	4	S^4 Sylvester mfd
point+line	rotation about optical center	image center shift+rotation	3	S^3 three-dimensional Sylvester mfd
plane	no feasible 3D rigid motion	quadratic warping	2	*so*(3) skew-symmetric unit-norm three-matrices

on the unknowns, and therefore it is not possible to unravel both the state of the model and its parameters [36].

At this point, we face a choice of two opposite strategies. We may "dynamically extend" the model, which means that we take the derivatives of the input to be the unknown parameter, rather than input itself. Then it is possible to extend the model by inserting the input into the state. Alternatively, we may try to "reduce" the original model by decoupling the states from the parameters. These alternative strategies are discussed in the coming Sections 2.3 and 3.

2.3 "Think Big": Dynamic Extension and Observers

In order to extend the state of the model described by (1)(5) or (2)(5) we have to assume some dynamics for the motion parameters:

$$\begin{cases} T(t+1) = f_T\big(T(t), n_T(t)\big) \\ R(t+1) = f_R\big(R(t), n_R(t)\big) \end{cases} \text{or} \begin{cases} \dot{V} = f_V\big(V, n_V\big) \\ \dot{\Omega} = f_\Omega\big(\Omega, n_\Omega\big) \end{cases} \tag{6}$$

since we do not know n_T, n_R, n_V, $n\Omega$, this is a purely formal step. If some a-priori information is available on how the motion parameters evolve, for instance the dynamics of the vehicle on which the camera is mounted, or a bound on acceleration, then it may be written in the form of f_V, $f\Omega$. The simplest possible model is constant velocity

$$\begin{cases} T(t+1) = T(t) \\ R(t+1) = R(t) \end{cases} \text{or} \begin{cases} \dot{V} = 0 \\ \dot{\Omega} = 0 \end{cases} \tag{7}$$

Alternatively, one may use a stochastic model, for instance a Brownian motion, where n_* are appropriately defined white, zero-mean and Gaussian noises. Although we are going to use a Brownian motion for the purpose of analysis, we stress that any other dynamical or statistical model may be inserted in place of f_*, as long as it preserves the observability properties of the original system. There is no "right" model for the motion parameters, and eventually validation must come from experiments; in the companion paper [39] we argue that a first-order random walk can give satisfactory results despite its simplicity.

In order to recover both structure and motion from the augmented model we need an observer[3] whose state-space is now quite complicated, for the motion parameters belong either to the Lie-group of Euclidean motions, $(T, R) \in SE(3)$, or to the corresponding Lie-algebra, $(V, \Omega\wedge) \in se(3)$:

$$\begin{cases} \mathbf{X}^i(t+1) = R(t)\mathbf{X}^i(t) + T(t) \\ T(t+1) = T(t) + n_T(t) \\ R(t+1) = R(t)e^{n_R\wedge(t)} \\ \mathbf{y}^i(t) = \pi\big(\mathbf{X}^i(t)\big) + n^i(t) \end{cases} \quad \forall i = 1 \dots N(t) \tag{8}$$

where n_T, n_R, and n^i are white, zero-mean Gaussian noises and $R(t) \in SO(3)$ and $T(t) \in \mathbb{R}^3$. This model underlies all recursive motion estimation methods seen in the literature. Nonstructural variations of this model include change of state coordinates (for instance, object-centered or world-centered reference coordinates), and a change of the parameter dynamics, for instance higher-order random walks. A change of the projection model (for instance, weak perspective or orthography) is significant from the modeling point of view; however, all the essential features of the problem are captured by the perspective projection, and all the concepts that we will treat in this paper can be extended to other projection models quite easily.

There are two problems with such an approach: the high-dimensionality of the models, and the lack of local observability. Suppose we are looking at number $N = 100$ of points, which is a conservative number of feature points in realistic image sequences. Then, the state of the filter just described has dimension 305, since there are 300 coordinates of the points, six motion parameters, and one unknown scaling factor that affects the depth of the scene and the norm of the translational velocity. Moreover, due to occlusions and appearance of new features, the number of visible features $N(t)$ changes in time, which causes the filter to have a variable dimension: when a new feature enters the state, it needs to be initialized and the estimation error for the position of that feature will have a discontinuity, which propagates onto the estimates of the motion parameters. Therefore, even when the motion is smooth but the set of feature points changes in time, the estimates of motion are subject to discontinuities. In [27], a method is proposed for dealing with such a situation, which uses a "variable state-dimension filter."

Local observers, such as EKF, update the estimates with the residual of the prediction multiplied by a gain computed from the linearization of the model. The model just described, however, is not locally observable [36]. As an intuitive argument, first observe that (8) is "block triangular," in the sense that the dynamics of each feature point $\mathbf{X}^i$

[3] We recall that an observer for a dynamical model is itself a dynamical system that takes as inputs the input/output pairs of the original model, and returns as output an estimate of its state. For in introduction to the basic concepts of linear observers, see, for instance [22]. The Kalman filter represents an instance of an observer for a special class of linear systems driven by white, zero-mean, and Gaussian noise. For an introduction to Kalman filtering, see, for instance, [20].

depends only on itself and on the motion parameters, but not on other points $\mathbf{X}^j \,|\, i \neq j$. This means that, as far as the observability is concerned, it does not matter how many points are visible. In particular, the observability of *motion parameters* does not depend upon the number of visible points, while it is intuitive that the more points are visible the better the perception of motion ought to be. For instance, consider a camera moving with constant velocity on a short interval of time while viewing a single point. If the image of the point moves along the horizontal axis x of the image plane in the positive direction, this could correspond, for instance, to the viewer translating along the opposite direction $-X$, or rotating about the axis Y. In few words, these two motions are *locally indistinguishable*. However, under the assumption of constant velocity, when we let the point move for a longer period of time we can *distinguish* these different motions, for translational motion along $-X$ produces a constant velocity motion on the image plane, while a rotational velocity along Y causes the projection to escape in finite time.

3 "Think Small": Reducing the Order of the Model

3.1 Explicit Reduction

The reduced-order observer [22] is a long-established technique for reducing the dimension of an observer for a dynamical system. The basic idea is to solve the measurement equation for some of the states, and then substitute into the model equation. The states that have been eliminated are no longer part of the state-space, and their state equation becomes a new measurement equation, which involves derivatives of the measurements. The original measurement equation becomes now trivial, for it has been used to define the states to be eliminated.

For instance, consider the simple linear model

$$\begin{cases} \dot{x}_1 = a_{11}x_1 + a_{12}x_2 \\ \dot{x}_2 = a_{21}x_1 + a_{22}x_2 \\ y = c_1x_1 + c_2x_2 \end{cases} \tag{9}$$

and "solve" the measurement equation for x_2, so that $x_2 \doteq \frac{y - c_1x_1}{c_2}$. If we now substitute x_2 into the dynamic equations, we get a new state model for x_1 which does not involve x_2 but has an "output injection" term, and a constraint involving the measurements y and $\dot{y}$ and the unknown state x_1:

$$\begin{cases} \dot{x}_1 = \left(a_{11} - a_{12}\frac{c_1}{c_2}\right)x_1 + \frac{a_{12}}{c_2}y \\ \frac{1}{c_2}\dot{y} - \left(a_{22}\frac{1}{c_2} + a_{12}\frac{c_1}{c_2^2}\right)y = \left(a_{11}\frac{c_1}{c_2} - a_{22}\frac{c_1}{c_2} - a_{12}\frac{c_1^2}{c_2^2} + a_{21}\right)x_1 \end{cases} \tag{10}$$

The previous measurement equation is now the identity $y = y$. We may rewrite the above model as

$$\begin{cases} \dot{x}_1 = \tilde{a}x_1 + ky \\ \tilde{y} = \tilde{c}x_1 \end{cases} \tag{11}$$

where $\tilde{y}$ hides a time-derivative of the measured output y. It is possible to get rid of this undesirable effect by either an output-dependent change of coordinates, as done in the original reduced-order observer [22], or by integrating the measurement equation over a sample time interval.

Let us apply this simple idea to the extended model (8) derived from (2)(5), after integrating it from the initial time t_0 to the current time t. In the simplest case of constant velocity, we have

$$\begin{cases} \dot{\mathbf{X}}^i(t_0) = 0 \\ \dot{\Omega} = 0 \\ \dot{V} = 0 \\ \mathbf{y}^i(t) = \pi\left({}^tR_{t_0}(\Omega)\mathbf{X}^i(t_0) + {}^tT_{t_0}(V,\Omega)\right) + n^i(t) \end{cases} \tag{12}$$

where $\left({}^tT_{t_0}, {}^tR_{t_0}\right)$ describes the change of coordinates between the initial (at t_0) and the current (at t) viewer's reference frame. After a change of coordinates $\mathbf{X}^i \mapsto \mathbf{x}^iZ^i$, we can solve the measurement constraint for $\mathbf{x}^i$, substitute into the state equation, and integrate the measurement equation starting from the initial time-instant. By doing so, we can eliminate $2N$ states, and be left with a model having $N + 6$ states, the depth of each point and the motion parameters:

$$\begin{cases} \dot{Z}^i(t_0) = 0 \\ \dot{\Omega} = 0 \\ \dot{V} = 0 \\ \mathbf{y}^i(t) = \pi\left({}^tR_{t_0}(\Omega)\mathbf{y}^i(t_0)Z^i(t_0) + {}^tT_{t_0}(V,\Omega)\right) + \\ \qquad\qquad n^i(t) + n_y^i \;\; ; \; t > t_0 \end{cases} \tag{13}$$

Since we cannot measure $\mathbf{x}^i(t_0)$, but only its noisy version $\mathbf{y}^i(t_0)$, we have to add a bias term n_y^i to the measurement equation.[4]

One may now write an EKF for such a model, where the constant states are modeled as first-order random walks, in order to estimate simultaneously depth and motion of the points. This approach has been pursued by Azarbayejani and Pentland. In [2], they consider an extended model that has a second-order random walk for the motion parameters, and an alternative projection model that allows orthography as a subcase (see the companion paper [39] for more details). Note that, since there is a scale factor ambiguity, the filter will estimate the depth of each point and the translational velocity modulo a one-dimensional subspace.

Model (13) is structurally similar to (8), and still suffers the shortcomings outlined in Section 2.3. These motivate us towards pushing the idea of the reduced-order observer one step further, in order to eliminate the structure parameters from the state, and be left with models that only involve motion and measured projections.

3.1.1 *Pushing the Model Reduction: Structure-Independent Motion Estimation*

Let us apply the idea of the reduced-order observer twice to the model of (2)(5). As we have seen in Section 3.1, in the first run we can eliminate $2N$ states, corresponding to the measured projections of each feature point, and be left with

[4] Such a bias, equal to the measurement error at time 0, is the price one pays for using the measurements to reduce the order of the model. In order to eliminate the bias, one can insert it in the state, thus obtaining model (12).

$N + 5$ states describing the depth of each point Z^i and the motion parameters. Now we can "solve" the new measurement equation, which in fact corresponds to the image motion field (and is approximated by the optical flow), for the depth parameters Z^i.

Since the expression of the image motion field $\dot{\mathbf{x}}$ is linear both in the inverse depth and the rotational velocity, one may eliminate both depth and rotation, as done in Adiv [1]. Heeger and Jepson [15] proposed to use orthogonal projections to perform such an elimination: consider the time-derivative of the projection of each feature point, which can be written in the form

$$\dot{\mathbf{x}}^i(t) = C^i\left(\mathbf{x}^i, V\right)\begin{bmatrix}\frac{1}{Z^i(t)}\\ \Omega(t)\end{bmatrix} \tag{14}$$

where $C^i(\mathbf{x}^i, V) = [\mathcal{A}^i V \mid \mathcal{B}^i]$, and

$$\mathcal{A}^i \doteq \begin{bmatrix}1 & 0 & -x^i\\ 0 & 1 & -y^i\end{bmatrix} \quad \mathcal{B}^i \doteq \begin{bmatrix}-x^i y^i & 1 + x^{i^2} & -y^i\\ -1 - y^{i^2} & x^i y^i & x^i\end{bmatrix}. \tag{15}$$

The derivative of the third (homogeneous) coordinate of $\mathbf{x}^i = [x^i \;\; y^i \;\; 1]^T$ is identically zero, and has therefore been neglected. Given a sufficient number of point-features, the equation

$$\dot{\mathbf{x}} = C(\mathbf{x}, V)\left[\frac{1}{Z^1}, \dots, \frac{1}{Z^N}, \Omega\right]^T, \tag{16}$$

where

$$C(\mathbf{x}, V) \doteq \begin{bmatrix}\mathcal{A}^1 V & & \mathcal{B}^1\\ & \ddots & \vdots\\ & \mathcal{A}^N V & \mathcal{B}^N\end{bmatrix}, \tag{17}$$

may be solved for the inverse depth parameters and the rotational velocity, provided that $N > 3$, and then substituted into the same equation, which becomes

$$\dot{\mathbf{x}} = CC^\dagger \dot{\mathbf{x}} \tag{18}$$

where $C^\dagger \doteq \left(C^T C\right)^{-1} C^T$ denotes the pseudo-inverse. This leaves us with a constraint involving only translation V and measured image-coordinates/flow:

$$\left[I - CC^\dagger\right]\dot{\mathbf{x}} \doteq C^\perp(\mathbf{x}, V)\dot{\mathbf{x}} = 0. \tag{19}$$

It can be shown that this operation does not alter the structure of the innovation. Since there is an overall scaling factor ambiguity, only the direction of translation $\frac{V}{\|V\|}$ can be recovered, which we represent by imposing $\|V\| = 1$. The above constraint describes a particular class of nonlinear dynamical system, called *Exterior Differential System* [8] (EDS), with the parameters V constrained on the unit-sphere $\mathbf{S}^2$. We may therefore write our dynamical model as

$$\begin{cases} C^\perp(\mathbf{x}, V)\dot{\mathbf{x}} = 0 & V \in \mathbf{S}^2\\ \mathbf{y}^i \doteq \mathbf{x}^i + n^i & \forall i = 1 \dots N\end{cases} \tag{20}$$

Now, estimating motion is equivalent to identifying the above EDS, with parameters V on a sphere. Once such parameters have been identified, the remaining ones can be recovered a posteriori through the "pseudo-measurement"

$$\left[\frac{1}{\hat{Z}^1} \;\; \cdots \;\; \frac{1}{\hat{Z}^N} \;\; \hat{\Omega}\right]^T = C^\dagger \dot{\mathbf{x}}. \tag{21}$$

We show in the companion paper [39] how to practically perform the identification of models of the form (20).

Discrete-Time: The Essential Model

The idea of the reduced-order observer may be applied also to the discrete-time system (1)(5). The tool used to "eliminate" the depth parameters is now the so-called "Epipolar geometry" (see [11] for a review).

When a rigid object is moving between two time instants t and $t + 1$, the coordinates $\mathbf{X}^i(t)$ of a point at time t, their correspondent $\mathbf{X}^i(t + 1)$ at time $t + 1$, and the translation vector T are coplanar. Their triple product is therefore zero. This is true of course also for $\mathbf{x}^i(t)$, $\mathbf{x}^i(t + 1)$ and T, since $\mathbf{x}^i$ is the projective coordinate of $\mathbf{X}^i$ and therefore the two represent the same direction in $\mathbb{R}^3$, interpreted as the "ray-space" model of $\mathbb{RP}^2$ [35]. When expressed with respect to a common reference frame, for example that at time t, we may write the triple product as

$$\mathbf{x}^i(t + 1)^T \left(T \wedge \left(R\mathbf{x}^i(t)\right)\right) = 0 \quad \forall\, i = 1 : N. \tag{22}$$

Let us define $\boldsymbol{Q} \doteq (T\wedge)R$, so that the above coplanarity constraint, which is also known as the "Essential constraint" or the "epipolar constraint," becomes

$$\mathbf{x}^i(t + 1)^T\, \mathbf{Q}\mathbf{x}^i(t) = 0 \quad \forall\, i = 1 \dots N. \tag{23}$$

The above constraint may be interpreted as a discrete-time implicit dynamical model, with unknown parameters constrained to be of the form $T \wedge R$. Estimating motion therefore corresponds to identifying the model

$$\begin{cases}\left(\mathbf{Q}\mathbf{x}^i(t)\right)^T \mathbf{x}^i(t + 1) = 0 & \mathbf{Q} \in E\\ \mathbf{y}^i = \mathbf{x}^i + n^i & \forall i = 1 \dots N,\; n^i \in \mathcal{N}\left(0, R_{n^i}\right)\end{cases} \tag{24}$$

where now the parameters $\mathbf{Q}$ are constrained to belong to the so-called *Essential manifold*

$$E \doteq \{SR \mid R \in SO(3),\; S = (T\wedge) \in so(3)\} \subset \mathbb{R}^{3\times 3} \tag{25}$$

normalized in order to take into account the scale factor $\|T\| = 1$. The Essential manifold is a differentiable manifold of dimension six (or five after normalization), which is isomorphic to the tangent bundle of the rotation group $TSO(3)$, and therefore to the Euclidean group of rigid motions $SE(3)$. For a discussion of the topological and differential properties of the Essential manifold, see [37], and for a thorough description of its algebraic structure, see for instance [11], [26].

Asymmetry Between Continuous and Discrete-Time

The application of the simple idea of the reduced-order observer led us to formulating two implicit dynamical models involving only motion parameters and image coordinates. In the continuous-time case we could push the idea of the reduced-order observer up to the point in which we had a model with only two parameters. This was reasonably simple, for the parameters of rotation appeared linearly in the reduced measurement equation [15]. This did not work in the discrete-time case. In fact, although the elements of the rotation matrix R appear linearly, the rotation parameters Ω appear through the exponential map $R = e^{\Omega\wedge}$, which we cannot invert in closed-form in order to substitute it into the model equation and apply the trick of the reduced-order observer.

Therefore, there is an asymmetry between the instantaneous case and the discrete-time case. This will motivate us to explore alternative methods for reducing the state of the observer, which we do in the next section.

3.2 Implicit Reduction: Motion From Fixation

3.2.1 Output Stabilization and Geometric Stratification

Suppose that we are told that some of the states of a dynamical model are fixed. Then we may constrain the observer to the remaining states, and eliminate the constant ones from the dynamical model. The same applies if a *function* of the states is held constant. In fact, consider a point in the state-space manifold $\mathbf{P} \in M$. If $f : M \rightarrow \mathbb{R}$ is smooth, and $0 = f(\mathbf{P})$ is a regular value, then the preimage $f^{-1}(0) \subset M$ is a submanifold of M [14], and the point $\mathbf{P}$ is constrained onto such a submanifold. In this case it is possible to find a set of coordinates where some of the parameters are constant, and we can therefore concentrate our attention on the remaining ones.

Therefore, if we view some function of the state as an *output* (measurement equation) of the dynamic system, and this output is held constant, or *stabilized*, we may identify a "slice" of the state-manifold, and constrain the model on such a slice.

Although the choice of which function to stabilize is arbitrary, we will consider three simple instances; the image-motion of a point, a point and a line through it, and a plane. By stabilizing such outputs, we identify slices of the Essential manifold which build a geometric stratification of the problem of estimating motion under fixation constraints.

3.2.2 Choosing a Control Action

In order to stabilize a particular function of the image, we could either actuate the camera, and move it in space ("mechanical control"), or preprocess the image by considering changes of coordinates that depend upon the outputs, without physically acting on the camera ("software control"). For instance, keeping a single feature point fixed on the image plane can be accomplished both by rotating the camera about the center of projection (or about another point in space), or by shifting the origin of the image-coordinates. As far as the effects on motion estimation are concerned, the two methods are equivalent. It is simple to design gaze-control techniques which guarantee exponential convergence, while image-shift registration techniques that achieve fixation in a single step are described, for instance, in [42].

Fixating a point and a line through it on the image plane may be easily achieved by fixating a point and then rotating the image until another point comes to the desired line. This may be accomplished both by rotating the camera about the fixation axis, or by rotating the image about the optical center with a purely software operation.

Fixating a plane in the image, however, can be only accomplished by manipulating, or preprocessing, the image as described in Section 3.2.5.

3.2.3 Stabilization of a Point (Fixation)

Let us assume that we have applied some fixation technique that provides us with a sequence of images where the projection of a given point remains fixed on the image-plane. Since the projection of the fixation point is stationary, the object (scene) is free only to rotate about this point, and to translate along the fixation line. Therefore there are overall 4 degrees of freedom left. These four degrees of freedom are encoded into the rotation matrix $R = e^{\Omega\wedge}$, and in the relative translation along the fixation axis $v \in \mathbb{R}$. The epipolar representation presented in the previous section applies immediately once we represent the translation T as

$$T(R, v) \doteq \left[{}_{-R_{13}\ \ -R_{23}\ \ -R_{33}+v} \right]^T, \tag{26}$$

and $v \doteq \frac{d(t+1)}{d(t)} \neq 0$ is the ratio between the distance of the fixation point at time $t + 1$ and the same distance at time t.

The coplanarity constraint (23) also holds in the case of fixation, once we have substituted the appropriate expression for T. Since there are four degrees of freedom, the parameters Ω and v will now lie on a four-dimensional subspace of the Essential manifold. Indeed, it can be shown that the Essential matrices under the fixation constraint are all and only the 3 × 3 Essential matrices that satisfy the following Sylvester's equation

$$\mathbf{Q}(R, v) = RS^T + vSR \tag{27}$$

where $S \doteq [0\ \ 0\ \ \alpha]^T\wedge$ and α is the arbitrary scaling factor due to the homogeneous nature of the coplanarity constraint. We will call $\mathcal{S}^4$ the four-dimensional submanifold of the Essential manifold which is defined by the above equation after normalization. The $\mathcal{S}^4$ manifold is locally diffeomorphic to $\mathbb{R} \times SO(3)$ and hence to $\mathbb{R}^4$.

Therefore, in order to estimate motion under the fixation constraint, it is sufficient to consider the epipolar constraint where now the parameters are constrained not on the Essential manifold, but on the $\mathcal{S}^4$-manifold. We have therefore to deal with a model of the form

$$\begin{cases} \left(\mathbf{Q}\mathbf{x}^i(t)\right)^T \mathbf{x}^i(t+1) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} \in \mathcal{S}^4 \tag{28}$$

where

$$\mathcal{S}^4 = \{\mathbf{Q} \in \mathbf{E} \mid \mathbf{Q} = RS^T + vSR,\ R \in SO(3),$$
$$v \in \mathbb{R},\ S = [0\ \ 0\ \ 1]^T \wedge\}. \tag{29}$$

Estimating motion reduces to identifying the above dynamical system with parameters on $\mathcal{S}^4$. At this stage we do not consider the unavoidable fixation error, which makes the analysis significantly more complicated. In the companion paper we evaluate experimentally the impact of the fixation error on the quality of the estimates.

3.2.4 Stabilization of a Point and a Line Through It

Suppose now that, in addition to fixating a point, we can maintain a line passing through it fixed in the image plane.

We are essentially in the same situation described in the previous section, once we have "frozen" the degree of freedom corresponding to cyclorotation (rotation about the optical axis). Therefore, there are overall 3 degrees of freedom. The Essential matrices corresponding to motions that obey the "point plus line" fixation constraint must lie on a three-dimensional submanifold of the submanifold S^4 of the Essential manifold **E**, since the point-fixation constraint described in the previous section is satisfied. The only modification that occurs is that now there is no cyclorotation. Therefore the parameter space becomes

$$S^3 = S^4 \cap \left\{ R = e^{[\omega_1 \;\; \omega_2 \;\; 0]^T \wedge} \right\}. \tag{30}$$

Hence, under the "point plus line" fixation assumption, we end up with a model of the form

$$\begin{cases} \left(\mathbf{Q}\mathbf{x}^i(t)\right)^T \mathbf{x}^i(t+1) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} \in S^3 \tag{31}$$

which needs to be identified in order to estimate the motion parameters.

3.2.5 Stabilization of a Plane

We now proceed in our stratification by assuming that we are able to "compensate" the image sequence in such a way that the points that lie on some plane (not necessarily a physical plane in the scene) remains fixed in the image plane. In this case there is no physical motion of the camera that achieves this compensation (besides locking the camera to the plane). Therefore we need to "warp" the images of the sequence in order to account for the motion of the plane.

Compensation of Plane-Motion: Warping

Let us assume, for the moment, that all points in the scene lie on a plane – not passing through the origin – described by $\Pi = \{\mathbf{X}_\pi \in \mathbb{R}^3 \,|\, \mathbf{a}^T\mathbf{X}_\pi = 1\}$. We indicate with $\mathbf{x}_\pi \in \mathbb{R}\mathrm{P}^2$ the projective coordinates of the generic point of the plane Π. We will now see that, as the plane Π moves rigidly in space, its image deforms according to a projective transformation, i.e., a linear transformation of the projective coordinates. In fact, we may write the evolution of the 3D points of the plane as

$$\mathbf{X}_\pi^i(t+1) = R(t)\mathbf{X}_\pi^i(t) + T(t)\mathbf{a}^T\mathbf{X}_\pi^i(t) \doteq A(t)\mathbf{X}_\pi^i(t) \tag{32}$$

where $A(t) = R(t) + T(t)\mathbf{a}^T$ is a 3×3 invertible matrix. The projective coordinates of the points on the plane obey a similar relation

$$\mathbf{x}_\pi^i(t+1) \sim A(t)\mathbf{x}_\pi^i(t) \tag{33}$$

where the symbol $\sim$ indicates equality up to a scaling factor (projective equivalence). Given four or more point-correspondences on the image-plane, we may solve the above equation for the eight parameters of A that are free after normalization.

Once the matrix A has been estimated, up to a scaling factor, we may *undo* the transformation by multiplying the transformed points by A^{-1}:

$$\mathbf{x}_\pi^i(t+1)^w \doteq A^{-1}\mathbf{x}_\pi^i(t+1) = \mathbf{x}_\pi^i(t). \tag{34}$$

Therefore, such a *warping* leaves the points of the plane fixed in the image [4], [32], [34].

Plane-Plus-Parallax Representation

Now, let us assume that we have compensated for some plane, for instance the average plane, and see what happens to the points $\mathbf{X}^i$ that do not lie on such a plane, after the warping with A^{-1}. In general, $\mathbf{x}^i(t + 1)^w \neq \mathbf{x}^i(t)$. More specifically, we have

$$\begin{aligned} \mathbf{x}^i(t+1)^{\mathrm{w}} &\sim A^{-1}\mathbf{x}^i(t+1) = (R + T\mathbf{a}^T)^{-1}\mathbf{x}^i(t+1) \\ &\sim (I - R^T T\mathbf{a}^T)^{-1} R^T [R\mathbf{X}^i(t) + T] \end{aligned} \tag{35}$$

where the matrix inversion lemma has been used [22] and $[\cdot]$ denotes the projective coordinates. If we call $T' \doteq R^T T$, then we can write

$$\begin{aligned} \mathbf{x}^i(t+1)^w &\sim \left(I - T'\mathbf{a}^T\right)^{-1}\left[\mathbf{X}^i(t) + T'\right] \\ &\sim \left(I + \frac{T'\mathbf{a}^T}{1 - \mathbf{a}^T T'}\right)\left[\mathbf{X}^i(t) + T'\right] \end{aligned} \tag{36}$$

which may be finally written as

$$\mathbf{x}^i(t+1)\}^w \sim \mathbf{x}^i(t) + \beta^i(t)\mathrm{T'} \tag{37}$$

where $\beta^i(t) = \left(1 + \frac{T'\mathbf{a}^T\mathbf{X}^i(t)}{1 - \mathbf{a}^T T'}\right)$ is a scalar factor. Therefore, the last term can be interpreted as a residual, which is in the direction of the epipole (the projective coordinates of the direction of translation T'). The derivation above is taken from [34].

This representation, consisting in the motion of a plane—encoded by the matrix A—and the residual parallax in the direction of the epipole—encoded by $\beta^i(t)$—is known in the literature as the "plane-plus-parallax" representation, and has been developed in [4], [32], [34].

Now, let us see how warping affects the setup of epipolar geometry. It is immediate to verify that

$$\mathbf{x}^{i^w}(t+1)^T (T' \wedge)\mathbf{x}^i(t) = 0 \qquad T' \in \mathbf{S}^2 \tag{38}$$

and, therefore, the effect of rotation has been canceled out by the image warping. We may represent the overall model as, again, an implicit dynamical system, with parameters on a manifold

$$\begin{cases} \left(\mathbf{Q}\mathbf{x}^i(t)\right)^T \mathbf{x}^{i^w}(t+1) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} = T'\wedge \in so(3) \cap \mathbf{S}^2 \equiv \mathbf{S}^2 \tag{39}$$

where the last equivalence follows from the isomorphism between $so(3)$ and $\mathbb{R}^3$ [5]. Thus, the plane-fixation constraint corresponds to Essential matrices which are of the form $\mathbf{Q} = T'\wedge$. Due to the normalization constraint on T', we have only two degrees of freedom left, and rotation has been fully decoupled from translation. This model may be considered the discrete-time equivalent of the subspace constraint, for it fully decouples structure and rotation, and leaves a dynamic constraint only in the direction of translation.

4 Conclusions

We have proposed a unified framework for modeling “Structure From Motion.” Most of the dynamic models currently used in the literature can be derived following very simple ideas from the theory of dynamical systems. The first unifying concept is the so-called “reduced-order observer,” which allows deriving the coplanarity constraint of Longuet-Higgins [17], [23], [46] and the subspace constraint of Heeger and Jepson [15] as a unique procedure from the basic dynamical model, which is essentially underlying all recursive structure and/or motion estimation techniques. The “Essential filter” [37], and the “Subspace filter” [38] are methods tailored for estimating motion from such constraints, interpreted as implicit dynamical models with parameters on a manifold.

We solve the asymmetry between the continuous-time case, where rotation is easily decoupled from translation, and the discrete-time case, where such a decoupling is not possible, within the context of output stabilization. The constraints resulting from fixating the motion of a point, a line and a plane are derived in a unified fashion as Essential filters constrained to submanifolds of the Essential manifold. This procedure generates a geometric stratification of the Essential manifold, which unifies the work on fixation [12], [33], [42] and the so-called “plane plus parallax” [34], [32] approach in the framework of epipolar geometry [11]. All of these models are no longer treated as *algebraic constraints* on motion and/or structure parameters from a number of views. Rather, they are dynamical systems with unknown parameters on differentiable manifolds. Such dynamical systems are in the particular form of Exterior Differential Systems:

$$\begin{cases} f\left(\mathbf{x}^i, \phi\right)\dot{\mathbf{x}}^i = 0 \\ \mathbf{y}^i = \mathbf{x}^i + n^i \quad \forall\, i = 1 \dots N \end{cases} \qquad \phi \in M \tag{40}$$

where $\mathbf{x}^i \in \mathbb{R}\mathrm{P}^2$ are the projective coordinates of each visible featurepoint and ϕ are the unknown parameters that encode the motion of the viewer relative to the scene. The only thing that changes among different models is the parameter manifold M. We derive similar models in the discrete-time case. The models (20), (24), (28), (31), (39) all fall within this category, where the manifold M is, in each instance, a submanifold of the Essential manifold E, defined in (25). In all cases, the motion parameters may be estimated by identifying the parameters of the corresponding model in the form (40), as we discuss in the companion paper [39].

Despite SFM being a somewhat “old” subject in computer vision, we believe that a few crucial issues are worth discussing:

1) the fundamental difference between *time integration* and *time averaging*,
2) the possibility of achieving time integration with time-discontinuous feature sets (or with optical flow in the limit),
3) the relationship between different SFM models.

With regard to issue 1, any algorithm that averages out estimates obtained from few frames can only operate to the extent in which the few-frames algorithms work. For instance, if we use a two-frame algorithms in the presence of small parallax and high noise, this will lead to biased estimates that, once averaged, will be meaningless. On the other hand, the models presented in this paper represent a way to effectively increase the baseline by integrating information over time. This allows us to work with small inter-frame motion (which makes the correspondence problem simple) while achieving accuracy and robustness typical of large baseline-motions. Furthermore, we can estimate motion even if the visible features drop below five in each frame. Averaging results from two-frames algorithms in this case would not be possible at all. Reduced models can integrate motion information over time even if the feature set changes (issue 2); this is a crucial issue, since it is very difficult to track features over an extended period of time. With regard to issue 3, we often see in the literature “new” algorithms being proposed, which turn out to be variations of existing models, often with a different choice of optimization technique. In this paper we have made an effort to provide the researcher in motion analysis with tools to judge whether a model proposed is in fact new, and how it compares with existing techniques.

References

[1] G. Adiv, “Determining Three-Dimensional Motion and Structure From Optical Flow Generated by Several Moving Objects,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 384-401, July 1985.

[2] A. Azarbayejani and A. Pentland, “Recursive Estimation of Motion, Structure and Focal Length,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, pp. 562-575, June 1995.

[3] J. Barron, D. Fleet, and S. Beauchemin, “Performance of Optical Flow Techniques,” *Int’l J. Computer Vision*, vol. 12, no. 1, pp. 43–78, 1994.

[4] J. Bergen, R. Kumar, P. Anandan, and M. Irani, “Representation of Scenes From Collections of Images,” *Internal Report, Sarnoff Research Center*, 1995.

[5] W. Boothby, *Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 1986.

[6] T. Broida, S. Chandrashekhar, and R. Chellappa, “Recursive 3D Motion Estimation From a Monocular Image Sequence,” *IEEE Trans. Aerospace and Electronic Systems*, vol. 26, no. 4, pp. 639-656, 1990

[7] T. Broida and R. Chellappa, “Estimating the Kinematics and Structure of a Rigid Object From a Sequence of Monocular Frames,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 497-513, June 1991.

[8] R.L. Bryant, S.S. Chern, R.B. Gardner, H.L. Goldshmidt, and P.A. Griffith, *Exterior Differential Systems*. Mathematical Research Institute, Springer Verlag, 1991.

[9] N Cui, J. Weng, and P. Cohen, “Recursive-Batch Estimation of Motion and Structure From Monocular Image Sequences,” *IEEE Trans. Aerospace and Electronic Systems*, 1990.

[10] E. Dickmanns and W. Graefe, “Dynamic Monocular Machine Vision,” *Machine Vision and Applications*, 1988.

[11] O.D. Faugeras, *Three Dimensional Vision, A Geometric Viewpoint*. MIT Press, 1993.

[12] C. Fermüller and Y. Aloimonos, “Tracking Facilitates 3D Motion Estimation,” *Biological Cybernetics*, vol. 67, pp. 259-268, 1992.

[13] E. Gennery, “Tracking Known Three-Dimensional Objects,” *Prof. AAAI Second Nat’l Conf. Artificial Intelligence*, 1982.

[14] V. Guillemin and A. Pollack, *Differential Topology*. Prentice-Hall, 1974.

[15] D. Heeger and A. Jepson, “Subspace Methods for Recovering Rigid Motion I: Algorithm and Implementationm,” *Int’l J. Computer Vision*, vol. 7, no. 2, 1992.

[16] J. Heel, “Direct Estimation of Structure and Motion From Multiple Frames,” *AI Memo 1190*, MIT AI Lab, Mar. 1990.

[17] B.K.P. Horn, "Relative Orientation," *Int'l J. Computer Vision*, vol. 4, pp. 59–78, 1990.

[18] A. Isidori, *Nonlinear Control Systems*. Springer Verlag, 1989.

[19] D. Jacobs, "Linear Fitting with Missing Data," *Proc. IEEE CVPR*, 1997.

[20] A.H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[21] A. Jepson and D. Heeger, "Linear Subspace Methods for Recovering Rigid Motion," *Spatial Vision in Humans and Robots*, Cambridge University Press, 1992.

[22] T. Kailath, *Linear Systems*. Prentice Hall, 1980.

[23] H.C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene From Two Projections," *Nature*, vol. 293, pp. 133–135, 1981.

[24] Q. Luong and O. Faugeras, In preparation, 1997.

[25] L. Matthies, R. Szeliski, and T. Kanade, "Kalman Filter-Based Algorithms for Estimating Depth From Image Sequences," *Int'l J. Computer Vision*, 1989.

[26] S.J. Maybank, *Theory of Reconstruction From Image Motion*. Springer Verlag, 1992.

[27] P. McLauchlan, I. Reid, and D. Murray, "Recursive Affine Structure and Motion From Image Sequences," *Proc. Third ECCV*, 1994.

[28] R.M. Murray, Z. Li, and S.S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[29] J. Oliensis, "Rigorous Bounds for 2-Frame Structure From Motion. *Proc. IEEE CVPR*, 1996.

[30] J. Oliensis and J. Inigo-Thomas, "Recursive Multiframe Structure From Motion Incorporating Motion Error," *Proc. DARPA Image Understanding Workshop*, 1992.

[31] C. Poelman and T. Kanade, "A Paraperspective Factorization Method for Shape and Motion Recovery," *Proc. Third ECCV, LNCS*, vol. 810, Springer Verlag, 1994.

[32] P. Anandan, R. Kumar, and K. Hanna, "Shape Recovery From Multiple Views: A Parallax Based Approach," *Proc. Image Understanding Workshop*, 1994.

[33] D. Raviv and M. Herman, "A Unified Approach to Camera Fixation and Vision-Based Road Following," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 24, no. 8, 1994.

[34] H.S. Sawhney, "Simplifying Motion and Structure Analysis Using Planar Parallax and Image Warping," *Proc. Int'l Conf. Pattern Recognition*, 1994.

[35] J.G. Semple and G.J. Kneebone, *Algebraic Projective Geometry*. Oxford, 1952.

[36] S. Soatto, "3D Structure From Visual Motion: Modeling, Representation and Observability," *Automatica*, vol. 33, no. 9, 1997.

[37] S. Soatto, R. Frezza, and P. Perona, "Motion Estimation Via Dynamic Vision," *IEEE Trans. Automatic Control*, vol. 41, no. 3, 1996.

[38] S. Soatto and P. Perona, "Recursive 3D Visual Motion Estimation Using Subspace Constraints," *Int'l J. Computer Vision*, vol. 22, no. 3, 1997.

[39] S. Soatto and P. Perona, "Reducing "Structure From Motion": General Framework for Dynamic Vision Part 2: Implementation and Experimental Assessment," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 943-961, Sept. 1998.

[40] M. Spetsakis and J. Aloimonos, "A Multiframe Approach to Visual Motion Perception," *Int'l J. Computer Vision*, vol. 6, no. 3, 1991.

[41] R. Szeliski, "Recovering 3D Shape and Motion From Image Streams Using Nonlinear Least Squares," *J. Visual Communication and Image Representation*, 1994.

[42] M.A. Taalebinezhaad, "Direct Recovery of Motion and Shape in the General Case by Fixation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 847-853, Aug. 1992.

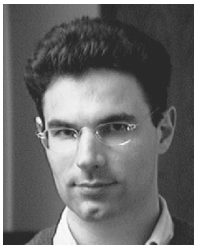
[43] I. Thomas and E. Simoncelli, "Linear Structure From Motion," *Technical Report IRCS 94-26*, Univ. of Pennsylvania, 1994.

[44] T. Tian and C. Tomasi, "Comparison of Approaches to Egomotion Computation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996.

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

[46] J. Weng, N. Ahuja, and T. Huang, "Motion and Structure From Point Correspondences With Error Estimation: Planar Surfaces," *IEEE Trans. Signal Processing*, vol. 39, no. 12, pp. 2,691–2,716, 1991.

[47] J. Weng, N. Ahuja, and T. Huang, "Optimal Motion and Structure Estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 864–884, 1993.



Stefano Soatto received a DEng in electrical engineering from the University of Padova in 1992 and a PhD in control and dynamical systems from the California Institute of Technology in 1996. In 1996-1997, he was with the Division of Applied Sciences at Harvard University. He is currently an assistant professor of electrical engineering at Washington University and research faculty at the University of Udine, Italy. His main research interests are in computational vision and in nonlinear systems and control theory.

Pietro Perona received a DEng in electrical engineering from the University of Padova in 1985 and a PhD in electrical engineering and computer science from the University of California at Berkeley in 1990. He is currently professor of electrical engineering with the California Institute of Technology and adjunct professor with the University of Padova. His main research interests are computational vision, visual psychophysics, and modeling of human vision.