Motion Estimation via Dynamic Vision*

Stefano Soatto† **Ruggero Frezza‡** **Pietro Perona†‡**

† California Institute of Technology 116-81, Pasadena–CA 91125

Tel. (818) 395-3272 Fax. (818) 395-2137 e-mail soatto@systems.caltech.edu

‡ Università di Padova, Dipartimento di Elettronica, Padova–Italy

keywords: Dynamic vision, motion estimation, nonlinear identification, estimation on manifolds, exterior differential systems, implicit Extended Kalman Filter.

Abstract

Estimating the three-dimensional motion of an object from a sequence of projections is of paramount importance in a variety of applications in control and robotics, such as autonomous navigation, manipulation, servo, tracking, docking, planning, surveillance. Although "visual motion estimation" is an old problem (the first formulations date back to the beginning of the century), only recently tools from nonlinear systems estimation theory have hinted at acceptable solutions.

In this paper we formulate the visual motion estimation problem in terms of identification of nonlinear implicit systems with parameters on a topological manifold, and go on to propose a dynamic solution either in the local coordinates or in the embedding space of the parameter manifold. Such a formulation has structural advantages over previous recursive schemes, since the estimation of motion is decoupled from the estimation of the structure of the object being viewed, and therefore it is possible to handle occlusions in a principled way.

1 Introduction

"Understanding" the geometry and kinematics of the environment is a basic requirement for humans to successfully accomplish tasks such as walking, driving, recognizing and grasping objects. It has been one of the principal goals of Artificial Intelligence, starting from the early seventies, to build machines that recognize the *shape* and *motion* of objects within the environment. The

*Research funded by the California Institute of Technology, a fellowship from the University of Padova, a fellowship from the "A. Gini" Foundation, an AT&T Foundation Special Purpose grant, ONR grant N0014-93-1-0990 and grant ASI-RS-103 from the Italian Space Agency. This work is registered as Technical Report CIT-CDS 94-004, California Institute of Technology, 1994. Reduced version in the Proceedings of the 33^{rd} IEEE Conf. on Decision and Control, Lake Buena Vista – FL, Dec. 1994.

goal is far from being reached and, indeed, it opens a new and exciting avenue of research in nonlinear systems theory.

Although the first formulations of the visual motion estimation problem date back to the beginning of the century [31, 77], only within recent years tools from control and estimation theory have been applied [3, 8, 9, 28, 29, 35, 52, 56, 60, 66], with rather encouraging results in traditionally difficult applications, such as autonomous vehicle navigation [18, 19, 20], vision-based tracking and servo [12, 21, 42, 44], vision-based manipulation [5, 21, 42], docking [19, 37], vision-based planning [14], active sensing [69].

As the reliability and the performance of the algorithms improves, vision starts being acknowledged in the automatic control community as a powerful and versatile sensor to measure motion, position and structure of the environment, and the appropriate tools from nonlinear estimation/identification theory start being exploited [16, 30, 59, 60]. The implementation of sophisticated vision algorithms running in real time is not too far from becoming reality and, due also to the evolution of computer hardware, vision will be soon included "in the loop" of many control systems.

"Vision in the loop" raises new and interesting problems of system theoretic flavor, ranging from distributed filtering and processing of large amounts of sensory data to the analysis and control of new classes of dynamical systems. Crucial issues in the use of vision as a sensor in control systems are, for example, nonlinear observability and identifiability in a projective geometric framework, and estimation and control on peculiar topological manifolds.

In this paper we will be mainly concerned with the "visual motion estimation" problem:

> *Given a sequence of images taken from a moving camera, reconstruct the relative 3-D motion between the camera and the environment (or scene).*

Since our goal is that of posing the visual motion estimation problem within a system-theoretical framework, we need to specify a "description" of the environment and the motion of the viewer. We will restrict our attention to "static" scenes or, equivalently, to portions of it which are moving rigidly relative to the viewer.

The existing methods for motion estimation may be classified, depending on the scene descriptors employed, as point-based, line-based, curve-based or model-based. We will focus on the simplest case, when the scene is described by a number of point features in the Euclidean 3-D space. For line-based schemes see [72, 81] and references therein. The curve-based approach has been addressed in [2, 13, 71].

The point-based methods may be further classified in terms of the camera model in question. The simplest cases assume either parallel projection [58, 73, 74, 75] or ideal perspective projection (pinhole model, see [23]). More articulated camera models in terms of projective transformations allow parallel and perspective projection as a subcase [3, 25, 61, 70]. We will be mainly concerned with the classical pinhole model; however, our schemes generalize to other camera representations and may estimate the camera model along with visual motion (camera self-calibration, see [25, 61]). Other schemes recover projective, non-metric structure and motion independent on the camera parameters [22, 54, 58].

Motion reconstruction methods may be further classified in terms of the data processing technique as 2-frames schemes (see for example [38, 49, 78]), multiframe-batch methods [70, 75] or recursive algorithms.

In the last decade a variety of schemes has been proposed for recursively reconstructing structure for known motion [52], motion for known structure [9, 28, 29] or both structure and motion [3, 35, 56, 60, 66]. In general, given either the relative motion or the shape of the object being viewed, the other can be recovered easily, since the problem can be reduced to a linear estimation task. When neither the motion nor the shape of the scene are known, the problem of estimating both of them from visual information becomes a remarkably difficult one. We find that a crucial step in tackling such an estimation task consists in being able to *decouple* the estimation of motion from the estimation of structure. This decoupling has dramatic consequences also from the practical standpoint, since it allows integrating motion information in the presence of occlusions in the image plane, whereas previous structure *and* motion estimation schemes could integrate motion information only to the extent in which all initial feature points were still visible (as in [3, 36]).

In this paper we present a framework for estimating rigid motion *independent* of the structure (shape) of the scene. The estimates of motion can later be fed to any recursive "structure from known motion" module in order to estimate scene structure [52, 56, 66].

Organization of the paper

The next section (2) of this paper has introductory purpose: we establish the notation and review some basic concepts in the representation of rigid motion. Section 2.2 describes an alternative representation based upon the so-called "essential matrices", which were introduced in [49]. In section 3, we introduce a novel nonlinear implicit dynamical model with motion coded as a vector of parameters constrained to the space of essential matrices.

We show in section 4 how to carry out the identification of the model introduced. There we propose two methods, one based upon a dynamic model in the local coordinates of the parameter manifold, the other based upon a dynamic model in its embedding space, projected onto the parameter manifold. The formulation makes use of results contained in the appendix. An alternative iteration for identifying the model in local coordinates is described in appendix A and tested in the experimental section 6. There we discuss benchmark experiments that highlight the peculiarity of each scheme. Extensive experiments in real-world scenes have been conducted by the authors as well as by other researchers [7], and show consistent performance in situations where previously proposed techniques fail.

The methods introduced have some degree of generality, as a number of other problems in Computational Vision may be cast in the same framework, as discussed in section 5.

Finally, conclusions are drawn in section 7.

2 Background and notation

2.1 Representation of rigid motion

A transformation $g : \mathbb{R}^3 \to \mathbb{R}^3$ of the three-dimensional Euclidean space is a *rigid motion* if it preserves the Euclidean distance between points p_i and the cross product between vectors q_i:

$$\begin{aligned} d(p_1, p_2) &= d\left(g(p_1), g(p_2)\right) \quad \forall\, p_1, p_2 \in \mathbb{R}^3 \\ g_*(q_1 \wedge q_2) &= g_*(q_1) \wedge g_*(q_2) \quad \forall\, q_1, q_2 \in \mathrm{T}\mathbb{R}^3 \sim \mathbb{R}^3 \end{aligned}$$

where g_* is the transformation induced on vectors: $q \doteq p_2 - p_1 \Rightarrow g_*(q) \doteq g(p_2) - g(p_1)$, and $\mathrm{T}\mathbb{R}^3$ is the tangent space to $\mathbb{R}^3$. If we represent the points p_i in coordinates $\mathbf{X}_i \doteq [X_i \ Y_i \ Z_i]^T$ relative to some orthonormal reference frame, we may characterize a rigid motion as a translation of the origin and a rotation of the reference frame. The matrices which represent the change of basis induced by a rotation of the reference are orthogonal with unit determinant; such matrices form a Lie group of dimension 3, called $SO(3)$ (Special Orthogonal group of transformations of $\mathbb{R}^3$) [1, 6, 68, 76]. We write the action of a rigid motion on $\mathbb{R}^3$ as $g = (T, R)$ with $T \in \mathbb{R}^3$ and $R \in SO(3)$, such that

$$g(\mathbf{X}) = R\mathbf{X} + T. \tag{1}$$

The set of rigid motions has the structure of a Lie group of dimension 6, and is called $SE(3)$, (Special Euclidean transformations of $\mathbb{R}^3$). It is sometimes useful to *embed* $SE(3)$ in the linear group $\mathcal{GL}(4)$ (the General Linear group of nonsingular 4×4 matrices) using *homogeneous coordinates* $\bar{\mathbf{X}} \doteq [\mathbf{X}^T \ 1]^T \in \mathbb{R}^4$. Each rigid motion g is then represented as a matrix:

$$G = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \mid T \in \mathbb{R}^3 \ , \ R \in SO(3).$$

We will use the notation $g(t) \doteq (T(t), R(t))$ when emphasizing the time-dependence of g. The group operations in $SE(3)$ coincide with the group operations of $\mathcal{GL}(4)$, so that the composition of rigid motions may be represented as a matrix multiplication: $g_1 \circ g_2 = G_1 G_2$. The price we pay for such a simplification is that we have to embed $SE(3)$, which is a six-dimensional manifold, into $\mathbb{R}^{4\times4}$, which has dimension 16.

The tangent space at the origin of $SE(3)$ has the structure of a Lie algebra, and is called $se(3)$. Elements of $se(3)$ are called “twists” in the robotics literature [55], and may be represented in so-called “Plücker coordinates” as $v\wedge \doteq \dot{g}g^{-1} = \begin{bmatrix} \Omega\wedge & V \\ 0 & 0 \end{bmatrix}$, where $V \in \mathbb{R}^3$ and $\Omega\wedge \doteq \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}$ belongs to the Lie algebra of the skew-symmetric matrices $so(3) \doteq \{S | S^T = -S\}$, which is isomorphic to $\mathbb{R}^3$ via $\Omega\wedge \leftrightarrow \Omega = [\omega_1 \ \omega_2 \ \omega_3]^T \in \mathbb{R}^3$. We will use the same symbol v for an element of $se(3)$ and its Plücker coordinates. The reader interested in a complete treatment of the concepts sketched here may consult for instance [6, 41, 45, 55, 68].

The reason why the representation introduced above is appealing is that all the (compact) one-parameter subgroups of a Lie group can be characterized using the *exponential map*. For instance, $\forall \ v \ \in \ se(3) \ , \ g(t) = e^{(v\wedge)t}$ is a one-parameter subgroup of $SE(3)$. An explicit expression for the exponential map on $SE(3)$ is given by

$$\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} = \exp\left(\begin{bmatrix} \Omega\wedge & V \\ 0 & 0 \end{bmatrix}\right)$$

where

$$R \doteq e^{(\Omega\wedge)} \tag{2}$$

$$T \doteq \mathcal{T}(\Omega)V \tag{3}$$

$$\mathcal{T}(\Omega) \doteq \frac{1}{\|\Omega\|}\left[\left(I - e^{(\Omega\wedge)}\right)(\Omega\wedge) + \Omega\Omega^T\right]. \tag{4}$$

The exponential map may be inverted locally for computing V and Ω from R and T, since the matrix $\mathcal{T}(\Omega)$ is invertible when $\|\Omega\| \in (0\ ,\ \pi)$. In the case $\|\Omega\| = 0$, the exponential map is defined simply by

$$R \doteq I \tag{5}$$
$$T \doteq V. \tag{6}$$

Note that the exponential map, together with the isomorphism of $so(3)$ with $\mathbb{R}^3$, gives a local coordinate parametrization of $SE(3)$, which in the robotics literature is called the "canonical" (exponential) representation. The Rodrigues' formulae [55] provide a convenient way of computing the exponential map.

If we consider the composite action of time on the Euclidean space through $SE(3)$, we can motivate the characterization of $v\wedge = \dot{g}g^{-1}$ as "velocity". Consider a point p which has moved between t_0 and t according to some motion: $p(t) = g(t)p(t_0)$. Then we have

$$\dot{p}(t) = \dot{g}(t)p(t_0) = \dot{g}(t)g^{-1}(t)g(t)p(t_0) = v(t) \wedge p(t)$$

and, in coordinates,

$$\dot{\mathbf{X}}(t) = \Omega(t) \wedge \mathbf{X}(t) + V(t), \tag{7}$$

where V and Ω represent the translational and rotational velocities of the viewer's moving frame [55].

2.2 The "essential manifold"

A rigid motion may be represented as a point in the Lie group $SE(3)$, which can be embedded in the linear space $\mathcal{GL}(4)$ (and hence exploit the matrix product as composition rule) and is in local correspondence with $\mathbb{R}^6$ via the exponential coordinates and the isomorphism between $se(3)$ and $\mathbb{R}^6$, as seen in the previous section. We now discuss an alternative matrix representation of rigid motion which is more "compact", in the sense that it can be embedded in a linear space of smaller dimensions. Such a representation is derived from the so-called "essential matrices" introduced by Longuet-Higgins [49].

Consider a point $g = (T, R) \in SE(3)$, then $T\wedge \in so(3)$ is a skew-symmetric matrix. Now define the space of *"essential matrices"* as

$$E \doteq \{SR \mid \ R \in SO(3)\ ,\ S = (T\wedge) \in so(3)\} \subset \mathbb{R}^{3\times 3}. \tag{8}$$

Clearly the essential space does not inherit the group structure from the sum of matrices in $\mathbb{R}^{3\times 3}$, since $\mathbf{Q}_1, \mathbf{Q}_2 \in E$ does not imply $\mathbf{Q}_1 + \mathbf{Q}_2 \in E$. One possible way of imposing the group structure is by forcing a group morphism with $SE(3)$, for which it is necessary to "unfold" T, R from $\mathbf{Q} = (T\wedge)R$, perform the group operation on $SE(3)$ and then collapse the result into $\mathbf{Q}$. We will see later in this section a way of unfolding an essential matrix into its rotation and translation components.

The essential space has many interesting geometrical properties: it is an algebraic variety [53] and a topological manifold of dimension 6. Later on we will provide a characterization of a local coordinate chart. The essential space may also be identified with $TSO(3)$, the tangent bundle of the rotation group, defined as $TSO(3) \doteq \cup_{R\in SO(3)} T_R SO(3)$. This proves that the essential space is indeed a differentiable manifold of dimension 6 [63].

The following theorem, due to Huang and Faugeras and reported by Maybank [53], gives a simple characterizing property of the essential space.

Theorem 2.1 *(Huang and Faugeras, 1989)*
Let $\mathbf{Q} = \mathbf{U}\Sigma\mathbf{V}^T$ be the Singular Value Decomposition (SVD) [32] of a matrix in $\mathbb{R}^{3\times 3}$. Then

$$\mathbf{Q} \in E \Leftrightarrow \Sigma = \Sigma_0 = diag\{\lambda\ \lambda\ 0\} \mid \lambda \in \mathbb{R}^+.$$

Proof:

($\Rightarrow$) let $\mathbf{Q} = SR \mid R \in SO(3), S \in so(3)$; $\sigma(\mathbf{Q})$, the set of singular values of $\mathbf{Q}$, is such that $\sigma(\mathbf{Q}) = \sqrt{\sigma(\mathbf{Q}\mathbf{Q}^T)}$. Next observe that $\mathbf{Q}\mathbf{Q}^T = SS^T = -S^2$. Also $\forall\ S \in so(3)\ \exists\ !\ T \in \mathbb{R}^3 \mid S = (T\wedge)$, and the singular values of S^2 are $\{\|T\|^2, \|T\|^2, 0\}$. Hence if $\mathbf{Q} \in E$, it has two equal singular values and a zero singular value.

($\Leftarrow$) let $\mathbf{Q} = \mathbf{U}\Sigma_0\mathbf{V}^T$ be a Singular Value Decomposition. Let furthermore $R_Z(\frac{\pi}{2}) = \begin{bmatrix} 0 & -1 & 0 \\ +1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ be a rotation of $\frac{\pi}{2}$ about the axis $[0\ 0\ 1]^T$, then

$$\mathbf{Q} = \mathbf{U}\Sigma_0\mathbf{V}^T = \mathbf{U}\Sigma_0 R_Z^T(\pm\frac{\pi}{2})\mathbf{U}^T\mathbf{U}R_Z(\pm\frac{\pi}{2})\mathbf{V}^T.$$

Now call $R \doteq \mathbf{U}R_Z^T(\pm\frac{\pi}{2})\mathbf{V}^T$ and $S \doteq \mathbf{U}\Sigma_0 R_Z(\pm\frac{\pi}{2})\mathbf{U}^T$; it is immediate to see that $RR^T = R^TR = I$ and $S^T = -S$. From the uniqueness of the SVD, it follows that this decomposition is unique, modulo the sign in $R_Z(\pm\frac{\pi}{2})$. **Q.E.D.**

Remark 2.1 *Note that, since $\mathbf{Q} \doteq \mathbf{U}\Sigma\mathbf{V}^T \in E \Leftrightarrow \Sigma = \mathrm{diag}\{\lambda\ \lambda\ 0\}$, there is one degree of freedom in defining the the basis components of the subspaces $< \mathbf{V}_{.3} >^\perp$ and $< \mathbf{U}_{.3} >^\perp$, which corresponds to rotating the orthogonal bases $< \mathbf{V}_{.1}, \mathbf{V}_{.2} >$ and $< \mathbf{U}_{.1}, \mathbf{U}_{.2} >$ about their orthogonal complements. However, the effects cancel out in the multiplications when defining R and S as in the above proof.*

Local coordinates of the essential manifold

For any given rigid motion $(T, R) \in SE(3)$, there exists an essential matrix $\mathbf{Q}$ defined by $\mathbf{Q} \doteq (T\wedge)R$. We are interested now in the inverse problem: *given an essential matrix $\mathbf{Q}$, can we extract its rotational and translational components? Is the correspondence $\mathbf{Q} \leftrightarrow (T, R)$ unique?*
Consider the map

$$\begin{aligned} \Phi : E &\rightarrow \mathbb{R}^3 \times SO(3) \rightarrow \mathbb{R}^3 \times \mathbb{R}^3 \\ \mathbf{Q} &\mapsto \begin{bmatrix} \pm\|\mathbf{Q}\|\mathbf{U}_{.3} \\ \mathbf{U}R_Z(\pm\frac{\pi}{2})\mathbf{V}^T \end{bmatrix} = \begin{bmatrix} T \\ e^{\Omega\wedge} \end{bmatrix} \mapsto \begin{bmatrix} T \\ \Omega \end{bmatrix} \end{aligned} \tag{9}$$

where $\mathbf{U}, \mathbf{V}$ are defined by the Singular Value Decomposition (SVD) [32] of $\mathbf{Q} = \mathbf{U}\Sigma\mathbf{V}^T$; $\mathbf{U}_{.3}$ denotes the third column of $\mathbf{U}$ and $R_Z(\frac{\pi}{2})$ is a rotation of $\frac{\pi}{2}$ about the axis $[0\ 0\ 1]^T$. Note that the map Φ defines the local coordinates of the essential manifold modulo two signs; therefore, the map Φ associates to each element of the essential space four distinct points in local coordinates. This ambiguity may be resolved in the context of the visual motion estimation problem by imposing the *"positive depth constraint"*, which means that each visible point lies in front of the viewer. In a case like this we will be able to identify a unique local coordinates homeomorphism, as discussed in section 3.3. The inverse map is simply

$$\begin{array}{rcl} \Phi^{-1} : \mathbb{R}^3 \times \mathbb{R}^3 & \rightarrow & E \\ \begin{bmatrix} T \\ \Omega \end{bmatrix} & \mapsto & (T\wedge)e^{(\Omega\wedge)}. \end{array}$$

Projection onto the essential manifold

Theorem 2.1 suggests a simple "projection" of a generic 3×3 matrix onto the essential manifold: let us define

$$\begin{array}{rcl} pr_{<E>} : \mathbb{R}^{3\times 3} & \rightarrow & E \\ M & \mapsto & \mathbf{U}\ \mathrm{diag}\{\lambda, \lambda, 0\}\ \mathbf{V}^T \end{array} \tag{10}$$

where $\mathbf{U}, \mathbf{V}$ are defined by the SVD of $M = \mathbf{U}\ \mathrm{diag}\{\sigma_1, \sigma_2, \sigma_3\}\ \mathbf{V}^T$, and $\lambda \doteq \frac{\sigma_1+\sigma_2}{2}$. It follows from the properties of the SVD [32] that $pr_{<E>}(M)$ minimizes the Frobenius distance from M to the essential manifold [33, 53].

3 Structure-independent motion estimation modeled as the recursive identification of nonlinear implicit systems

In this section we start off with the constraints of rigid motion and perspective projection, which define a "natural" nonlinear dynamical model for the 3-D coordinates of each visible feature point (structure). Motion is an unknown parameter of the model which is constrained on $SE(3)$. If we represent motion on the essential manifold, instead, it is possible to remove the 3-D structure of the scene from the model, ending up with a nonlinear and implicit dynamical model for the (measured) projective coordinates of the visible features, with motion as an unknown parameter constrained on E. This allows us to decouple the estimation of motion from the 3-D structure of the scene, which has many advantages, for it allows dealing with occlusions of feature points and to cross regions of motion-space which render the "natural" model unobservable [59].

Consider the position of a rigid set of feature points in 3-D space. We call $\mathbf{X} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T \in \mathbb{R}^3$ the coordinates of a generic point 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 (see figure 1). The relative motion between the

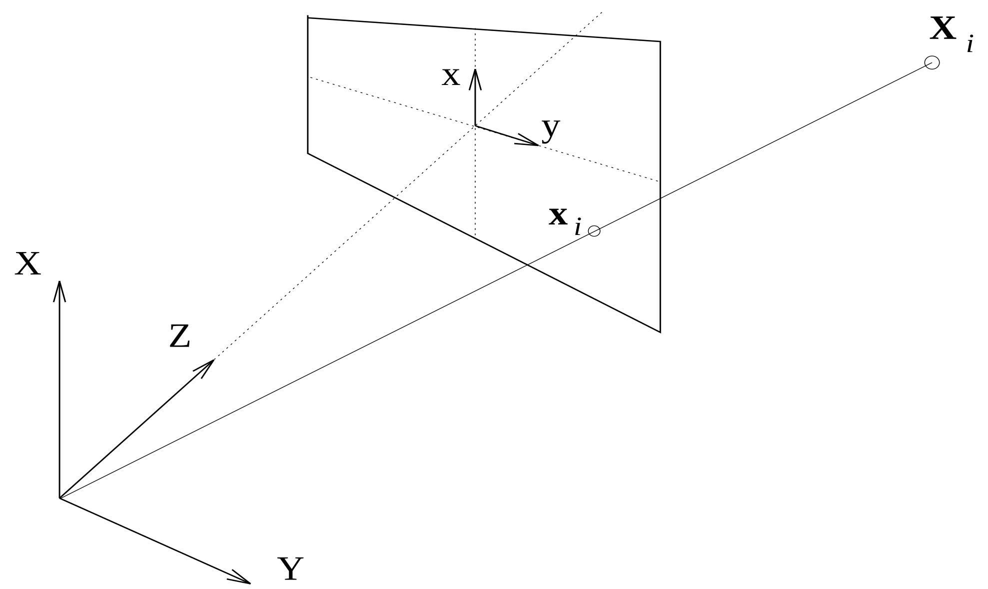

Figure 1: *Point-based visual motion estimation; the viewer-centered reference frame.*

camera and the object (or scene) is described by a **rigid motion** $g(t) = (T(t), R(t)) \in SE(3)$ which induces an instantaneous velocity $(V(t), \Omega(t))$ such that

$$\dot{\mathbf{X}}(t) = \Omega(t) \wedge \mathbf{X}(t) + V(t). \tag{11}$$

If we assume the velocity to be constant between the samples, so that $\Omega(t)$ and $V(t)$ represent the local coordinates of the rigid motion of the camera between time t and $t+1$, the we can write

$$\mathbf{X}(t+1) = R(t)\mathbf{X}(t) + T(t) \tag{12}$$

where (T, R) are related to (V, Ω) via the exponential map.

What we are able to measure is the **perspective projection** π of the point features onto the image plane, which for simplicity we represent as the real projective plane. The projection map π associates to each $p \neq 0$ its projective coordinates as an element of $\mathbb{R}\mathrm{P}^2$ (see figure 1):

$$\begin{array}{rcl} \pi : \mathbb{R}^3 - \{0\} & \to & \mathbb{R}\mathrm{P}^2 \\ \mathbf{X} & \mapsto & \mathbf{x} \doteq \left[\begin{array}{ccc} x & y & 1 \end{array} \right]^T = \pi(\mathbf{X}) \doteq \left[\begin{array}{ccc} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{array} \right]^T. \end{array} \tag{13}$$

We usually measure $\mathbf{x}$ up to some error n, which is well modeled as a white, zero-mean and normally distributed process with covariance R_n:

$$\mathbf{y} = \mathbf{x} + n \qquad\qquad n \in \mathcal{N}(0, R_n).$$

In summary, when we represent the scene structure using points in the Euclidean 3-D space, the visual motion estimation problem is *defined* by the constraints of rigid motion and perspective projection

$$\left\{ \begin{array}{l} \mathbf{X}_i(t+1) = R(t)\mathbf{X}_i(t) + T(t) \ ; \ \mathbf{X}_i(0) = \mathbf{X}_{i0} \\ \mathbf{y}_i(t) = \pi(\mathbf{X}_i(t)) + n_i(t) \qquad \forall i = 1 \ldots N. \end{array} \right. \tag{14}$$

The above is a nonlinear dynamical model having the 3-D structure of the scene in the state. Estimating motion amounts to identifying the above model with the parameters T, R constrained on $SE(3)$. However, we do not know $\mathbf{X}_{i0}$, so that we end up with a mixed estimation/identification task which proves extremely difficult [59].

In the next section we will show how representing motion on the essential manifold allows us to decouple the estimation of motion from the structure parameters $\mathbf{X}_i$.

3.1 Structure-independent models for motion estimation

When a rigid object is moving between two time instants t and $t+1$, the coordinates $\mathbf{X}$ of a point at time t, their correspondent $\mathbf{X}'$ at time $t+1$, and the translation vector T are coplanar (fig. 2). Their triple product is therefore zero. This is true of course also for $\mathbf{x}$, $\mathbf{x}'$ and T, since $\mathbf{x}$ is the projective coordinate of $\mathbf{X}$ and therefore the two identify the same direction in $\mathbb{R}^3$, interpreted as the "ray-space" model of $\mathbb{RP}^2$ [57]. When expressed with respect to a common reference frame, for example that at time t, we may write the triple product as

$$\mathbf{x}_i'^T \left(T \wedge (R\mathbf{x}_i)\right) = 0 \qquad\qquad \forall\ i \ = \ 1 : N. \tag{15}$$

As it turns out, the above constraint is not only a consequence of rigid motion, but also suffices to characterize it, once five or more such constraints are given [53, 49]. Let us define $\mathbf{Q} \doteq (T\wedge)R$, so that the above coplanarity constraint, which is known as the "essential constraint" or the "epipolar constraint", becomes

$$\mathbf{x}_i'^T \mathbf{Q} \mathbf{x}_i = 0 \qquad\qquad \forall i = 1 \dots N. \tag{16}$$

Estimating motion corresponds to identifying the model

$$\begin{cases} (\mathbf{Q}\mathbf{x}_i)^T \mathbf{x}_i' = 0 & \mathbf{Q} \in E \\ \mathbf{y}_i = \mathbf{x}_i + n_i & \forall i = 1 \dots N, \quad n_i \in \mathcal{N}(0, R_{n_i}). \end{cases}$$

Since the constraint (16) is linear in $\mathbf{Q}$, we use the improper notation

$$\chi(t+1)\mathbf{Q}(t) \doteq \chi_{\mathbf{x}'(t),\mathbf{x}(t)} \mathbf{Q}(t) = 0 \qquad \chi \in \mathbb{R}^{N \times 9}$$

where χ is an $N \times 9$ matrix combining $\mathbf{x}_i, \mathbf{x}_i'$ and $\mathbf{Q}$ is interpreted as a nine-dimensional vector obtained by stacking the columns of the 3×3 matrix $\mathbf{Q}$ on top of each other. In the following we will not distinguish between $\mathbf{Q}$ interpreted as a matrix in $\mathbb{R}^{3\times 3}$ and a nine-dimensional column vector. The generic row of χ has the form $[xx',\ yx',\ x',\ xy',\ yy',\ y',\ x,\ y,\ 1\ ]$. We will use the notation $\chi(t)$ when emphasizing the time-dependence, while we will write $\chi_{\mathbf{x}'(t),\mathbf{x}(t)}$ when highlighting which vectors are used for constructing χ.

3.2 The "essential filter"

Since the essential constraint is an homogeneous equation, and hence defined only up to a scale factor, we may restrict $\mathbf{Q}$ to belong to $\mathbf{S}^8$ instead of $\mathbb{R}^9$. It is customary to set the norm of translation to be unitary; this can be done without loss of generality, as long as translation is

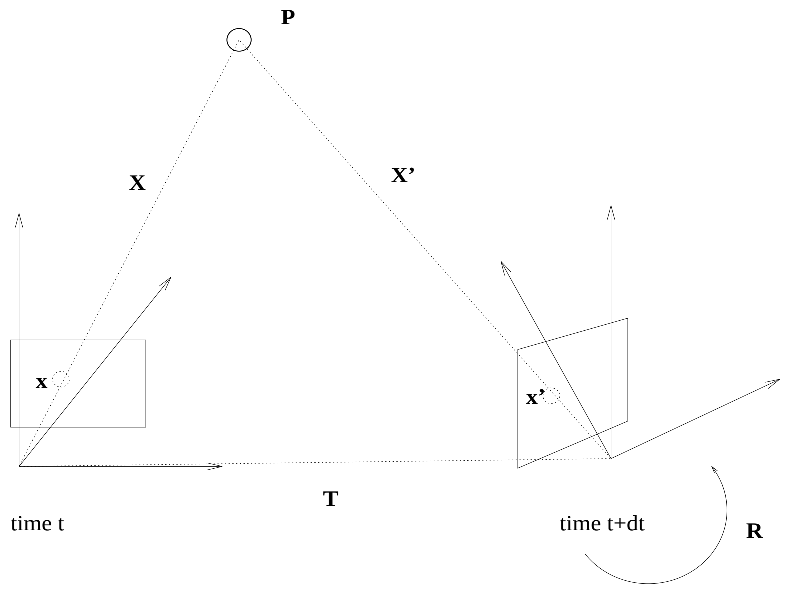

Figure 2: *The coplanarity constraint*

not zero. The zero-norm translation case can be dealt with separately, and we discuss it in section 3.5. For simplicity we now assume $\|\mathbf{Q}\|_2 = \|T\| = 1$. At each time instant we have a set of N constraints in the form

$$\chi_{\mathbf{x}'(t),\mathbf{x}(t)}\mathbf{Q}(t) = 0,$$

therefore, $\mathbf{Q}$ lies at the intersection between the essential manifold and the linear variety $\chi^{-1}_{\mathbf{x}'(t),\mathbf{x}(t)}(0)$ (see fig. 3).

Note that, even after imposing unit norm, there is still a sign indeterminacy in $\mathbf{Q}$, which accounts for the two possible solutions $\mathbf{Q}_1 = +\mathbf{Q}$ and $\mathbf{Q}_2 = -\mathbf{Q}$ of the essential constraint. These become four after being transformed to local coordinates. This ambiguity can be overcome by imposing the positive depth constraint as it will be done in section 3.3.

As time progresses, the point $\mathbf{Q}(t)$, corresponding to the actual motion, describes a trajectory on E (and a corresponding one in local coordinates) according to

$$\mathbf{Q}(t+1) \doteq \mathbf{Q}(t) + n_{\mathbf{Q}}(t).$$

The last equation is indeed just a *definition* of the right-hand side, as we do not know $n_{\mathbf{Q}}(t)$. The identity of $n_{\mathbf{Q}}(t)$ will be unrevealed in section 4.2. For now, we will consider the previous equation to be a discrete-time dynamical model for $\mathbf{Q}$ on the essential manifold, with $n_{\mathbf{Q}}$ as *unknown* input. If we accompany it with the essential constraint, we get

$$\begin{cases} \mathbf{Q}(t+1) \doteq \mathbf{Q}(t) + n_{\mathbf{Q}}(t) & \mathbf{Q} \in E \\ 0 = \chi_{\mathbf{y}'(t),\mathbf{y}(t)}\mathbf{Q}(t) & \\ \mathbf{y}_i = \mathbf{x}_i + n_i & \forall\ i = 1 \ldots N. \end{cases} \tag{17}$$

Now the visual motion estimation problem is characterized as the estimation of the state of the above model, which is defined on the essential manifold. It can be seen that the

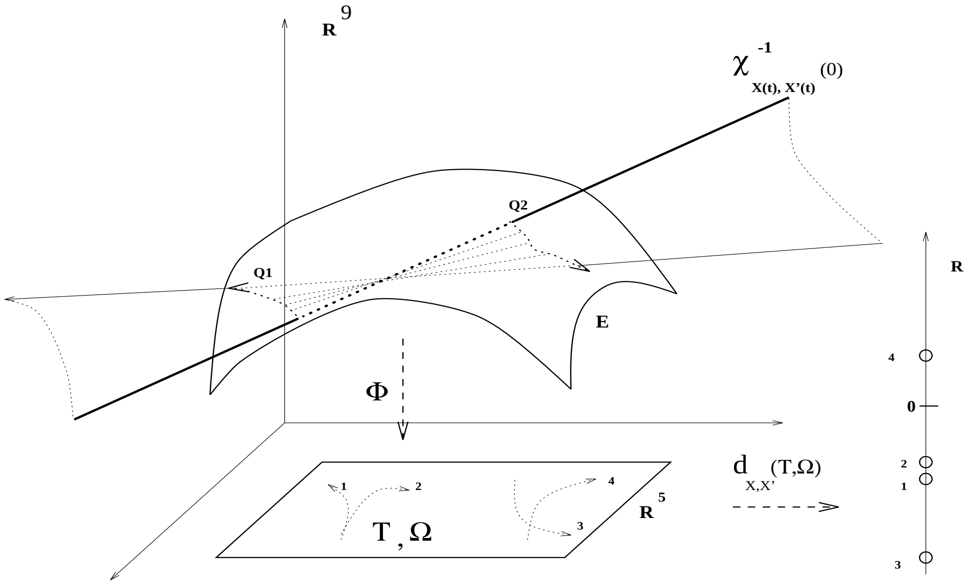

Figure 3: *Structure of the motion problem on the essential space.*

system is "linear" (both the state equation and the essential constraint are linear in $\mathbf{Q}$). E, however, is not a linear space. We will see how to solve the estimation task in section 4. The observability/identifiability of the essential models is addressed in [59]. It is proven that the model is globally observable under general position conditions. Such conditions are satisfied if the viewer's path and the visible objects cannot be embedded in a (proper) quadric surface of $\mathbb{R}^3$ and if all the visible points cannot be embedded on a plane [50, 59].

3.3 Choosing the local coordinates for the essential manifold

The map Φ introduced in eq. (9) defines the local coordinates of the essential space modulo a sign in the direction of translation and in the rotation angle of R_Z. Therefore, the map Φ associates to each element of the essential space 4 distinct points in local coordinates. This ambiguity can be resolved by imposing the *"positive depth constraint"*, i.e. that each visible point lies in front of the viewer [27, 33, 49, 50, 79]. Consider one of the four local counterparts of $\mathbf{Q} \in E$, and the triangulation function $d_{\mathbf{x},\mathbf{x}'} : E \to \mathbb{R}^{1+1}$, with $d_{\mathbf{x},\mathbf{x}'}(\mathbf{Q}) = [Z, Z']^T$, which gives the depth of each point as a function of the projection and the motion parameters (see figure 2). Note that it is locally smooth away from zero translation. Therefore, given any N point-matches with projective coordinates $\mathbf{x}^i, \mathbf{x}'^i$ we may use Φ as a local coordinate chart for the following set, which we call the *"normalized essential manifold"*:

$$\begin{aligned} \mathbf{E} &\doteq E \cap d^{-1}_{\mathbf{x},\mathbf{x}'}(\mathbb{R}^2_+)^N \cap \mathbf{S}^8 = \\ &= \{\mathbf{Q} = SR | R \in SO(3), S \doteq T\wedge \in so(3), \|T\| = 1, d_{\mathbf{x}^i,\mathbf{x}'^i}(\mathbf{Q}) > 0 \;\forall i = 1..N\} \end{aligned} \quad (18)$$

where $\mathbb{R}_+$ is the positive open half space of $\mathbb{R}$, and $d^{-1}_{\mathbf{x},\mathbf{x}'}$ denotes the preimage of $d_{\mathbf{x},\mathbf{x}'}$. Consider Φ restricted to $\mathbf{E}$. It follows from the properties of the SVD that Φ is continuous and, furthermore, bijective. The normalized essential manifold thus defined is a topological manifold of dimension 5, since we have imposed the metric constraint $\|T\| = 1$.

3.4 Propagating scale information

It is well known [49] that from visual information it is only possible to recover the structure and the motion modulo a scale factor multiplying the translational velocity and the depth of the visible points. In fact, we cannot distinguish between a car moving on a street, and a car which is "twice as big, twice as far, and moving twice as fast". Such a scale ambiguity is captured by the homogeneous nature of the essential constraint (16). However, as soon as we are given some scaling information about the scene *at one time instant* – for example the size of the car – we can rescale the scene and the estimated velocity to its appropriate values.

Suppose we are given the distance between two visible "reference" points in space $\|\mathbf{X}_{r1} - \mathbf{X}_{r2}\| = \rho$. Once the motion has been estimated, with a normalized translational velocity, it can be used to estimate the "normalized structure" $\tilde{\mathbf{X}}_i$ via triangulation [66]. By matching the distance between the reference points in the normalized structure with its reference value, we can rescale both the depth of each point and the direction of translation simply by $\|\tilde{\mathbf{X}}_{r1} - \tilde{\mathbf{X}}_{r2}\| = \rho\|T\|$.

3.5 Dealing with zero-translation

So far we have assumed that $\|T\| \neq 0$, and we have defined the normalized essential manifold based upon the constraint $\|T\| = 1$. It is easy to see that the condition $\|T\| = 0$ defines a "thin-set" in the parameter space. Due to the noise in the measurements, there is always a translation which is least-squares compatible with the observations. However, one may ask what happens when the system is close to such a configuration. When translation is almost zero, there is little parallax in the projected coordinates of the visible objects, which makes the estimates of the depth and those of the direction of translation ill-conditioned.

Luckily enough, we do not need to worry about the structure of the scene, since it does not enter our dynamic model, and about the direction of translation, since its estimate will be weighted by the scale, which is exactly $\|T\| \cong 0$. However, we would still like to estimate the correct rotational velocity. Here the definition of the normalized essential manifold comes at hand. In fact, the estimation scheme will estimate some direction of translation $\hat{T}$ such that $\|\hat{T}\| = 1$ regardless the scale of T, so that the correct rotational component of the local coordinates can be computed. In the experimental section we will show an experiment in which the system crosses a region in the parameter space where $T = 0$ and $\Omega \neq 0$.

Remark 3.1 *The one just described is a crucial feature of the method proposed. In fact, all previous schemes based upon simultaneous structure and motion estimation [3, 52, 56] become ill-conditioned when close to zero-norm translation, since it is a non-observable configuration for the model (14) and the measurements can no longer be used for updating (or maintaining) the current estimates of structure.*

Remark 3.2 *The essential constraint (16) defines a unique essential matrix (up to scale) only if 8 or more point matches are given. If 5 or more matches are available, one may extract directly the motion parameters from the essential constraint (up to a finite number* *of solutions). Early motion estimation schemes from two frames, based upon the essential matrices, needed at least 5 or 8 point matches in order to estimate motion [38, 49, 78]. However, since the essential filter integrates motion over time, it does not need to have a minimum number of features visible at each time instant, as long as the observability conditions are satisfied [59]. Therefore, using the essential filter allows us to maintain the motion estimates even when crossing regions of the ambient space with less than 5 visible features.*

4 Solving the estimation task

At this point we are ready to address the problem of recursively estimating motion from an image sequence. There are two approaches that may be derived naturally from the formulation introduced in section 3.

The first approach we describe consists of composing the equations (17) with the local coordinate chart Φ, ending up with a *nonlinear* dynamical model for motion in $\mathbb{R}^5$. At this point we have to make some assumptions about motion: since we do not have any dynamical model, we will assume a statistical model. In particular, we will assume that motion is a *first order random walk in* $\mathbb{R}^5$ (see fig. 4 top). The problem then is to estimate the state of a nonlinear system on a linear space driven by white, zero-mean Gaussian noise (see fig. 4 bottom).

In the second approach we change the model for motion: in particular we assume motion to be a *first order random walk in* $\mathbb{R}^9$ *projected onto the essential manifold* (see fig. 4 top). We will see that this leads to a method for estimating motion that consists in solving at each step a *linear estimation* problem in the linear embedding space and then "projecting" the estimate onto the essential manifold (see fig. 4 bottom).

It is very important to understand that these are <u>modeling assumptions</u> about motion which can be validated only a-posteriori. In general we observe that the first method solves a strongly nonlinear problem with techniques which are based upon the linearization of the system about the current reference trajectory, so that the linearization error may be relevant. The second method does not involve any linearization, whereas it imposes the constraint of belonging to the essential manifold in a weaker way. Note that each method produces, together with the motion estimates, the variance of the estimation error, which is to be used by the subsequent modules of the structure from motion estimation scheme [66].

4.1 Estimation in local coordinates

Consider composing the system (17) with the map Φ defined in (9) restricted to the normalized essential manifold $\mathbf{E}$:

$$\begin{aligned} \Phi : \mathbf{E} &\rightarrow \mathbf{S}^2 \times \mathbb{R}^3 \rightarrow \mathbb{R}^5 \\ \mathbf{Q} &\mapsto \xi \doteq \begin{bmatrix} T \\ \Omega \end{bmatrix} \end{aligned}$$

where T is expressed in spherical coordinates of radius one. Then the system in local coordinates becomes

$$\begin{cases} \xi(t+1) = \xi(t) + n_\xi(t) \; ; \; \xi(t_0) = \xi_0 \\ 0 = \chi_{\mathbf{y}(t),\mathbf{y}'(t)} \mathbf{Q}(\xi(t)) + \tilde{n}(t). \end{cases} \tag{19}$$

Motion may be modeled as a first order random walk: $n_\xi(t) \in \mathcal{N}(0, R_\xi)$ for some R_ξ which is referred to as the variance of the model error. While the above assumption is somewhat arbitrary and can be validated only a posteriori, it is often safe to assume that the noise in the measurements $\mathbf{y}(t), \mathbf{y}'(t)$ are white zero-mean Gaussian processes with variance R_n. The second order statistics of the induced noise $\tilde{n}$ is a delicate issue that is discussed in appendix A.

The estimation scheme for the model above, which takes into account the correlation of the error $\tilde{n}$, is reported in appendix A. A simplified version is obtained by approximating $\tilde{n}$ with a white process (note that $\tilde{n}$ is correlated only within one time step). The resulting scheme is based upon an Implicit Extended Kalman Filter (IEKF), which is derived in appendix B. We summarize here the equations of the estimator. Call $C \doteq \left(\frac{\partial \chi \mathbf{Q}}{\partial \xi}\right)$ and $D \doteq \left(\frac{\partial \tilde{\chi} \mathbf{Q}}{\partial \mathbf{x}}\right)$, then we have

Prediction step:

$$\begin{aligned} \hat{\xi}(t+1|t) &= \hat{\xi}(t|t) \; ; \; \hat{\xi}(0|0) = \xi_0 & (20) \\ P(t+1|t) &= P(t|t) + R_\xi \; ; \; P(0|0) = P_0 & (21) \end{aligned}$$

Update step:

$$\begin{aligned} \hat{\xi}(t+1|t+1) &= \hat{\xi}(t+1|t) - L(t+1)\chi(t+1)\mathbf{Q}(\hat{\xi}(t+1|t)) & (22) \\ P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)R_{\tilde{n}}(t+1)L^T(t+1) & (23) \end{aligned}$$

Gain:

$$\begin{aligned} L(t+1) &= P(t+1|t)C^T(t+1)\Lambda^{-1}(t+1) & (24) \\ \Lambda(t+1) &= C(t+1)P(t+1|t)C^T(t+1) + R_{\tilde{n}}(t+1) & (25) \\ \Gamma(t+1) &= I - L(t+1)C(t+1) & (26) \end{aligned}$$

Residual variance:

$$R_{\tilde{n}}(t+1) = D(t+1)R_n D^T(t+1). \tag{27}$$

Note that $P(t|t)$ is the variance of the motion estimation error which is used as variance of measurement error from subsequent modules of the structure from motion estimation scheme [66]. A similar formulation of the IEKF was used by Di Bernardo et al. [17]. Similar expressions were also used before in the literature on specific applications; the first instance to our knowledge was in the recursive computation of the Hough transform [15].

4.2 Estimation in the embedding space

Suppose that motion, instead of being a random walk in $\mathbb{R}^5$, is represented in the essential manifold as the "projection" of a random walk through $\mathbb{R}^9$ (see fig. 4 top).
We define the operator $\oplus$ that takes two elements in $\mathbb{R}^{3\times3}$, sums them and then projects the result onto the essential manifold:

$$\begin{aligned} \oplus : \mathbb{R}^{3\times3} \times \mathbb{R}^{3\times3} &\rightarrow E \\ M1, M2 &\mapsto \mathbf{Q} = pr_{<E>}(M1 + M2), \end{aligned}$$

where the symbol "$+$" is the usual sum in $\mathbb{R}^{3\times3}$. With the above definitions our model for motion becomes simply

$$\mathbf{Q}(t+1) = \mathbf{Q}(t) \oplus n_{\mathbf{Q}}(t), \tag{28}$$

where $n_{\mathbf{Q}}(t) \in \mathcal{N}(0, R_{n_{\mathbf{Q}}})$ is a white zero-mean Gaussian noise in $\mathbb{R}^9$. If we substitute the above equation into (17), we have again a dynamical model on a Euclidean space (in our case $\mathbb{R}^9$) driven by white noise. The essential estimator is the least variance filter for the above model, and corresponds to a linear Kalman filter update in the embedding space, followed by a projection onto the essential manifold. In principle, an approximate gain could be precomputed offline for each possible configuration of motion and feature positions.

Prediction step:

$$\begin{aligned} \hat{\mathbf{Q}}(t+1|t) &= \hat{\mathbf{Q}}(t|t)\ ;\ \hat{\mathbf{Q}}(0|0) = \mathbf{Q}_0 & (29) \\ P(t+1|t) &= P(t|t) + R_{\mathbf{Q}}\ ;\ P(0|0) = P_0 & (30) \end{aligned}$$

Update step:

$$\begin{aligned} \hat{\mathbf{Q}}(t+1|t+1) &= \hat{\mathbf{Q}}(t+1|t) \oplus L(t+1)\chi(t+1)\hat{\mathbf{Q}}(t+1|t) & (31) \\ P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)R_{\tilde{n}}(t+1)L^T(t+1) & (32) \end{aligned}$$

Gain:

$$\begin{aligned} L(t+1) &= -P(t+1|t)\chi^T(t+1)\Lambda^{-1}(t+1) & (33) \\ \Lambda(t+1) &= \chi(t+1)P(t+1|t)\chi^T(t+1) + R_{\tilde{n}}(t+1) & (34) \\ \Gamma(t+1) &= I - L(t+1)\chi(t+1). & (35) \end{aligned}$$

5 Further problems in dynamic vision which may be formulated as identification of Exterior Differential Systems

5.1 Dynamic self-calibration

So far, we have taken the camera to be an ideal perspective projection of unit focal length. When the camera model is a more general projective transformation in $\mathbb{RP}^2$, eq. (16) does not hold.

However, a similar constraint may be derived based on the epipolar geometry as

$$\mathbf{x}'^T_i F \mathbf{x}_i = 0 \qquad\qquad \forall i = 1 \dots N. \tag{36}$$

The matrix F is called "fundamental matrix". It specifies the relation between each point and its corresponding epipolar line [25]. If the camera is represented as a 3×4 matrix $\left[\begin{array}{c|c} \mathbf{A} & 0 \end{array}\right]$ where

$$\mathbf{A} \doteq \begin{bmatrix} fs_x & 0 & -i_0 \\ 0 & fs_y & -j_0 \\ 0 & 0 & 1 \end{bmatrix}$$

is the internal parameter matrix[1], then it can be shown that

$$\mathbf{A}^T F \mathbf{A} \ \in \mathbf{E} \tag{37}$$

is an *essential matrix.*

The Fundamental matrix has been originally introduced by Faugeras. In [25] the matrix F is estimated from the (linear) constraint (36), and then its structure (37) is *imposed a posteriori,* by solving a set of polynomial equations known as Kruppa equations. Such equations are, unfortunately, poorly conditioned, and the scheme is extremely sensitive to noise. Furthermore, temporal coherence of the camera model is not exploited. If we substitute (37) into (36) we get a dynamic model

$$\begin{cases} (\mathbf{A}^{-T} Q \mathbf{A}^{-1} \mathbf{x}_i)^T \mathbf{x}'_i = 0 & Q \ \in \mathbf{E} \\ \mathbf{y}_i = \mathbf{x}_i + n_i & \forall\ i = 1 \dots N. \end{cases}$$

Estimating the camera parameters, along with rigid motion, may then be formulated as *identification of the above exterior differential model, where the parameters are on the manifold* $\mathbf{E} \times AF$, and AF is the set of affine transformations of $\mathbb{R}^2$ represented in homogeneous coordinates. This formulation has been derived in [61].

5.2 Motion from weak-perspective

Alternative camera models may be employed in the same framework, for example the so-called "weak" or "affine" perspective, consisting of a parallel projection onto a plane, followed by a perspective projection of the plane onto the image. In such a case, the fundamental matrix has the simple form [25, 43]

$$F = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{bmatrix},$$

where

$$a \ = \ -R_{23}$$

[1] f is the focal length, (i_0, j_0) the coordinates of the intersection between the optical axis and the image plane, and (s_x, s_y) the pixel sizes along the image plane coordinates. The deviation from 90^o of the angle between the optical axis and the CCD surface is usually on the order of 1^o, and we may therefore neglect it.

$$
\begin{aligned}
b &= R_{13} \\
c &= sR_{23}R_{11} - sR_{13}R_{21} \\
d &= sR_{23}R_{12} - sR_{13}R_{22} \\
e &= R_{23}T_x - R_{23}T_y,
\end{aligned}
$$

and s is a scale factor. The state manifold in this case is $\mathbf{S}^4$ [62].

5.3 Subspace motion factorization for estimating direction of heading

Consider the derivative of the output of the differential version of the basic model (14):

$$\dot{\mathbf{x}}_i(t) = [\mathcal{A}(\mathbf{x}_i, V(\theta,\phi)) \mid \mathcal{B}(\mathbf{x}_i)] \begin{bmatrix} \frac{1}{Z(t)_i} \\ \Omega(t) \end{bmatrix},$$

where $\mathcal{A}_i \doteq \mathcal{A}(\mathbf{x}_i, V) \doteq \begin{bmatrix} V_1 - x_i V_3 \\ V_2 - y_i V_3 \end{bmatrix}$, $\mathcal{B}_i \doteq \mathcal{B}(\mathbf{x}_i) = \begin{bmatrix} -x_i y_i & 1 + x_i^2 & -y_i \\ -1 - y_i^2 & x_i y_i & x_i \end{bmatrix}$ and $V \in \mathbf{S}^2$ is represented in local coordinates as $V(\theta,\phi)$. Observing N points, one may write

$$\dot{\mathbf{x}} = \mathcal{C}(V, \mathbf{x})[\frac{1}{Z_1}, \ldots, \frac{1}{Z_N}, \Omega]^T \doteq \mathcal{C}\mathbf{d}^T,$$

where

$$\mathcal{C}(V,\mathbf{x}) \doteq \begin{bmatrix} \mathcal{A}_1 & & & \mathcal{B}_1 \\ & \ddots & & \vdots \\ & & \mathcal{A}_N & \mathcal{B}_N \end{bmatrix}.$$

Under the usual rank conditions, we may compute the least-squares approximation of $\mathbf{d}$ as

$$\hat{\mathbf{d}} = \mathcal{C}^\dagger \begin{bmatrix} \dot{\mathbf{x}}_1 \\ \vdots \\ \dot{\mathbf{x}}_N \end{bmatrix} \doteq \mathcal{C}^\dagger \dot{\mathbf{x}}$$

where $\dagger$ indicates the pseudo-inverse. Therefore, the motion field specifies the constraint [34]

$$\dot{\mathbf{x}} = \mathcal{C}\mathcal{C}^\dagger \dot{\mathbf{x}} \Rightarrow \mathcal{C}^\perp(V,\mathbf{x})\dot{\mathbf{x}} = 0,$$

where $\mathcal{C}^\perp \doteq I - \mathcal{C}\mathcal{C}^\dagger$ indicates the orthogonal complement. Heeger and Jepson [34] proposed to estimate the direction of translation by minimizing the two norm of the above constraint over $V \in \mathbf{S}^2$. They minimize by extensive search over all possible directions (θ, ϕ).

Indeed, it is immediate to see [65] that the problem of estimating the direction of translation can be rephrased as the problem of *identifying the following Exterior Differential System [10], with parameters V on a sphere*, embedded in $\mathbb{R}^3$:

$$\begin{cases} \mathcal{C}^\perp(V,\mathbf{x})\dot{\mathbf{x}} = 0 & V \in \mathbf{S}^2 \\ \mathbf{y}_i = \mathbf{x}_i + n_i & \forall i = 1 \ldots N. \end{cases}$$

The "projection" onto the manifold is defined, in this case, simply as $pr_{\mathbf{S}^2}(V) \doteq \frac{V}{\|V\|}$, and the same techniques described in the previous section can be used for carrying on the estimation.

6 Experiments

In this section we describe two experiments on real image sequences and one simulation experiment, in order to unreveal the different features of each scheme and their behavior when close to singular configurations in the motion space (e. g. pure rotation about the optical axis).

6.1 Simulation experiments

We have generated a cloud of 20 feature points at random within a cubic volume of side 1 m, placed 1.5 m ahead of the viewer. The scene was viewed under perspective projection from an image plane of 500×500 pixels with a focal length of 1, corresponding to a visual field of approximately 50^o. Gaussian noise with 1 pixel std was added to the measured projections, according to the performance of the most current feature-tracking schemes [4]. The observer was then made to navigate around the cloud with constant velocity for 50 time instants (frames), after which, the observer stopped translating and only rotated about its center of projection for 25 frames, inverting the direction after 15 of them. Finally, the observer resumed its roto-traslational motion in order to return to the initial configuration.

This experiment is interesting from many extents: first of all, for part of the sequence the model is in a singular configuration, since the translational velocity is zero. Indeed, as we have discussed in section 3.5, the schemes proposed still recover some normalized direction of translation, and the correct rotational velocity. Once the appropriate scaling information has been inserted, full translation is correctly estimated. Secondly, in the first and the last part of the experiment, the motion is designed such that the effects of translation and rotation produce the same variation, up to first order, in the derivative of the observations. This is a well-known ambiguous stimulus in which it is difficult to distinguish the effects of rotation from those of translation.

We have systematically varied the conditions of the experiments, by changing the distance in space from the cloud of dots between 1 m and 5 m , the initial conditions between 0% and 1000% off the true value, the level of measurement noise between 0 and 2 pixels and the number of visible points between 1 and 100.

It is interesting to notice that, while previous schemes based upon the essential matrix needed at least 8 [49] or 5 [38] visible points at each time instant, here we can allow any number of points even below the threshold of 5, since we integrate over time the motion information.

The behavior of the different filters was consistent, with a graceful degradation of the estimates as the noise level increases, and a need for more precise initial conditions as the noise increases and the number of visible points diminishes. The performance of the filter saturates as the number of visible points increases beyond 20. The performance also degrades as the points move far away from the viewer and as the structure approaches a plane. Under these conditions, in fact, the matrix χ approaches rank 6, rather than its normal rank of 8 [24, 26].

We have tested the essential filter in local coordinates, both implemented using the IEKF and the two-D iteration described in appendix A, and the essential filter in the embedding space. We now comment on the performance of each filter on the reference simulation experiment, highlighting some of the features peculiar to each scheme. The performance of the filters is compared on the table reported at the end of this subsection.

The local coordinate estimator

In figures 5-6 we show the six components of translational and rotational velocity as estimated by the local coordinates estimator. Ground truth is plotted in dotted lines. Convergence is reached in less than 20 steps from an initial condition within 20% off the true state. Initialization is performed using one step of the traditional Longuet-Higgins' algorithm [49]. Tuning of the filter has been performed, as with the other schemes, within an order of magnitude. It must be pointed out that we have observed a better behavior by increasing the variance of the pseudoinnovation. This is due to the fact that the EKF relies on the hypothesis that the measurement noise is white and the linearization error is negligible, while this is often not the case. An increase in the variance of the measurement noise accounts for the residual of the linearization. The computational cost of one iteration is of about 100 Kflops for 20 points.

The estimator in the embedding space

In fig. 9 we show the 9 components of the essential matrix as estimated by the essential estimator in the embedding space. Since convergence is about 4 times slower than the local coordinate version, but each step requires 4 times less computation, we have sampled the measurements four times faster, ending up with a 400 frames-long sequence.

Note first that, between the frames 200 and 300, the true value of the state is zero. However, the estimates of the filter drift off to non-zero values, since the essential matrices are defined as to have unit norm. Such non-zero values are those that allow estimating correctly the rotational velocity and a dummy direction of translation even in the case of pure rotation about the optical axis, as discussed in section 3.5. Once transformed the state into local coordinates and inserted the appropriate scale, it is possible to recover the correct rotational and translational components of motion, as shown in figures 7-8.

The homeomorphism Φ defined in (9) may have singularities due to noise when the last eigenspace is exchanged with one of the other two. In fact, due to the presence of noise, the third singular value of the estimated essential matrix is non-zero, and occasionally may even become bigger than the other two. Since the SVD sorts the singular values in decreasing order, the eigenvectors – which encode the motion information – may be interchanged.

This causes the spikes observed in the estimates of motion. However, there is no transient to recover, since the errors do not occur in the estimation step, but only in transferring to local coordinates. The switching can be avoided by a higher level control on the continuity of the singular values. The only significant error in the local coordinates occurs at around frame 260, when the translation is zero and the direction of rotation is inverted. The smoothness imposed by the dynamics of the parameters is responsible for the transient in the estimates of the rotation, which propagates onto the estimates of translation, causing a visible spike with a significant transient. Note that a much less relevant spike was also present in the estimate of the filter in local coordinates (figure 5).

The computational cost of our current implementation of the filter in the embedding space amounts to circa 41 Kflops per each step for 20 points. Initialization was performed within 20%, as in the previous case, using one step of the algorithm of Longuet-Higgins [49].

The 2-D iteration

The iteration of the essential filter in local coordinates has been implemented using the double iteration described in appendix A. The results are reported in figures 10-11. This scheme reaches similar accuracy to the local filter after proper initialization, even though the error analysis used for calculating the variance of the estimates at each fixed time was only approximate. Speed may be adjusted by varying the number of iterations at each fixed time. We have noticed that a number of steps between 3 and 7 is sufficient. The cost of the scheme for 7 iterations and 20 points is 100 Kflops. The simulations reported were performed using a constant variance of the error of the k-iteration.

We now summarize the performance of the three schemes: mean (M) and standard deviation (Std) of the estimation error are computed in steady state between time 30 and 50 for the local coordinate scheme and the 2-D iteration, while between time 180 and 260 for the estimator in the embedding space.

Scheme	T_X	T_Y	T_Z
Local	M: .0002 Std:.0004	M:-.0015 Std: .0048	M: .0002 Std: .0004
Essential	M: 3.9754E-5 Std: .0001	M: .0017 Std: .0013	M: .0002 Std: .0001
2-D	M: .376E-3 Std: .0009	M: -.0835E-3 Std: .0071	M: .2851E-3 Std: .0009

Scheme	Ω_X	Ω_Y	Ω_Z
Local	M:.0008 Std: .0022	M:.0002 Std:.0002	M:-.0002 Std:.0008
Essential	M:-.0008 Std: .0004	M: 3.9949E-6 Std: .0002	M: -1.6107E-5 Std: .0004
2-D	M: .2156E-3 Std: .0034	M: .2261E-3 Std: .0006	M:.0073E-3 Std:.0006

6.2 Experiments on real image sequences

In a first experiment, we have tested our schemes on a sequence of 10 images taken at the University of Massachussets at Amherst (see fig 12). There are 22 feature points visible; ground truth and feature tracking have been provided. Due to the limited length of the sequence, we have run it on the local coordinates estimator which, however, has a transient of about 10-20 steps to converge from arbitrary initial condition. Hence we have run the local estimator on the 10 images starting from zero initial condition, and we have used the final estimate as initial condition for a new run, whose results we report in figures 13-15. We did not perform any ad hoc tuning, and the setting was the same used in the simulations described in the previous paragraphs. In fig. 13 we report the 6 motion components as estimated by the local coordinate estimator and the corresponding ground truth (in dotted lines). The estimation error is plotted in figure 14. As it can be seen the estimates are within 5% error, and the final estimate is less than 1% off the true motion. Finally, in fig. 15 we display the norm of the pseudo-innovation of the filter, which converges to a value of about 10^{-3} in less than 10 + 5 steps. In this experiment, we have used the true norm of translation as the scale factor.

In a second experiment we have taken a box, attached some texture to it and generated a sequence of images by rotating the box on top of a revolving chair placed in front of the camera.

We have selected and tracked automatically feature-points using a multiscale implementation of a standard scheme proposed by Lucas and Kanade [51], which gave us a number of good features as well as a number of spurious one (like the "T"-junctions between the chair and the horizontal lines in the background wall) and points in the background (figure 16 top). A simple on-line statistical analysis on the innovation process of the filter allows us to easily reject these points as outliers [64]. The motion components of the remaining points, the ones attached to either the box or the chair, are estimated and plotted in figure 16 (bottom-left; error-bars indicate twice the variance of the estimates) along with the top-view of the structure as estimated by a simple EKF using the estimated motion, in the lines of [66] (figure 16, bottom-right).

Benchmark experiments on the scheme's performance on real-life situations have been conducted also by other researchers. For instance, Bouguet [7] considered a sequence of over 4000 frames taken from a camera mounted on a cart which moved inside a building, eventually returning to its initial position. The aim of the experiment was to assess how accurate the reconstruction of the trajectory was by integrating over time the velocity (or relative instantaneous configuration) through the sequence. Individual features, tracked using standard feature tracking schemes [4], have a relatively short lifetime (in the order of ten frames for that particular experiment). Therefore, traditional motion estimation schemes, having structure encoded in the state of the filter, would have discontinuities in the states corresponding to the structure parameters whenever a feature appears or disappears, which affects the estimates of motion through the coupling present in the model (14). On the contrary, the velocity of the cart is of course continuous, and only by decoupling motion from the estimation of structure it is possible to integrate information throughout the sequence without having to deal with a variable number of discontinuous states.

7 Conclusions

The problem of estimating three-dimensional motion from a sequence of images can be naturally set in the framework of dynamic estimation and identification. Under the assumption of a static scene, the rigid motion constraint and the perspective projection map *define* in a natural way a nonlinear dynamical model, and estimating motion is equivalent to a mixed estimation/identification task.

Motivated by the structural limitations of the natural model (see [59]), we have proposed a new formulation for structure-independent motion estimation, based upon the representation of motion via the "essential matrices", introduced by Longuet-Higgins [49]. Motion estimation is equivalent to the identification of a nonlinear implicit model with parameters on the essential manifold. Other problems in computer vision may be cast as the identification of a nonlinear implicit model, as for example dynamic self-calibration, subspace motion factorization, partial motion reconstruction from weak perspective.

We have proposed an algorithm which solves the identification task by estimating the state of a model defined on the parameter manifold. We perform the estimation either in the local coordinates or in the embedding space of the parameter manifold.

We are now in the process of implementing the proposed schemes on real-time hardware. We believe that the simplicity and robustness of the methods proposed, along with the power of modern architectures, will soon allow us to insert them into the control loop of mechanical systems. The flexibility of vision as a sensor, once brought to real-time operation, opens up a number of applications ranging from visually-guided navigation, manipulation, surveillance, active sensing and recognition.

Acknowledgements

We wish to thank Prof. Giorgio Picci for his advice, Prof. Karl Åström for his discussions on implicit Kalman filtering, Prof. Richard Murray and Prof. Shankar Sastry for their observations and useful suggestions. Finally, we thank John Oliensis and J. Inigo-Thomas for providing the rocket sequence.

References

[1] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, tensor analysis and applications.* Springer Verlag, 2 edition, 1988.

[2] E. Arbogast and R. Mohr. An ego-motion algorithm based on the tracking of arbitrary curves. *Proc. of the 2nd Europ. Conf. on Computer Vision*, 1992.

[3] A. Azarbayejani, B. Horowitz, and A. Pentland. Recursive estimation of structure and motion using relative orientation constraints. *Proc. CVPR*, New York, 1993.

[4] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. RPL-TR 9107, Queen's University Kingston, Ontario, Robotics and perception laboratory, 1992. Also in Proc. CVPR 1992, pp 236-242.

[5] A. Blake, M. Taylor, and A. Cox. Grasping visual symmetry. *Proc. of the ICCV*, 1993.

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

[7] J.-Y. Bouguet and P. Perona. A visual odometer and gyroscope. *Proceedings of the IEEE Int. Conf. Comp. Vision*, 1995.

[8] 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.

[9] T. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE trans. PAMI*, Jan. 1986.

[10] 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.

[11] R.S. Bucy. Non-linear filtering theory. *IEEE Trans. A.C. AC-10, 198*, 1965.

[12] F. Chaumette and A. Santos. Tracking a moving object by visual servoing. *Proc. of the 12th IFAC World Congr. Vol. 9 pp. 409–414*, 1993.

[13] R. Cipolla and A. Blake. Surface orientation and time to crash from image divergence and deformation. *Proc. of the European Conf. on Comp. Vision*, 1992.

[14] R. Curwen, A. Blake, and A. Zisserman. Real-time visual tracking for surveillance and path planning. *Proc. of the ECCV*, 1992.

[15] F. Darmon. A recursive method to apply the hough transform to a set of moving objects. *Proc. IEEE, CH 1746 7/82*, 1982.

[16] W. Dayawansa, B. Ghosh, C. Martin, and X. Wang. A necessary and sufficient condition for the perspective observability problem. *Systems and Control Letters, in press*, 1994.

[17] E. Di-Bernardo, L. Toniutti, R. Frezza, and G. Picci. Stima del moto dell'osservatore e della struttura della scena mediante visione monoculare. *Tesi di Laurea–Università di Padova*, 1993.

[18] E. D. Dickmanns and Th. Christians. Relative 3d-state estimation for autonomous visual guidance of road vehicles. In *Intelligent autonomous system 2 (IAS-2)*, Amsterdam, 11-14 December 1989.

[19] E. D. Dickmanns and V. Graefe. Applications of dynamic monocular machine vision. *Machine Vision and Applications*, 1:241–261, 1988.

[20] E. D. Dickmanns and V. Graefe. Dynamic monocular machine vision. *Machine Vision and Applications*, 1:223–240, 1988.

[21] K. Hashimoto T. Kimoto T. Ebine and H. Kimura. Image-based dynamic visual servo for a hand-eye manipulator. *Kodama Kimura, editor,* Recent advances in mathematical theory of systems, control, networks, and signal processing II*, pages 609–614. Proceedings of the international symposium of MTNS, Mita Press*, 1991.

[22] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. *Proc. of the 2 ECCV*, 1992.

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

[24] O. D. Faugeras, N. Ayache and B. Faverjon. Building visual maps by combining noisy stereo images. *Proc. of the Int. Conf. on Robotics and Automation*, 1986.

[25] O. D. Faugeras, Q.T. Luong, and S. J. Maybank. Camera self-calibration: theory and experiments. *Proc. of the ECCV92, Vol. 588 of LNCS, Springer Verlag*, 1992.

[26] O. D. Faugeras, F. Lustman, and G. Toscani. Motion and structure from motion from point and line matches. *Proc. of the IEEE Conf. ICCV*, 1987.

[27] O. D. Faugeras and S. J. Maybank. Motion from point matches: multiplicity of solutions. *Int. J. of Computer Vision*, 1990.

[28] D.B. Gennery. Tracking known 3-dimensional object. In *Proc. AAAI 2nd Natl. Conf. Artif. Intell.*, pages 13–17, Pittsburg, PA, 1982.

[29] D.B. Gennery. Visual tracking of known 3-dimensional object. *Int. J. of Computer Vision*, 7(3):243–270, 1992.

[30] B. Ghosh, M. Jankovic, and Y. Wu. Perspective problems in systems theory and its application in machine vision. *Journal of Math. Systems, Est. and Control*, 1994.

[31] E.J. Gibson, J. J. Gibson, O. W. Smith, and H. Flock. Motion parallax as a determinant of perceived depth. *J. Exp. Psych. Vol.45*, 1959.

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

[33] R. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Proc. 2^{nd} Europ. Conf. Comput. Vision, G. Sandini (Ed.), LNCS-Series Vol. 588, Springer-Verlag*, 1992.

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

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

[36] J. Heel. Temporal integration of 3-d surface reconstruction. *To appear on IEEE trans. PAMI, special issue on the interpretation of 3-D scenes.*, March 1991.

[37] C.C. Ho and N.H. McClamrock. Autonomous spacecraft docking using a computer vision syste¡m. *Proc. of the 31st CDC – Tucson, AZ*, 1992.

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

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

[40] R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. of the ASME-Journal of basic engineering.*, 35-45, 1960.

[41] A. Karger and J. Novak. *Space kinematics and Lie groups*. Gordon and Brech Science Publ., 1985.

[42] K. Hashimoto T. Kimoto T. Ebine H. Kimura. Manipulator control with image-based visual servo. *IEEE intl' Conference on Robotics and Automation, pages 2267–2272*, 1991.

[43] J. J. Koenderink and A. J. Van Doorn. Affine structure from motion. *J. Optic. Soc. Am.*, 1991.

[44] M. Lei and B. K. Ghosh. A new nonlinear feedback controller for visually-guided robotic motion tracking. *Proc. of the ECC*, 1993.

[45] P. Libermann and C. M. Marle. *Symplectic geometry and analytical mechanics*. Reidel, Dordrecht, 1987.

[46] Y. Liu and O.D. Faugeras T.S. Huang. Determination of camera location from 2d to 3d line and point correspondences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):28–37, 1990.

[47] L. Ljung. *Theory and Practice of Recursive Identification.* MIT press, 1983.

[48] L. Ljung. *System Identification: theory for the user.* Prentice Hall, 1987.

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

[50] H.C. Longuet-Higgins. Configurations that defeat the eight-point algorithm. *Mental processes: studies in cognitive science, MIT press*, 1987.

[51] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. 7th Int. Joint Conf. on Art. Intell.*, 1981.

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

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

[54] R. Mohr. Relative positioning with uncalibrated cameras. *in Geometric Invariance in Computer Vision*, 1992.

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

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

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

[58] A. Shashua. On geometric and algebraic aspects of 3d affine and projective structure from perspective 2d views. *AI Memo 1405, MIT AI Lab*, March 1993.

[59] S. Soatto. Observability/identifiability of rigid motion under perspective projection. In *33 IEEE Conf. on Decision and Control*, pages 3235–3240, Dec. 1994. Extended version submitted to the IFAC Journal "Automatica".

[60] S. Soatto, R. Frezza, and P. Perona. Motion estimation on the essential manifold. In *Proc. 3^{rd} Europ. Conf. Comput. Vision, J.-O. Eklundh (Ed.), LNCS-Series Vol. 800-801, Springer-Verlag*, pages II–61–72, Stockholm, May 1994.

[61] S. Soatto and P. Perona. Recursive estimation of camera motion from uncalibrated image sequences. In *Proc. 1^{st} IEEE Int. Conf. on Image Processing*, pages III–58–62, Austin – TX, November 1994. Extended version in: Technical Report CIT-CDS 94-003, California Institute of Technology.

[62] S. Soatto and P. Perona. Dynamic 3-D rigid motion estimation from weak perspective. In *Proc. of the IEEE Int. Conference on Computer Vision ICCV, Boston – MA*. To appear, June 1995.

[63] S. Soatto and P. Perona. Structure-independent visual motion control on the essential manifold. In *Proc. of the IFAC Symposium on Robot Control (SYROCO), Capri, Italy*, pages 869–876, Sept. 1994.

[64] S. Soatto and P. Perona. Three dimensional transparent structure segmentation and multiple 3d motion estimation from monocular perspective image sequences. In *IEEE Workshop on Motion of Nonrigid and Articulated Objects, Austin – TX*, pages 228–235. IEEE Computer Society, November 1994.

[65] S. Soatto and P. Perona. Visual motion estimation from subspace constraints. In *Proc. 1st IEEE Int. Conf. on Image Processing*, pages I–333–337, Austin, November 1994. Extended version in: Technical Report CIT-CDS 94-005, California Institute of Technology, submitted to the Int. Journal of Computer Vision.

[66] 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.

[67] T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.

[68] M. Spivak. *A comprehensive introduction to differential geometry– Voll.I-V*. Publish or perish, 1970-75.

[69] M. Swain and M. Stricker (editors). Promising directions in active vision. Technical Report T.R. CS 91-27, University of Chicago, November 1991. Written by the attendees of the NSF Active Vision Workshop – August 5-7 1991.

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

[71] G. Taubin. Estimation of planar curves, surfaces and nonplanar space curves defined by implicit equations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1991.

[72] C. J. Taylor and D. J. Kriegman. Structure and motion from line segments in multiple views. *Technical Report 9402, Yale University*, 1994.

[73] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method – 3. detection and tracking of point features. CMU-CS 91-132, School of CS – CMU, April 1991.

[74] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method – 2. point features in 3d motion. CMU-CS 91-105, School of CS – CMU, January 1991.

[75] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method – 1. planar motion. CMU-CS 90-166, School of CS – CMU, September 1990.

[76] V. S. Varadarajan. *Lie groups, Lie algebras and their representation.* Springer Verlag, GTE, 1984.

[77] H. von Helmholtz. *Treatise on Physiological Optics.* 1910.

[78] J. Weng, T. Huang, and N. Ahuja. Motion and structure from two perspective views: algorithms, error analysis and error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(5):451–476, 1989.

[79] J. Weng, T.S. Huang, and N. Ahuja. Motion and structure from line correspondences: closed-form solution, uniqueness and optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(3):318–336, 1992.

[80] Z. Zhang and O. Faugeras. *3D dynamic scene analysis*, volume 27 of *Information Sciences*. Springer-Verlag, 1992.

[81] Z. Zhang and O. D. Faugeras. Estimation of displacement from two 3d frames obtained form stereo. *TR 1440 – INRIA*, 1991.

A Recursive local identification of implicit systems using prediction error criteria

Suppose $\{x(t)\} \in \mathbb{R}^N$ is a trajectory on a linear state-space, which is subject to an implicit dynamic constraint of the form

$$h\left[x(t), dx(t), a\right] = 0 \qquad x(0) = x_0 \qquad a \in M \tag{38}$$

where a are some unknown parameters which may move (slowly) on some topological manifold M. Call $\alpha \doteq \psi(a) \in \mathbb{R}^m$ the local coordinates correspondent of a. Suppose we are able to measure x up to some white, zero-mean Gaussian noise:

$$y(t) = x(t) + n(t) \qquad\qquad n \in \mathcal{N}(0, R_n).$$

We are interested in identifying the parameters a recursively from the measurements $\{y(t)\}$ based on the minimization of some cost function of the prediction error (for a classical treatment of prediction error methods (PEM) for linear explicit models see for example [67, 48, 47]).

A common paradigm for PEM identification consists in forcing a Kalman Filter to work as a parameter estimator. The state of the filter is augmented with the unknown parameters, which are described using a random walk model. In this section we will extend this paradigm to nonlinear implicit dynamics and parameters living on a topological manifold. We will restrict our attention to discrete time dynamics, although the same analysis may be carried out for continuous time models.

First we proceed in analogy with the linear-explicit case: we describe the local coordinates of the parameters as first order random walk, and use the dynamic constraint as an implicit measurement constraint:

$$\begin{cases} \alpha(t+1) = \alpha(t) + n_\alpha(t) & \alpha(0) = \alpha_0 \\ h\left[y(t) - n(t), y(t-1) - n(t-1), \psi^{-1}(\alpha(t))\right] = 0 \end{cases} \tag{39}$$

where we have substituted the index t with $t-1$ in the measurements $\{y\}$ (or equivalently the estimator runs with one step delay). We assume n_α, the noise driving the random walk, to be white zero-mean and Gaussian; its variance R_α may be regarded as a tuning parameter. The noise process $\{n(t)\}$ induces a residual in the measurement equation: if we approximate $x(t)$ with $y(t)$, in general we will observe $h\left[y(t), y(t-1), a\right] = \tilde{n} \neq 0$, where $\tilde{n}$ depends on $\{n\}\{y\}$ and a. This residual – as we will see – is the prediction error (or pseudo-innovation) when choosing a least-squares criterion in the PEM.

Let us collect the measurements into a vector $\bar{y}(t) \doteq \left[\begin{array}{cc} y^T(t) & y^T(t-1) \end{array}\right]^T$, and similarly with $\bar{n}(t) \doteq [n^T(t)\ n^T(t-1)]^T$. Our task is to estimate α from the model

$$\begin{cases} \alpha(t+1) = \alpha(t) + n_\alpha(t) & \alpha(0) = \alpha_0 \\ h\left[\bar{y}(t) - \bar{n}(t), \psi^{-1}(\alpha(t))\right] = 0. \end{cases} \tag{40}$$

In order to follow the course of the linear-explicit case, we have to solve a number of problems:

1. the noise $\bar{n}$ is not white: $E[\bar{n}(t)\bar{n}^T(s)] = \left[\begin{array}{cc} R_n\delta(t-s) & R_n\delta(t-s+1) \\ R_n\delta(t-s-1) & R_n\delta(t-s) \end{array}\right]$

2. the error $\bar{n}$ does not appear additively in the measurement equation

3. the measurement equation is nonlinear and implicit.

The Extended Kalman Filter (EKF) [40, 11, 39] is a general-purpose local extension to nonlinear systems of the traditional Kalman Filter. It is based on a variational model about the best current trajectory. The systems is linearized at each step around the current estimate in order to calculate a correcting gain; the update of the previous estimate is then performed on the original (nonlinear) equations. In order to solve step 3 we need to further extend the EKF to cope with the implicit measurement constraint. This is done in appendix B. We call the result Implicit Extended Kalman Filter (IEKF); some variations of the scheme have been used in different applications in the last years, see for example [15, 17, 36, 26]. The derivation is based on the simple fact that the variational model about the current trajectory is *linear and explicit*, so that the a pseudo-innovation process may be defined analogously to the explicit case.

The derivation of the IEKF in appendix B does not address the fact that the noise $\bar{n}$ is correlated (see point 2. above). The residual of the measurement equation $\tilde{n}$, which is in fact the pseudo-innovation of the filter, is characterized in terms of $\bar{n}$, provided that the last is white, zero-mean and uncorrelated with n_α. In the following section we will show how to whiten $\bar{n}$ and therefore reduce the problem to a form suitable for using the IEKF as derived in appendix B. Later on we will see how the problem simplifies by assuming that $\bar{n}$ is white.

A.1 Uncorrelating the model from the measurements

Consider a first order expansion of the measurement equation about the point $\bar{y}(t), \alpha(t)$:

$$h\left[\bar{y}(t), \psi^{-1}(\alpha(t))\right] - D_+(t)n(t) - D_-(t)n(t-1) = \mathcal{O}(\|\bar{n}\|^2) \cong 0$$

where the limit implicit in $\mathcal{O}$ is intended in the mean-square sense, and where we have defined

$$D_+(t) \doteq \left(\frac{\partial h[x(t), x(t-1), a]}{\partial x(t)}\right)_{|\bar{y}(t), \psi^{-1}(\alpha(t))} \tag{41}$$

$$D_-(t) \doteq \left(\frac{\partial h[x(t), x(t-1), a]}{\partial x(t-1)}\right)_{|\bar{y}(t), \psi^{-1}(\alpha(t))}. \tag{42}$$

Here the residual $\tilde{n}(t) = -D_+(t)n(t) - D_-(t)n(t-1)$ is clearly correlated. In order to estimate the dynamics of $n(t)$, we may insert it into the state: call $z(t) \doteq n(t-1)$.

$$\begin{cases} \alpha(t+1) = \alpha(t) + n_\alpha(t) & \alpha(0) = \alpha_0 \\ z(t+1) = n(t) & z(0) = 0 \\ 0 = h\left[\bar{y}(t), \psi^{-1}(\alpha(t))\right] - D_-(t)z(t) + w(t) \end{cases} \tag{43}$$

where we have defined $w(t) \doteq -D_+(t)n(t)$. Now the measurement error w is white; however, it is correlated with the model error $v \doteq [n_\alpha^T, n^T]^T$. We may therefore project the model error onto the span of the measurement error, $H(w)$, in order to make the two orthogonal. We define $\tilde{v}(t) \doteq v(t) - \hat{E}[v(t)|H(w)]$. Since $w(t), n(t)$ and $n_\alpha(t)$ are white, it is easily seen that $\hat{E}[v(t)|H(w)] = \hat{E}[v(t)|w(t)] = E[v(t)w^T(t)]\left(E[w(t)w^T(t)]\right)^{-1} w(t) \doteq \Sigma_{vw}\Sigma_w^{-1}w(t)$. If we define

$$Q(t) \doteq \begin{bmatrix} R_\alpha & 0 \\ 0 & R_n \end{bmatrix} \tag{44}$$

$$R(t) \doteq D_+(t)R_n(t)D_+^T(t) \tag{45}$$

$$S(t) \doteq \begin{bmatrix} 0 \\ -R_n(t)D_+^T(t) \end{bmatrix} \tag{46}$$

it is easy to see that $\Sigma_{vw}\Sigma_w^{-1} = S(t)R^{-1}(t)$; furthermore $\Sigma_{\tilde{v}} \doteq \tilde{Q}(t) = Q(t) + S(t)R^{-1}(t)S^T(t)$. Now $\tilde{v} \doteq v - SR^{-1}w$ is by construction orthogonal (uncorrelated) to w.

A.2 A model for PEM identification of nonlinear implicit models

In the previous paragraph we have derived an extended model (up to first order) with the model error uncorrelated from the measurement error:

$$\begin{cases} \alpha(t+1) = \alpha(t) + n_\alpha(t) & \alpha(0) = \alpha_0 \\ z(t+1) = K(t)\left(h\left[\bar{y}(t), \psi^{-1}(\alpha(t))\right] - D_-(t)z(t)\right) + n(t) & z(0) = 0 \\ 0 = h\left[\bar{y}(t), \psi^{-1}(\alpha(t))\right] - D_-(t)z(t) + w(t) \end{cases} \tag{47}$$

where we have defined

$$K(t) \doteq R_n(t)D_+^T(t)\left(D_+(t)R_n(t)D_+^T(t)\right)^{-1} \tag{48}$$

$$w(t) \doteq -D_+(t)n(t). \tag{49}$$

By applying the results of appendix B, we can derive a pseudo-optimal PEM identification scheme described by the following iteration:

Prediction step

$$\begin{cases} \hat{\alpha}(t+1|t) = \hat{\alpha}(t|t) & \hat{\alpha}(0|0) = \alpha_0 \\ \hat{z}(t+1|t) = K(t)\left(h\left[\bar{y}(t), \hat{\alpha}(t|t)\right] - D_-(t)\hat{z}(t|t)\right) & \hat{z}(0|0) = 0 \\ P(t+1|t) = F(t)P(t|t)F^T(t|t) + \tilde{Q}(t) & P(0|0) = P_0 \end{cases} \tag{50}$$

where $F \doteq \begin{bmatrix} I & 0 \\ K(t)\left([C(t) \right. & \left. - D_-(t)]\right) \end{bmatrix}$ and $C(t) \doteq \left(\frac{\partial h[\bar{y}, \psi^{-1}(\alpha)]}{\partial \alpha}\right)_{|\hat{\alpha}(t|t), \bar{y}(t)}$.

Update step

$$\begin{cases} \begin{bmatrix} \hat{\alpha}(t+1|t+1) \\ \hat{z}(t+1|t) \end{bmatrix} = \begin{bmatrix} \hat{\alpha}(t+1|t) \\ z(t+1|t+1) \end{bmatrix} + L(t+1)\left(h\left[\bar{y}(t), \hat{\alpha}(t+1|t)\right] - D_-(t+1)\hat{z}(t+1|t)\right) \\ P(t+1|t+1) = \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)D_+(t+1)R_n(t+1)D_+^T(t+1)L^T(t+1) \end{cases} \tag{51}$$

where

$$L(t+1) \doteq P(t+1|t)C^T(t+1)\Lambda^{-1}(t+1) \tag{52}$$

$$\Lambda(t+1) \doteq C(t+1)P(t+1|t)C^T(t+1) + D_+(t+1)R_n(t+1)D_+^T(t+1) \tag{53}$$

$$\Gamma(t+1) \doteq I - L(t+1)C(t+1) \tag{54}$$

Note that we are trying to estimate a process $\{z(t)\}$ which is nearly white noise ($n(t)$ is correlated only within one step). Furthermore, if we expect a large number of measurements, the cost in updating a large state and tuning a large number of model-variance parameters may be relevant. In practical applications the approximation $\bar{n}$ as white noise are often better behaved. In the following section we show how the structure of the filter simplifies under such an approximation.

A.3 A simplified version: approximate least-squares PEM identification

In this section we report the equation of the parameter estimator which are obtained supposing that the residual $\tilde{n}$ is white. This correspond to applying the results of appendix B directly to the model of eq. (40), assuming that $\{\bar{n}\}$ is a white process:

Prediction step

$$\begin{cases} \hat{\alpha}(t+1|t) = \hat{\alpha}(t|t) & \hat{\alpha}(0|0) = \alpha_0 \\ P(t+1|t) = P(t|t) + R_\alpha(t) & P(0|0) = P_0 \end{cases} \tag{55}$$

Update step

$$\begin{cases} \hat{\alpha}(t+1|t+1) = \hat{\alpha}(t+1|t) + L(t+1)h\left[\bar{y}(t), \psi^{-1}(\hat{\alpha}(t+1|t)\right] \\ P(t+1|t+1) = \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)D_+(t+1)R_n(t+1)D_+^T(t+1)L^T(t+1) \end{cases} \tag{56}$$

where now the quantities $L(t+1), \Lambda(t+1)$ and $\Gamma(t+1)$ are defined according to appendix B. Note that we have reduced the size of the state from $n+m$ down to m.

Detecting outliers

Note that each component of the pseudo-innovation is a measure of the consistency of each datum with the current parameter estimates. This proves useful when applied to the motion problem because it allows us to detect outliers and also segment the scene into a number of independently moving objects [64].

A.4 An iterative scheme for computing the update

The IEKF update seen in the previous section may be substituted with a Gauss-Newton iteration, as it is customary in recursive ID of linear models:

$$\hat{\alpha}(k+1) = \hat{\alpha}(k) - L_{NR}(k)h(\hat{\alpha}(k))$$

where $L_{NR} = J_h^{-1}(\hat{\alpha}(k))$ and J_h is the jacobian of h.

Note that at each fixed time we could perform a Newton-Raphson iteration on the function $h(\bar{y}, \alpha)$, for which local convergence results are known as well as bounds on the convergence rate. This suggests, as an alternative to the IEKF, to fix t and perform a Newton-Raphson iteration along the k coordinate. Once this is done, we propagate the estimate across time with an iteration which now is *linear*, and has all the desirable asymptotic properties.

A.4.1 Iteration at each fixed time

At each time instant a new set of measurements $\bar{y}(t)$ becomes available. The constraint imposes

$$h[\bar{y}(t), \alpha] = 0 \quad \forall t$$

Define $T_\alpha h : \mathbb{R}^m \rightarrow \mathbb{R}^n$ to be the derivative of the map h and $J_h(\alpha)$ the Jacobian matrix calculated at the point α. Suppose that there exists some α^* such that $h(\bar{y}(t), \alpha^*) = 0$ for our particular (fixed) t. Then we may write a first order expansion around the point α^*, starting from some point α_0 (we neglect time indices for the remainder of this section); the resulting iteration, which is obtained by neglecting the second order term of the expansion, is defined by

$$h[\alpha_k] \doteq J_h(\alpha_k)[\alpha_{k+1} - \alpha_k].$$

At each iteration we solve for Y the linear problem

$$J_h(\alpha_k)Y = h[\alpha_k]$$

and then define $\alpha_{k+1} \doteq \alpha_k + Y$. In general, also due to noise, we can expect $h[\alpha_k] \notin Im(J_h(\alpha_k))$, so that we will be seeking for Y such that $J_h(\alpha_k)Y$ is the projection of $h[\alpha_k]$ onto the range space of $J_h(\alpha_k)$:

$$\alpha_{k+1} \doteq \alpha_k - L_{NR}(k)h[\alpha_k].$$

where $L_{NR}(k) \doteq \left(J_h^T(\alpha_k)J_h(\alpha_k)\right)^{-1} J_h^T(\alpha_k)$. The map defined by the right-hand side of the above equation is contractive as long as $J_h(\alpha_k)$ has full rank, in which case the scheme is guaranteed to converge to some (possibly local) minimum.

At each time the scheme will converge to some α^*, which best explains the noisy measurements $y_i(t), y_i(t-1)$; hence we have $\alpha^* = \alpha + n_\alpha$ where n_α is an error term, and can be interpreted as a white noise whose variance can be inferred from the variance of n and the linearization of the scheme about zero-noise. The estimate obtained at each fixed time, together with its variance, is fed to a time-integration step, which we describe next.

A.4.2 Propagation along time

Suppose at each fixed time the iteration along k described above converges to a fixed point $\alpha^*(t)$, then we may propagate the information across time with a similar iteration:

$$\hat{\alpha}(t+1) = \hat{\alpha}(t) + L(t)\left[\alpha^*(t) - \hat{\alpha}(t)\right]$$

which realizes a linear Kalman filter based upon the model

$$\begin{cases} \alpha(t+1) = \alpha(t) + n_\alpha(t) \\ \alpha^*(t) = \alpha(t) + n_0(t) \end{cases} \tag{57}$$

where n_α is the noise driving the random walk model for the parameters, which we assume to be white zero-mean and Gaussian, and n_0 is the error made by the fixed-time iteration. $L(t)$ is the usual linear Kalman gain [40, 39]. The above model has all the desirable properties, as it satisfies the conditions of the asymptotic theorem of Kalman Filtering.

Suppose now that the $k-$iteration has converged to a local minimum, which is compatible with the current observations. At the next step the $t-$iteration will predict an estimate which is in general no longer compatible with the current observations. This should help to disambiguate local minima as the measurements accumulate in time.

B Extended Kalman Filtering for implicit measurement constraints

We are interested in building an estimator for a process $\{\alpha\}$ which is described by a stochastic difference equation of the form

$$\alpha(t+1) = f(\alpha(t)) + v(t)\ ;\ \alpha(t_0) = \alpha_0$$

where $v(t) \in \mathcal{N}(0, Q_v)$ is a white, zero-mean Gaussian noise with variance Q_v. Suppose there is a measurable quantity $x(t)$ which is linked to α by the constraint

$$h[\alpha(t), x(t)] = 0\ \forall t. \tag{58}$$

We will assume throughout $f, h \in C^r\ ;\ r \geq 1$. Usually x is known via some noisy measurement:

$$x(t) = y(t) + w(t)\ :\ w(t) \in \mathcal{N}(0, R_w) \tag{59}$$

where the variance/covariance matrix R_w is derived from knowledge of the measurement device. The model we consider is hence of the form

$$\begin{cases} \alpha(t+1) = f(\alpha(t)) + v(t)\ ;\ \alpha(t_0) = \alpha_0 \\ h[\alpha(t), y(t) + w(t)] = 0 \end{cases}. \tag{60}$$

Construction of the variational model about the reference trajectory

Consider at each time sample t a reference trajectory $\bar{\alpha}(t)$ which solves the difference equation

$$\bar{\alpha}(t+1) = f(\bar{\alpha}(t))$$

and the jacobian matrix

$$F(\bar{\alpha}(t)) \doteq F(t) = \left(\frac{\partial f}{\partial \alpha}\right)_{|\bar{\alpha}(t)}.$$

The linearization of the measurement equation about the point $(\bar{\alpha}(t), y(t))$ is

$$h[\alpha(t), x(t)] = h[\bar{\alpha}(t), y(t)] + C(\bar{\alpha}, y)(\alpha(t) - \bar{\alpha}(t)) + D(\bar{\alpha}, y)(x(t) - y(t)) + \mathcal{O}(\mathcal{E}^2)$$

where

$$\begin{aligned} C(\bar{\alpha}, y) &\doteq \left(\frac{\partial h}{\partial \alpha}\right)_{|\bar{\alpha}(t), y(t)} \\ D(\bar{\alpha}, y) &\doteq \left(\frac{\partial h}{\partial x}\right)_{|\bar{\alpha}(t), y(t)} \\ \mathcal{E}^2 &\doteq \{\|\alpha - \bar{\alpha}\|^2, \|x - y\|^2\} \end{aligned}$$

and the limit implicit in $\mathcal{O}$ is intended in the mean-square sense. Exploiting the fact that $h[\alpha, x] = 0$, calling $\delta\alpha(t) \doteq \alpha(t) - \bar{\alpha}(t)$ and neglecting the arguments in C and D, we have, up to second order terms

$$h[\bar{\alpha}(t), y(t)] = -C\delta\alpha(t) - Dw(t).$$

Prediction Step

Suppose at some time t we have available the best estimate $\hat{\alpha}(t|t)$; we may write the variational model about the trajectory $\bar{\alpha}(t)$ defined such that

$$\bar{\alpha}(t+1) = f(\bar{\alpha}(t)) \ ; \ \bar{\alpha}(t) = \hat{\alpha}(t|t).$$

For small displacements we may write

$$\delta\alpha(t+1) = F(\bar{\alpha}(t))\delta\alpha(t) + \tilde{v}(t) \tag{61}$$

where the noise term $\tilde{v}(t)$ may include a linearization error component.

Note that with such a choice we have $\delta\hat{\alpha}(t|t) = 0$ and $\delta\hat{\alpha}(t+1|t) = F(\bar{\alpha}(t))\delta\hat{\alpha}(t|t) = 0$, from which we can conclude

$$\hat{\alpha}(t+1|t) = \bar{\alpha}(t+1) = f(\bar{\alpha}(t)) = f(\hat{\alpha}(t|t)). \tag{62}$$

The variance of the prediction error $\delta\hat{\alpha}(t+1|t)$ is

$$P(t+1|t) = F(t)P(t|t)F^T(t) + \tilde{Q} \tag{63}$$

where $\tilde{Q} = var(\tilde{v})$. The last two equations represent the prediction step for the estimator and are equal, as expected, to the prediction of the explicit EKF [40, 39, 11].

Update Step

At time $t+1$ a new measurement becomes available $y(t+1)$, which is used to update the prediction $\hat{\alpha}(t+1|t)$ and its error variance $P(t+1|t)$. Exploiting the linearization of the measurement equation about $\bar{\alpha}(t+1) = \hat{\alpha}(t+1|t)$, we obtain, letting $\hat{\alpha} \doteq \hat{\alpha}(t+1|t)$ and $y \doteq y(t+1)$,

$$h[\hat{\alpha}, y] = -C(\hat{\alpha}, y)\delta\alpha(t+1) - n(t+1) \tag{64}$$

where we have defined $n(t+1) \doteq D(\hat{\alpha}, y)w(t+1)$. This, together with the equation (61) defines a linear and *explicit* variational model, for which we can finally write the update equation based on the traditional linear Kalman filter:

$$\delta\hat{\alpha}(t+1|t+1) = \delta\hat{\alpha}(t+1|t) + L(t+1)\left[h[\hat{\alpha}, y] + C(\hat{\alpha}, y)\delta\hat{\alpha}(t+1|t)\right] \tag{65}$$

where

$$\begin{aligned} L(t+1) &= -P(t+1|t)C(\hat{\alpha}, y)^T \Lambda^{-1}(t+1) && (66)\\ \Lambda(t+1) &= C(\hat{\alpha}, y)P(t|t)C(\hat{\alpha}, y)^T + R_n(t+1) && (67)\\ P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)R_n(t+1)L(t+1)^T && (68)\\ \Gamma(t+1) &= (I - L(t+1)C(\hat{\alpha}, y)). && (69) \end{aligned}$$

Since $\delta\hat{\alpha}(t+1|t) = 0$ and $\delta\hat{\alpha}(t+1|t+1) = \hat{\alpha}(t+1|t+1) - \hat{\alpha}(t+1|t)$, we may write the update equation for the original model:

$$\hat{\alpha}(t+1|t+1) = \hat{\alpha}(t+1|t) + L(t+1)h\left[\hat{\alpha}(t+1|t), y(t+1)\right]. \tag{70}$$

In this formulation the quantity $h\left[\hat{\alpha}(t+1|t), y(t+1)\right]$ plays the role of the pseudo-innovation. The noise n defined in (64) has a variance which is calculated from its definition:

$$R_n(t) = D(\hat{\alpha}, y)R_w(t)D^T(\hat{\alpha}, y). \tag{71}$$

The update of the variance $P(t+1|t+1)$ is computed from the standard equations of the linear Kalman filter. The implicit Kalman filter was used by other researchers such as Darmon [15], Faugeras [26, 46, 80] and Heel [36], although in slightly different formulations, and always without a consistent derivation of the form of the update.

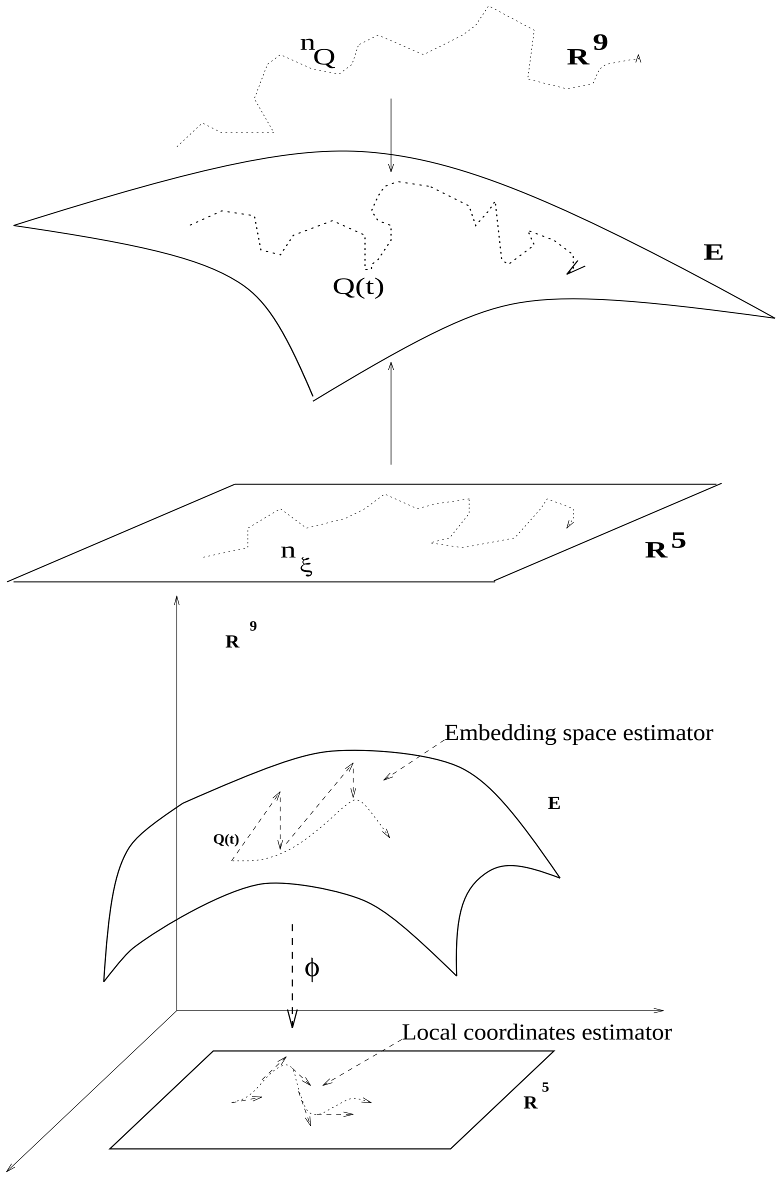

Figure 4: *(top) Model of motion as a random walk in* $\mathbb{R}^5$ *lifted to the manifold or as a random walk in* $\mathbb{R}^9$ *projected onto the manifold. (bottom) Estimation on the essential space.*

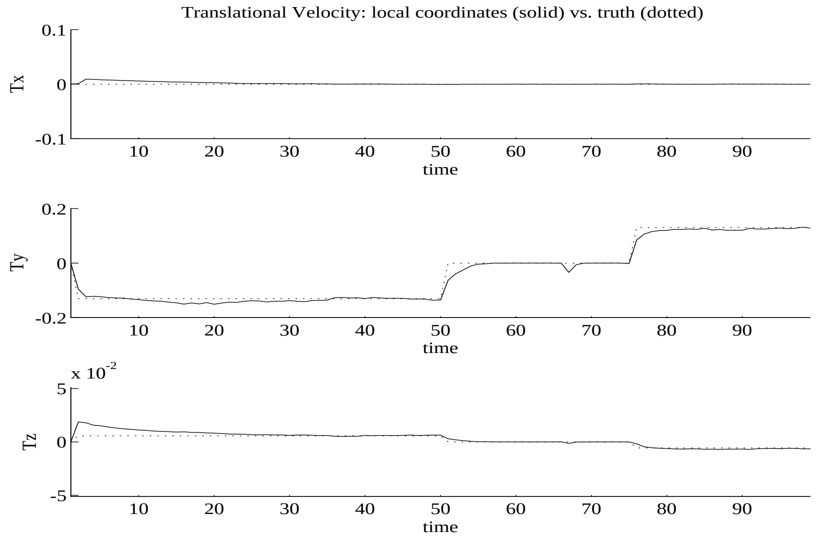

Figure 5: *Components of translational velocity as estimated by the local coordinate estimator. The ground truth is shown in dotted lines.*

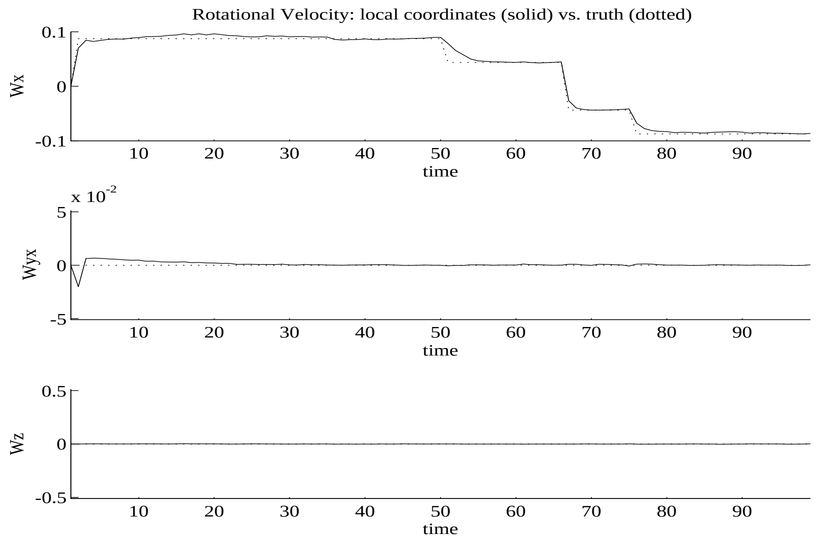

Figure 6: *Components of rotational velocity as estimated by the local coordinate estimator.*

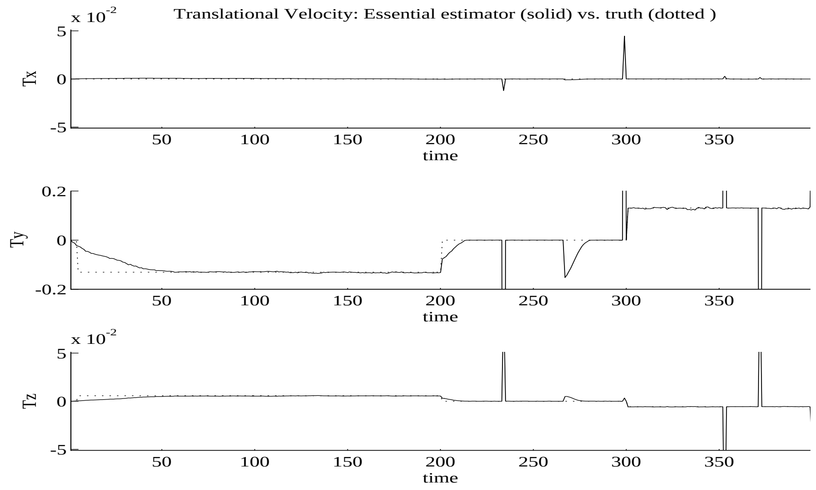

Figure 7: *Components of translational velocity as estimated by the essential estimator. Note the spikes due to the local coordinate transformation. Note also that such spikes do not affect convergence since they do not occur in the estimation process, but while transferring to local coordinates. The switching can be avoided by a higher level control on the continuity of the singular values of the estimated state. There is a significant error in the local coordinates at around frame 260, when the translation is zero and the direction of rotation is inverted. The smoothness imposed by the dynamics of the parameters is responsible for the transient in the estimates of the rotation, which propagates onto the estimate of translation, causing a visible spike with a significant transient.*

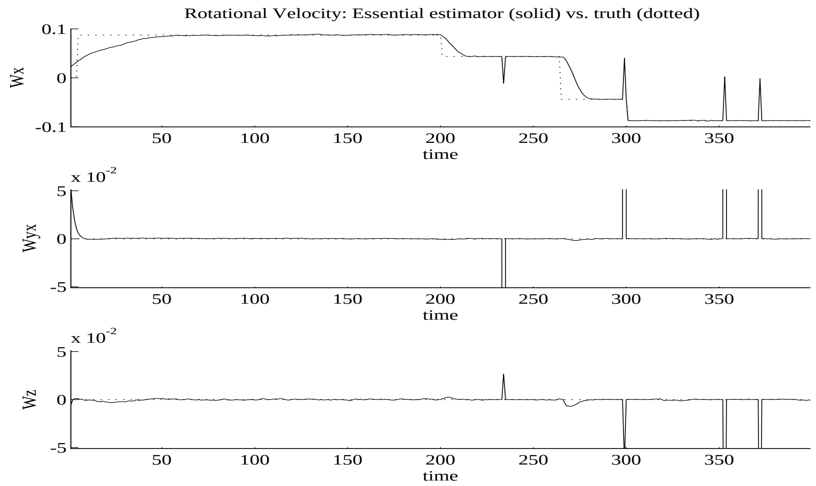

Figure 8: *Components of rotational velocity as estimated by the local coordinate estimator. The ground truth is shown in dotted lines. Note the spikes due to the local coordinate transformation. Note also that there is no transient to recover since they do not occur in the estimation process.*

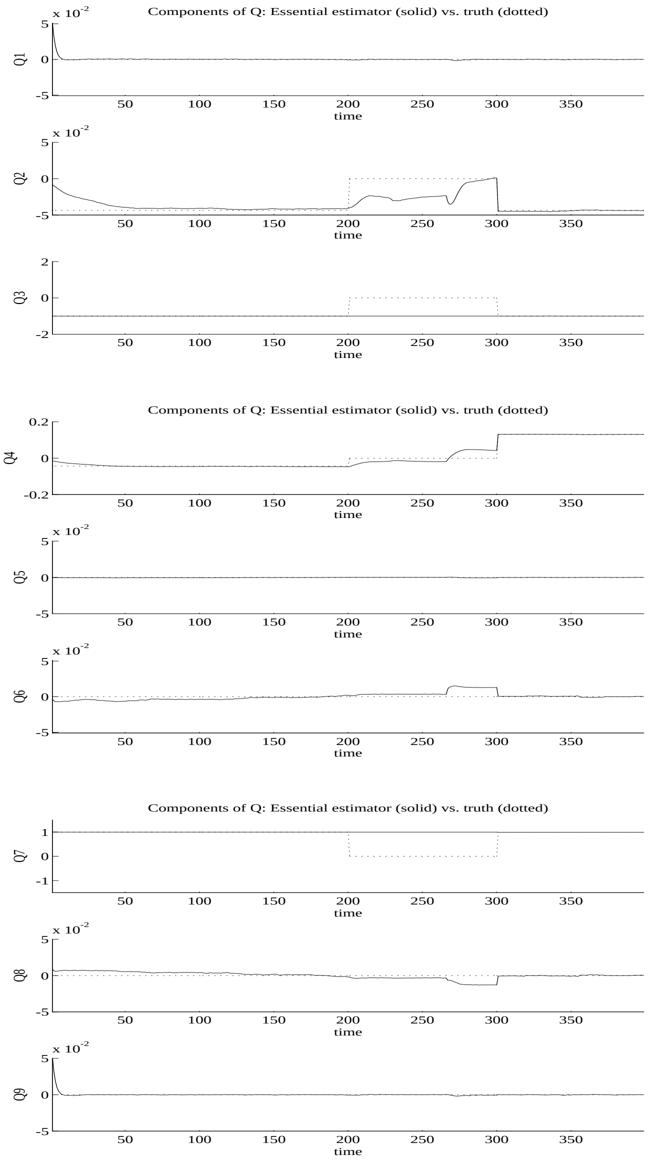

Figure 9: *Components of the essential matrix as estimated by the essential estimator. Note that there are no spikes. Note that the estimates between time 200 and 300 are non-zero, despite the ground truth (dotted line) is, since the essential space is normalized to unit-norm. The value of the components of the estimates of* $\mathbf{Q}$ *in the singular region* $T = 0$ *allow us to recover correctly the rotational velocity, once transformed to local coordinates.*

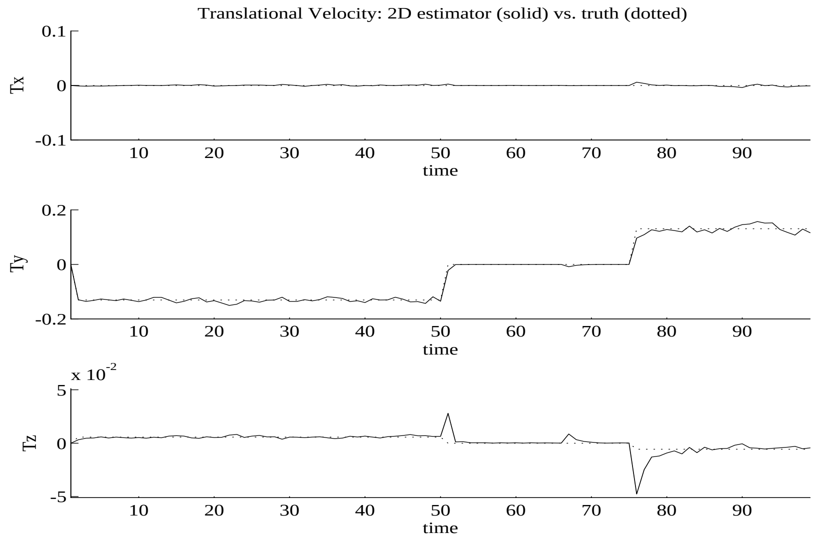

Figure 10: *Components of translational velocity as estimated by the double iteration estimator.*

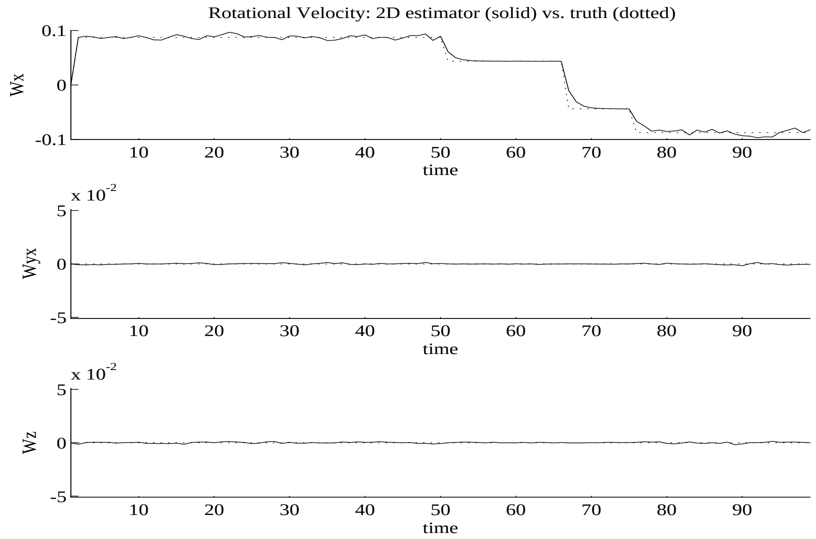

Figure 11: *Components of rotational velocity as estimated by the double iteration estimator.*

Figure 12: *One image of the rocket scene.*

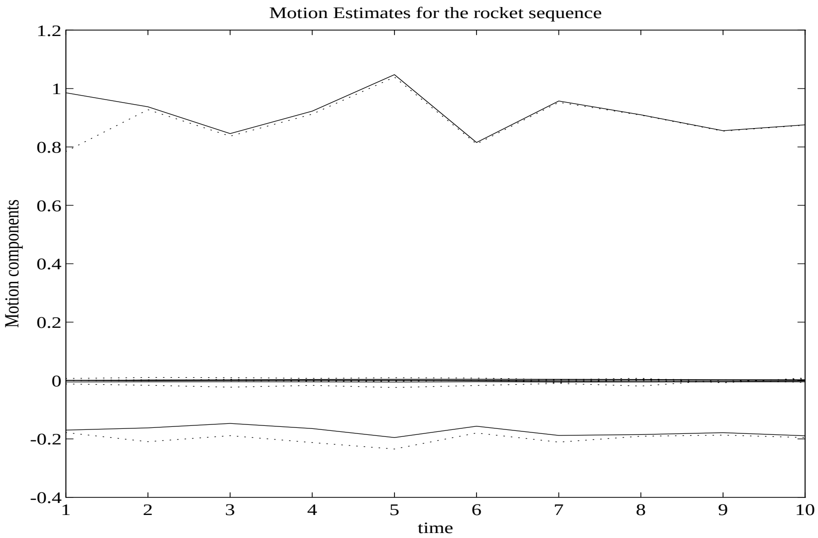

Figure 13: *Motion estimates for the rocket sequence: The six components of motion as estimated by the local coordinate estimator are showed in solid lines. The corresponding ground truth is in dotted lines.*

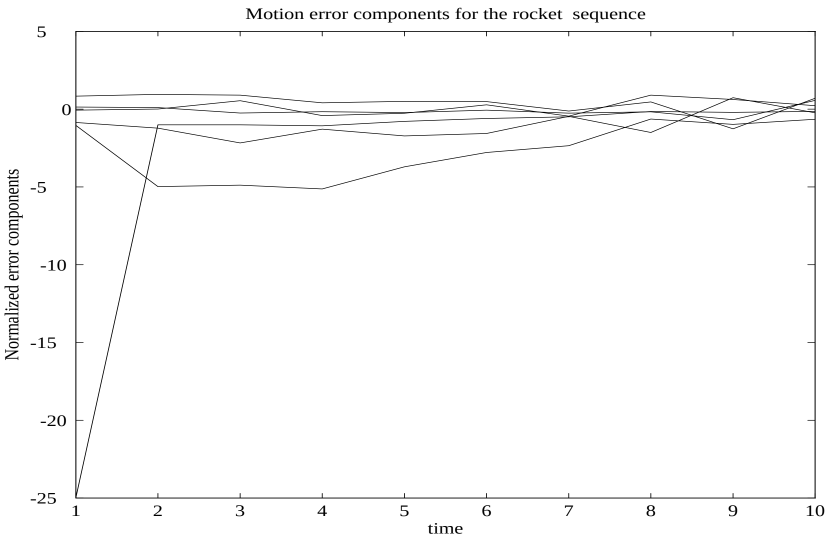

Figure 14: *Error in the motion estimates for the rocket sequence. All components are within 5% of the true motion.*

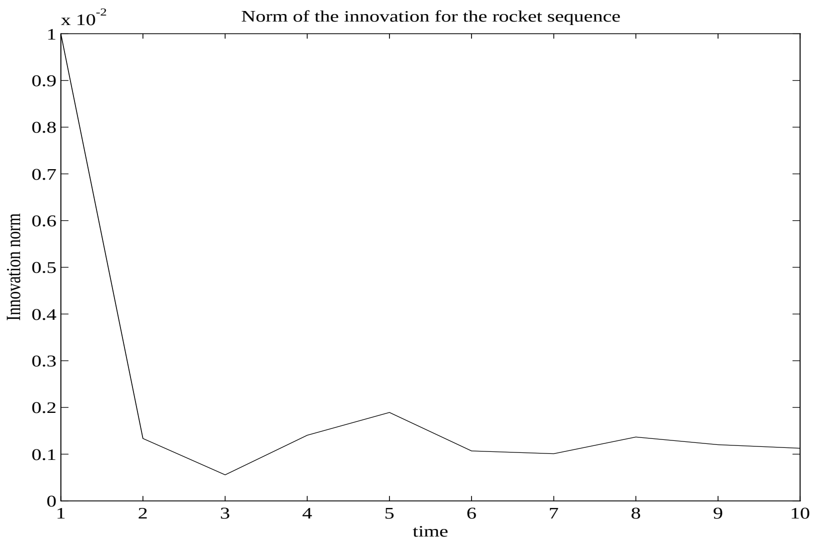

Figure 15: *Norm of the pseudo-innovation process of the local estimator for the rocket scene. Convergence is reached in less than 5 steps.*

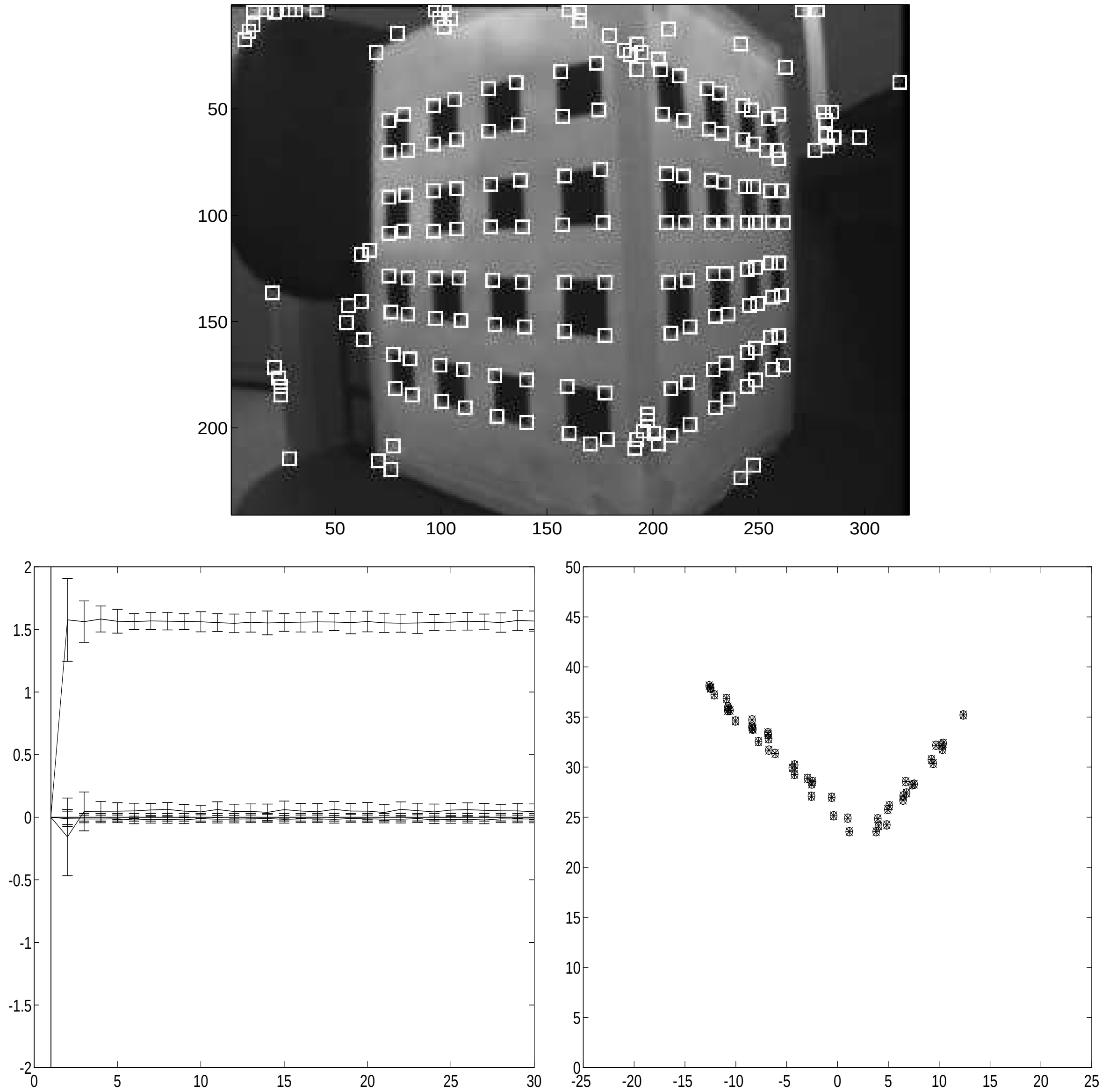

Figure 16: *(top) one frame of the original sequence with the feature points highlighted. (Lower-left) five components of the estimated motion (vertical units are rad/frame for the rotational velocity and rad/frame for the components of the direction of translation, the horizontal axis is the frame number). (Lower-right) reconstructed scene viewed from the top (the horizontal axis is a slice of the image plane, and the vertical axis is the depth of each feature point in cm).*