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 different 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 by the rigidity constraint and the perspective projection, 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 all models seen so far in the literature, but we can also derive novel ones.

1 Introduction

Suppose that we are looking at a scene through a moving camera. The problem of "structure from motion" deals with reconstructing both the relative motion between the scene and the camera, and the structure of the scene. We represent the structure of the scene as the position of a number N of *point-features* in 3-D space, and we assume to be able to measure their *perspective projection* onto the 2-D image plane. We also assume that we are able to assess which feature corresponds to which across different views. Alternatively, we may assume that we can measure the *optical flow*, which is the image velocity of brightness patches at a number N of locations on the image, as an approximation of the projection of the 3-D velocity of feature points (see [3] for a review of optical flow/feature tracking techniques).

The basic constraints of rigid motion and the projection map describe in a natural way a dynamical model, whose state encodes the structure of the scene, and whose inputs (or parameters) describe the motion relative to the viewer. 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 *methods*. Which one is the "correct" one? We feel the need to understand the relationships between such methods, and to assess the qualitative and quantitative properties of each one by comparing them on a common ground. Such comparison is not a trivial matter, for any estimation *method* involves two aspects: a *model* that describes the constraints involved in the problem, and an *estimation technique*, for reconstructing the unknows from the model and the data. For each model one may employ different estimation techniques.

We stress the fact that we do not wish to compare existing motion estimation *methods*, for there are many different ones that are based upon non-structural variations of the same models or that employ different estimation techniques. Rather, we wish to evaluate *models* that are structurally different, develop a framework that allows us to justify them all, and to compare their geometric substance and engineering value on a common experimental ground.

We will start from the constraints that "define" structure from motion, namely the rigidity constraint and the perspective projection, and see how they naturally define a dynamical model with unknown parameters. Such a model has structural limitations that do not allow us to estimate its state and identify its parameters from the measurements. Two alternative strategies may be chosen at this point: either we *extend* the dynamical model so as to include in the state the unknown parameters, or we *reduce* it so as to decouple the states from the parameters. The extended model has some shortcomings, which motivate us towards the reduced one. Simple reduction strategies may be applied both for discrete-time models and for continuous-time ones. However, they lead to different outcomes. It is possible to settle such an asymmetry only by allowing an "implicit" reduction of the the dynamical model, by enforcing that some function of its states is held constant.

This paper is concerned with *modeling*. We will see how all models for estimating motion from a dynamical system fall into a special class of implicit dynamical systems with unknown parameters on a manifold. Once a model is proposed, an optimization technique needs to be employed for estimating structure and motion. We do so in a companion paper [35], where we also evaluate all methods on a common experimental ground, which highlights some caveats when reduction is performed with an output-dependent change of coordinates.

1.1 Motion and structure estimation as an optimization problem

Once the geometric constraints involved in the problem (namely the rigidity constraint and the point-wise representation of structure) and the measurement model (for instance perspective projection) have been formalized, one may set up an optimization problem in order to estimate $3N + 6M$ unknown parameters (3 space coordinates for each feature-point and 6 components of motion across M time instants), from $2NM$ image projections of the N points at each of the M images.

A variety of models have been proposed involving structure, motion, and images of feature-points, for instance the coplanarity constraint [20], the subspace constraint [18, 14, 40], the so-called "plane plus parallax" representation [4, 27, 29] and fixation constraints [11]. These constraints have then been exploited for estimating structure and/or motion from image sequences using a number of optimization schemes, either batch, or recursively. Batch optimization techniques from two consecutive frames, based upon the coplanarity constraint, have been presented both in closed-form [20, 42], or iterative [16, 43]. The same holds for the subspace constraint [14]. Multi-frame batch techniques have also been presented, both in closed-form under orthographic or affine projection [26, 41], and iteratively for the case of full perspective projection [1, 23, 25, 37, 38]. In this paper we will be dealing with causal dynamic models for multi-frame processing. In a companion paper [35] we will use such models for designing local recursive observers, such as the Extended Kalman Filter (EKF) [17]. Relatively few schemes for recursive motion estimation exist in the literature, see for instance [2, 6, 7, 9, 15, 23, 25, 32, 37].

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 space of unknown parameters, in order to see whether there are "slices" where the parameters evolve independently in the cost objective. This responds to the need of decomposing a high-dimensional optimization task into the solution of a number of smaller, simpler and better conditioned problems.

In the case of structure and motion estimation, the work of Longuet-Higgins [20] (L-H) pioneered this approach, by decoupling structure from the motion parameters, which he encoded in a 3×3 matrix, called *Essential matrix*. Adiv [1] and Heeger and Jepson [14] (H-J) further decoupled the translational velocity from the rotational velocity.

We will re-derive the constraints of H-J and L-H within a unified procedure. We will start from the dynamical model determined by the rigidity constraint and the perspective projection, and construct the so-called *reduced-order observer* [19] both for the continuous-time and the discrete-time models. These result, respectively, in the subspace constraint and the coplanarity constraint, now interpreted as nonlinear implicit models of a special class (so-called Exterior Differential Systems [8]) with parameters on a manifold. Such a manifold is a 5-dimensional space, called Essential manifold, in the discrete-time case of L-H and the 2-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 manifolds are 4 and 3-dimensional submanifolds of the Essential manifold, when a point or a line are fixated, and the 2-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 4, 3 and 2-dimensional dynamic constraints, the latter being the discrete-time equivalent of the H-J constraint.

1.4 Relation to previous work

The literature on 3-D visual motion estimation comprises a large variety of apparently unrelated constraints involving rigid motion and projection of point-features. We consider equivalent all systems whose state-spaces are identified modulo a diffeomorphism. These include changes of coordinates (world-centered vs. viewer-centered), changes of the reference on the image plane etc. . This paper starts with the standard rigid motion and perspective projection constraints, which are the essential ingredients of the problem and common to all recursive schemes, for instance [2, 7, 15, 21, 23, 36], and derives the constraints of Longuet-Higgins [16, 20, 42] and Heeger and Jepson [14], 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 [11, 28, 39]. 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 [11, 28, 39], the so-called "plane-plus-parallax" representation [4, 27, 29], 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

In this section we are going to establish the notation and formalize the basic constraints that "define" the problem of structure and motion estimation. Such constraints naturally result in a dynamical model. However, we argue that such a model has limited engineering value; this motivates us towards the reduction strategy described in the next sections 3 and 4.

2.1 The basic ingredients: rigid motion and projection

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

We call $\mathbf{X}^i = \begin{bmatrix} X^i & Y^i & Z^i \end{bmatrix}^T \in \mathbb{R}^3$ 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

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

The matrix R belongs to the space of unit-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 feature-point can be written as

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

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 [24]. 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 equation (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. The parameters (T, R) that describe a rigid motion form 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 [24].

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

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

where $\mathbf{x} = \pi(\mathbf{X}) \doteq \begin{bmatrix} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{bmatrix}^T$. $\mathbf{x}$ is usually measured up to some error n, which is well modeled as a white, zero-mean and normally distributed process with covariance Σ_n:

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

In practice, feature tracking and optical flow are subject to various sorts of errors: (a) pixel noise in the image, (b) erroneous correspondence and (c) violations of the brightness constancy assumption [3]. Any algorithm for reconstructing 3-D motion and/or structure in real-time must handle such errors in an automatic fashion, by rejecting outlier measurements due to mismatches, and by exploiting the statistics of the localization error and the redundancy in the measurements in order to minimize their effects. We will briefly discuss how to reject outliers in the companion paper [35].

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 rigid motion. Equations (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 easily by estimating the state of the above dynamical systems (1)(5) or (2)(5) using an observer, for instance in the form of an EKF as in [21, 25, 36]. Vice-versa, if the trajectory of the points in space was known, their motion parameters could be estimated by solving (2) as a linear algebraic equation. 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[1] 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. Unfortunately, the model that comes out of the basic constraints is "driftless", in the sense that all of its dynamics depends upon the unknown parameters: if we call ξ the state of our system, and u the unknown parameters, then the dynamic equation of the model can be written in the form $\dot{\xi} = f(\xi) + g(\xi)u$ with the drift vector field $f(\xi) = 0$. This means that all constraints obtained from time-derivatives of the output couple the unknown states with the unknown parameters. Furthermore, it can be proven that only the first derivative produces independent constraints on the unknowns, and therefore it is not possible to unravel both the state of the model and its parameters [31].

[1] Such a space is called the "observability space", and is constructed by computing Lie derivatives of the output along the state vector field.

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 u to be the unknown parameters, rather than u itself. Then it is possible to insert u into the dynamical model, and make simplifying assumptions about its time derivatives. Alternatively, we may try to "reduce" the original model by decoupling the states from the parameters. These alternative strategies are discussed in the next two sections 2.3 and 2.4.

2.3 "Think big": dynamic extension and observers

Let us enlarge the state of the model described by (1)(5) or (2)(5) by including the unknown motion parameters into the state. To do so, we have to assume some dynamics for such parameters, i.e. some model of the form

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

where in essence we have transferred our ignorance on $T, R, V, \Omega,$ onto f_T, f_R, f_V, f_Ω and n_T, n_R, n_V, n_Ω, which we do not know. 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 a dynamic system and inserted into the model. For instance, the simplest constraint of constant velocity may be written as

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

and inserted in the state of the model (1) or (2). In such a case f_* is a linear map, and $n_* = 0$, where $*$ stands for T, R, V, Ω. The next simplest model is a first-order random walk (Brownian motion), where n_* are appropriately defined white, zero-mean and Gaussian noises. It is important to 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. If the reader is not comfortable with modeling motion as a first-order random walk, we suggest reading the companion paper [35] first.

Once we have inserted the parameters into the state, the problem of recovering simultaneously motion and structure becomes that of estimating the state of the augmented model using an observer[2], whose state-space is now a bit more complicated than it used to, for the motion parameters belong either to the Lie-group of Euclidean motions, $(T, R) \in SE(3)$, or to the corresponding Lie-algebra, $(V, \Omega\wedge) \in se(3)$. If the motion parameters are modeled as a first-order random walk, and the measurement noise is white, zero-mean and Gaussian, then one may set up an observer (or "filter", for instance an EKF) for estimating both structure and motion simultaneously from the augmented model:

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

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 is essentially common to all recursive motion estimation methods seen in the literature. Non-structural 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 of points $N = 100$, which is a typical number of feature-points in images of realistic sequences. Then the state of the filter just described has dimension 305, since there are 300 coordinates of the points, 6 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

[2] 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 an introduction to the basic concepts of linear observers, see for instance [19]. 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 on Kalman filtering, see for instance [17].

number of visible features $N(t)$ changes in time, which causes the filter to have a variable dimension, with the problem of initializing new states without affecting the continuity of the existing states. 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 [23] a method is proposed for dealing with such a situation, which uses a "variable state-dimension filter".

Furthermore, the EKF performs a local update on the residual of the prediction with a gain computed on the linearization of the model which, in the case just described, is not locally observable [31]. As an intuitive argument, first observe that the model described by (8) is "block triangular", in the sense that the dynamic of each feature point $\mathbf{X}^i$ depends only on itself and on the motion parameters, but not on other points $\mathbf{X}^j \mid i \neq j$. This means that, as far as the observability is concerned, it does not matter how many points are visible (of course accuracy is affected). 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 about Y causes the projection to escape in finite time.

2.4 "Think small": reducing the order of the model

The alternative to extending the original model (1)(5) or (2)(5) is to try to decouple the states from the unknown parameters, and reduce the original mixed estimation/identification task into either a state estimation independent of the unknown parameters, or a parameter identification independent of the unknown states. The states or parameters that have been eliminated can be recovered a-posteriori, once the remaining states or parameters have been estimated, using a standard observer.

In the next two sections we will see two different approaches for reducing the model by either *explicitly* decoupling the states of the original model from its unknown parameters, or *implicitly* imposing that some function of the states and the parameters is held constant.

3 Explicit reduction

In this section we will explore techniques for decoupling the unknown states of the original models (1)(5) or (2)(5) from the unknown parameters. We will first apply "verbatim" the idea of the so-called "reduced-order observer" for eliminating two out of the three space-coordinates for each point. We will then push the same idea for further decoupling all the states corresponding to structure and end up with a dynamical model where the only unknowns are the motion parameters. In the continuous-time case we will end up with a model having only two unknown parameters, which correspond to the direction of translation, while in the discrete-time case it is not possible to decouple the unknown rotation parameters from the model. Such an asymmetry motivates alternative decoupling methods, which we discuss in the next section 4.

3.1 The basic reduced-order observer: simultaneous depth and motion estimation

The reduced-order observer [19] is a long-established technique for reducing the dimension of an observer for a dynamical system. The basic idea consists in "solving" 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:

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

The previous measurement equation is now the identity $y = y$. We may re-write 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 [19], 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 $({}^tT_{t_0}, {}^tR_{t_0})$ 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:

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

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

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 et al., although derived with different motivations. In [2], an extended model is considered 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 [35] 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. One possible way of getting rid of such an ambiguity is to saturate the filter along any direction corresponding to a state subject to the ambiguity by setting the variance of the model error to zero [17]. For instance, one may initialize an arbitrary point to be at distance one.

The model above (13) is structurally similar to (8), and still suffers the shortcomings outlined in section 2.3, for it includes the structure parameters and the state equation is "diagonal". The model lacks local observability [31], and it makes it difficult to handle occlusions and appearance of new features in a principled way, since the errors in the transient following the introduction of any new feature propagate into the current estimate of the motion parameters. These are the main reasons that 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.

3.2 Pushing the model reduction: structure-independent motion estimation

In the previous sections we have seen how the constraints of rigid motion and perspective projection naturally define a nonlinear dynamical system, whose state comprises the structure and motion parameters. We have also seen how the dimension of such a state can be reduced by the number of measurements, using the concept of the *reduced-order observer*.

Although we have reduced the dimension of the state, it still involves structure parameters and, therefore, it can vary in time due to occlusions and appearance of new features. The next step consists in applying the same idea of the reduced-order observer to the already-reduced model in order to get rid of structure parameters altogether.

3.2.1 Continuous-time: the Subspace model

Let us apply the idea of the reduced-order observer twice to the model of equation (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 $N+5$ 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 [14] 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

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

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

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

where

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

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

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

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

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

Since there is an overall scaling factor ambiguity, only the direction of translation $\frac{V}{\|V\|}$ can be recovered, which we represent by imposing $V \mid \|V\| = 1$. The above constraint describes a nonlinear dynamical system of a very peculiar kind, called *Exterior Differential Systems* [8] (EDS), with the parameters V constrained on the unit-sphere $\mathbf{S}^2$. We may therefore write our dynamical model as

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

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"

$$\begin{bmatrix} \frac{1}{\hat{Z}^1} & \dots & \frac{1}{\hat{Z}^N} & \hat{\Omega} \end{bmatrix}^T = \mathcal{C}^{\dagger}\dot{\mathbf{x}}. \tag{21}$$

We will see in the companion paper [35] how to perform the identification of models of the form (20).

3.2.2 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 to be used for eliminating the depth parameters is the so-called "Epipolar geometry" (see [10] for a review), which essentially resorts to the well-known coplanarity constraint, first derived by Longuet-Higgins [20].

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 $\mathbb{RP}^2$ [30]. When expressed with respect to a common reference frame, for example that at time t, we may write the triple product as

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

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

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

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

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

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 6 (or 5 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 [32], and for a thorough description of its algebraic structure, see for instance [10, 22].

3.3 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 [14]. 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 is what we do in the next section.

4 Implicit reduction: motion from fixation

In this section we explore how to reduce the order of the observer by stabilizing some particular functions of the state.

4.1 Output stabilization and geometric stratification

Suppose that we are told that some of the states of a dynamical model are fixed. Then we may as well 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 $f : M \to \mathbb{R}$ is smooth, and $0 = f(\mathbf{P})$ is a regular value, then the pre-image $f^{-1}(0) \subset M$ is a submanifold of M [13], 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, 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.

4.2 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 pre-process the image by considering changes of coordinates that depend upon the outputs, without acting on the support of 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

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

Figure 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.**

estimation are concerned, the two methods are equivalent. A few gaze-control techniques which guarantee exponential convergence are described in [33], while image-shift registration techniques that achieve fixation in a single step are described, for instance, in [39].

Fixating a point and a line 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 pre-processing, the image, as described in section 4.5.1.

4.3 Stabilization of a point (fixation)

Let us assume that we have applied any 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 from the fixation loop. 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

$$T(R, v) \doteq \begin{bmatrix} -R_{13} & -R_{23} & -R_{33} + v \end{bmatrix}^T, \tag{26}$$

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 [33] 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

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

where

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

Estimating motion reduces to identifying the above dynamical system with parameters on $\mathcal{S}^4$.

4.4 Stabilization of a point and a line

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 $\mathcal{S}^4$ of the Essential manifold $\mathbf{E}$, 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

$$\mathcal{S}^3 = \mathcal{S}^4 \ \cap \ \{R = e^{\left[\begin{array}{ccc} \omega_1 & \omega_2 & 0 \end{array} \right]^T \wedge}\}. \tag{30}$$

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

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

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

4.5 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 "deform" the images of the sequence in order to account for the motion of the plane.

4.5.1 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 $\Pi = \{\mathbf{X}_\pi \in \mathbb{R}^3 \mid \mathbf{a}^T\mathbf{X}_\pi = 1\}$. We indicate with $\mathbf{x}_\pi \in \mathbb{RP}^2$ 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 3-D 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 4 or more point-correspondences on the image-plane, we may solve the above equation for the 8 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, 27, 29].

4.5.2 Plane-plus-parallax representation

In the previous subsection, we have assumed that all points of the scene lie on the plane Π. Note that, if we apply the above procedure to an unstructured cloud of points, and we estimate the matrix A using total-least-squares [12] from equation (33), then we compensate for the *average plane* in the scene.

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

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

where $[\cdot]$ denotes the projective coordinates. If we call $T' \doteq R^T T$, then we can write

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

which may be finally written as

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

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 [29].

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, 27, 29].

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

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

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.

5 Conclusions

In this paper 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 [16, 20, 42] and the subspace constraint of Heeger and Jepson [14] as a unique procedure from the basic dynamical model, which is essentially common to all recursive structure and/or motion estimation techniques. The "Essential filter" [32], and the "Subspace filter" [34] are methods tailored for estimating motion from such constraints, interpreted as implicit dynamical models with parameters on a manifold.

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, is resolved in 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 [11, 28, 39] and the so-called "plane plus parallax" [29, 27] approach in the framework of epipolar geometry [10].

The novelty is that 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 of a very peculiar form, which is that of Exterior Differential Systems:

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

where $\mathbf{x}^i \in \mathbb{RP}^2$ are the projective coordinates of each visible feature-point 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 a companion paper [35].

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1985.

[2] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure and focal length. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1995.

[3] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *Int. J. of Computer Vision*, 12(1):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 3-d motion estimation from a monocular image sequence. *IEEE trans. AES*, 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 Anal. Mach. Intell.*, 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. AES*, 1990.

[10] O. D. Faugeras. *Three dimensional vision, a geometric viewpoint.* MIT press, 1993.

[11] C. Fermüller and Y. Aloimonos. Tracking facilitates 3-d motion estimation. *Biological Cybernetics (67), 259-268*, 1992.

[12] G. Golub and C. Van Loan. *Matrix computations.* Johns Hopkins University Press, 2 edition, 1989.

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

[14] D. Heeger and A. Jepson. Subspace methods for recovering rigid motion i: algorithm and implementation. *Int. J. Comp. Vision vol. 7 (2)*, 1992.

[15] J. Heel. Direct estimation of structure and motion from multiple frames. *AI Memo 1190, MIT AI Lab*, March 1990.

[16] B.K.P. Horn. Relative orientation. *Int. J. of Computer Vision*, 4:59–78, 1990.

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

[18] A. Jepson and D. Heeger. Linear subspace methods for recovering rigid motion. *Spatial Vision in Humans and Robots, Cambridge University Press*, 1992.

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

[20] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[21] L. Matthies, R. Szeliski, and T. Kanade. Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. of computer vision*, 1989.

[22] S. J. Maybank. *Theory of reconstruction from image motion.* Springer Verlag, 1992.

[23] P. McLauchlan, I. Reid, and D. Murray. Recursive affine structure and motion from image sequences. *Proc. of the 3 ECCV*, 1994.

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

[25] J. Oliensis and J. Inigo-Thomas. Recursive multi-frame structure from motion incorporating motion error. *Proc. DARPA Image Understanding Workshop*, 1992.

[26] C. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. *Proc. of the 3 ECCV, LNCS Vol 810, Springer Verlag*, 1994.

[27] P. Anandan R. Kumar and K. Hanna. Shape recovery from multiple views: a parallax based approach. *Proc. of the Image Understanding Workshop*, 1994.

[28] D. Raviv and M. Herman. A unified approach to camera fixation and vision-based road following. *IEEE Trans. on Systems, Man and Cybernetics vol. 24, n. 8*, 1994.

[29] H. S. Sawhney. Simplifying motion and structure analysis using planar parallax and image warping. *Proc. of the Int. Conf. on Pattern Recognition*, 1994.

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

[31] S. Soatto. Observability/identifiability of rigid motion under perspective projection. In *Proc. of the 33rd IEEE Conf. on Decision and Control*, pages 3235–3240, Dec. 1994.

[32] S. Soatto, R. Frezza, and P. Perona. Motion estimation via dynamic vision. *IEEE Trans. on Automatic Control*, in press, March 1996.

[33] S. Soatto and P. Perona. Motion from fixation. CDS Technical report CIT-CDS-95-006, California Institute of Technology, February 1995, submitted to ECCV 96.

[34] S. Soatto and P. Perona. Recursive 3-d visual motion estimation using subspace constraints. *Int. J. of Computer Vision*, in press, 1996.

[35] S. Soatto and P. Perona. Reducing "structure from motion" 2: experimental evaluation. *submitted to the IEEE trans. PAMI*, Nov. 1995.

[36] S. Soatto, P. Perona, R. Frezza, and G. Picci. Recursive motion and structure estimation with complete error characterization. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR*, pages 428–433, New York, June 1993.

[37] M. Spetsakis and J. Aloimonos. A multi-frame approach to visual motion perception. *Int. J. Computer Vision 6 (3)*, 1991.

[38] R. Szeliski. Recovering 3d shape and motion from image streams using nonlinear least squares. *J. visual communication and image representation*, 1994.

[39] M. A. Taalebinezhaad. Direct recovery of motion and shape in the general case by fixation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1992.

[40] I. Thomas and E. Simoncelli. Linear structure from motion. *Technical Report IRCS 94-26, Univ. of Pennsylvania*, 1994.

[41] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. of Computer Vision*, 9(2):137–154, 1992.

[42] J. Weng, N. Ahuja, and T. Huang. Motion and structure from point correspondences with error estimation: planar surfaces. *IEEE Trans. Signal Processing*, 39(12):2691–2716, 1991.

[43] J. Weng, N. Ahuja, and T. Huang. Optimal motion and structure estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15:864–884, 1993.