Recursive Motion Estimation on the Essential Manifold *

Stefano Soatto† Ruggero Frezza‡ Pietro Perona†‡
† California Institute of Technology 116-81, Pasadena–CA 91125
‡ Università di Padova, Dipartimento di Elettronica, Padova–Italy
soatto@caltech.edu

Technical Report **CIT-CNS 32/93** and **CIT-CDS 93-021**
California Institute of Technology
October 28, 1993 – revised November 7,1993

Abstract

Visual motion estimation can be regarded as estimation of the state of a system of difference equations with unknown inputs defined on a manifold. Such a system happens to be "linear", but it is defined on a space (the so called "Essential manifold") which is not a linear (vector) space.

In this paper we will introduce a novel perspective for viewing the motion estimation problem which results in three original schemes for solving it. The first consists in "flattening the space" and solving a nonlinear estimation problem on the flat (euclidean) space.
The second approach consists in viewing the system as embedded in a larger euclidean space (the smallest of the embedding spaces), and solving at each step a *linear estimation* problem on a *linear* space, followed by a "projection" on the manifold (see fig. 5).
A third "algebraic" formulation of motion estimation is inspired by the structure of the problem in local coordinates (flattened space), and consists in a double iteration for solving an "adaptive fixed-point" problem (see fig. 6).

Each one of these three schemes outputs motion estimates together with the joint second order statistics of the estimation error, which can be used by any structure from motion module which incorporates motion error [20, 23] in order to estimate 3D scene structure.

The original contribution of this paper involves both the problem formulation, which gives new insight into the differential geometric structure of visual motion estimation, and the ideas generating the three schemes. These are viewed within a unified framework. All the schemes have a strong theoretical motivation and exhibit accuracy, speed of convergence, real time operation and flexibility which are superior to other existing schemes [1, 20, 23].

Simulations are presented for real and synthetic image sequences to compare the three schemes against each other and highlight the peculiarities of each one.

*Research founded by the California Institute of Technology, an AT&T Foundation Special Purpose grant and grant ASI-RS-103 from the Italian Space Agency

1 Introduction

Consider a camera (or a human eye) moving inside a scene. The objects populating the ambient space are projected onto the CCD surface (or the retina), and their projection changes in time as the camera moves. The visual motion problem consists in reconstructing the motion of the camera and the "structure" of the scene from its projection. We will try here to formalize the problem to its essentials. Our "structure" consists in the position of a rigid set of feature points in 3D space with respect to some cartesian frame, for example the one moving with the observer. We call $\mathbf{X}^i = \begin{bmatrix} X & Y & Z \end{bmatrix}_i^T \in \mathbb{R}^3$ the coordinates of the i^{TH} point, and we let $i = 1 : N$. As the camera moves between two discrete time instants, with rotation R and translation T, the coordinates change according to the rigid motion constraint:

$$\mathbf{X}^i(t+1) = T(t) + R(t)\mathbf{X}^i(t) \quad \forall i = 1 : N \tag{1}$$

where $T \in \mathbb{R}^3$ and R belongs to the group of rotation matrices, which is called $SO(3)$ (Special Orthogonal group of transformations in $\mathbb{R}^3$). The rigid motion is hence represented by (R, T), which belongs to $SE(3)$, the Special Euclidean group of rigid motions in $\mathbb{R}^3$.For a detailed study of these groups the reader can refer to [18].

The camera (or eye) is represented by a map from the 3D space onto some 2D surface. We adopt for simplicity the ideal perspective projection model, and consider the camera as a map to the real projective space of dimension 2 [3]:

$$\begin{array}{rcl} \pi : \mathbb{R}^3 & \rightarrow & \mathbb{R}P^2 \\ \mathbf{X} & \mapsto & \pi(\mathbf{X}) \doteq \begin{bmatrix} x & y & 1 \end{bmatrix} = \mathbf{x} \doteq \begin{bmatrix} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{bmatrix}^T \end{array} \tag{2}$$

This representation is the very simplest one can immagine, however we will show that it is not the most appropriate for motion estimation.

It is well known [15] that from the projection of 5 or more points it is possible to reconstruct motion and structure (position of points in 3D space) between two views up to a scale factor multiplying the inverse depth and the translation. Such ambiguity can be overcome as soon as some scale information is available as the size of an object, the norm of translation etc. The recent literature proposes a variety of techniques for recovering structure and motion recursively [4, 16, 1, 20, 23]. All of these schemes are essentially based on the same formalization of the problem (1,2). In particular [1] is based on the model (1,2,3,4) which will be described later, with the structure referred to the observer's reference at time 0 and a more general model of perspective projection. [20] recovers instantaneous motion from 2 frames and feeds it to a model similar to (1,2), hence at each step motion is considered known and it does not exploit a dynamical model. [23] also computes instantaneous motion at each step as [20], and then inserts it into the state dynamics with a model similar to the one used by [1].

In this paper we will discuss the fundamental limitations of the model (1,2), which will lead to a new and general perspective for viewing the motion problem. We will then present three different ways of approaching it which come naturally after the new formalization of the problem. Our new approach is driven by the same goals of recursiveness, optimal noise rejection, real time operation, and is inspired by the work of Longuet-Higgins [15] for representing rigid motions.

We first discuss few different interpretations of structure and motion estimation which serve to motivate the schematization of structure and motion estimation. The remainder of the paper is devoted to motion estimation. We introduce and describe the properties of the essential manifold, and we discuss how motion is represented and estimated on the essential manifold. We introduce then three approaches for solving the motion problem which are unified within this representation. Finally we compare the three schemes against each other and other existing motion estimation schemes.

1.1 Few different interpretations of motion estimation: observability and the separation of motion from structure estimation

The equations (1,2) can be regarded as a dynamical system describing the motion of points in 3D space, having a projection as measurement equation. In this framework motion can be viewed as the input of the system, and hence a motion estimator should "invert" such a system and produce motion from time varying projection of feature points. Since the initial condition of such a model (structure at time zero) is not known, we have an "unknown-input/present-state" observability problem. It has been shown by the authors, and will be presented elsewhere, that the inverse system is essentially instantaneous, and hence it does not exploit recursiveness and its benefits. This is due to the fact that the system (1,2) is *driftless* [9, 19]; a common trick to overcome such a problem is to use *dynamic extension*, i.e. to consider the derivative of the input as driving the system and to include the true input into the state dynamic. We augment (1) with the equations

$$\dot{R}(t) = n_R(t) \quad (3)$$

$$\dot{T}(t) = n_T(t) \quad (4)$$

and leave the measurement equation (2) unchanged. Since we do not know n_R and n_T, we need to make some hypothesis. The trick is to suppose n_R and n_T are are particular instances of a stochastic process, for example the image of a zero-mean white gaussian noise, which corresponds to modeling motion as a first order random walk (brownian motion). This approach is used in [23, 1].

The above is completely equivalent to viewing motion as unknown parameter in the model (1,2), which needs to be indentified. Motion can hence be viewed as a mixed estimation-identification process.

Once inserted motion into the state dynamics we have transformed the motion problem to a state estimation problem for a dynamical system driven, for example, by zero-mean white gaussian noise. A fundamental issue in state estimation is of course *observability*, which for linear systems is a necessary and sufficient condition for the existence of an observer with spectrally assignable error dynamics. For nonlinear systems the issue is more subtle [9, 19], however the traditional methods for state estimation are based upon linearizing the trajectory about the current state and hence suffer the limitations of the local observers [11].

The system under investigation (1,2,3,4) has the peculiarity of not only having a linearization which is not observable, but of also being non "locally weakly observable", hence the local linearization-based methods are not guaranteed to work.

Observability is of course a property of the *model*, not of the system itself. In fact the visual motion problem is nonlinearly observable, even though it is non locally weakly observable. In [23] we have presented an architecture which is based upon splitting motion estimation from true "structure from motion", so that each module gains observability properties. This is correct as long as each step is accompanied by a complete error characterization (at least up to second order statistics), so that information can propagate across modules with proper weighting.

1.2 Rigid motion and the essential constraint

Suppose the scene is a single rigid object, moving with $T(t), R(t)$ between two time instants. Then it is immediate (see fig. 1) to see that the vector $\mathbf{X}$, describing the coordinates of the generic point

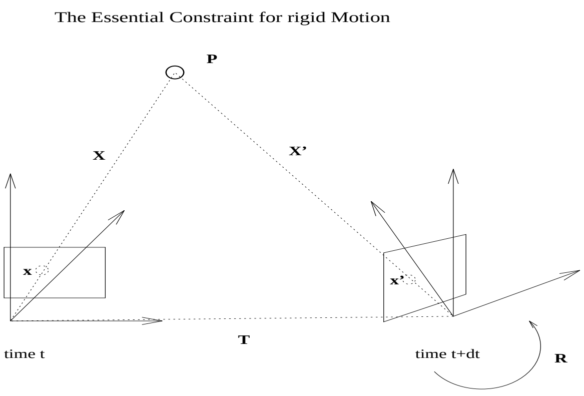

Figure 1: *The essential constraint*

at time t, the vector $\mathbf{X}'$ of coordinates at time $t+1$ and T, are coplanar, and therefore their triple product is zero. This is true of course also for $\mathbf{x}$, $\mathbf{x}'$ and T, since $\mathbf{x}$ is the projective line of $\mathbf{X}$ and has its same direction. When expressed with respect to a common reference, for example that at time t, we write the triple product as

$$\mathbf{x}'^T_i R(T \wedge \mathbf{x}_i) = 0 \; \forall \; i \;\; = \;\; 1 : N. \tag{5}$$

It turns out that the above constraint is also sufficient to characterize rigid motions [17, 15]. The operator $T\wedge$ is a skew symmetric matrix

$$T\wedge = \begin{bmatrix} 0 & -T_3 & T_2 \\ T_3 & 0 & -T_1 \\ -T_2 & T_1 & 0 \end{bmatrix} \doteq S.$$

S belongs to the lie algebra of skew symmetric 3×3 matrices, which is called $so(3)$ [18]. Following Longuet-Higgins we call

$$\mathbf{Q} \doteq RS$$

so that the above constraint, which we will call the "essential constraint", becomes

$$\mathbf{x}'^T_i \mathbf{Q} \mathbf{x}_i = 0. \tag{6}$$

with $\mathbf{Q} \doteq R \circ (T\wedge) \doteq RS$; $R \in SO(3)$; $S \in so(3)$. Since the constraint is linear in $\mathbf{Q}$, we can rewrite it as

$$\chi(x'(t), x(t))q(t) = 0$$

where χ is an $N \times 9$ matrix combining x_i, x'_i and q is a nine-vector obtained by stacking the columns of $\mathbf{Q}$. We will also use the notation $F_{x'(t),x(t)}(\mathbf{Q}(t)) \doteq \chi(x'(t), x(t))q(t) = 0$. The generic row of χ is $[x_1 x'_1 \; x_2 x'_1 \; x'_1 \; x_1 x'_2 \; x_2 x'_2 \; x'_2 \; x_1 \; x_2 \; 1 \,]$.

2 The Essential Space

We have seen that a rigid motion can be represented as an element of the Lie group $SE(3)$, which is naturally embedded in $GL(4)$, the linear group of real 4×4 matrices, via homogeneous coordinates:

$$\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \in SE(3) \subset GL(4) \sim \mathbb{R}^{16}.$$

We have indeed seen that rigid motion can be encoded using the essential constraint (6) based on the 3×3 matrix $\mathbf{Q} \doteq R(T\wedge) \in GL(3) \sim \mathbb{R}^9$. Since we can reconstruct translation only up to a scale factor, we can consider $\mathbf{Q}$ to belong to $\mathbb{R}P^8$ instead than $\mathbb{R}^9$. 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 not zero. The zero-norm translation case can be dealt with separately, and we will discuss it later. Now for simplicity we assume $\|\mathbf{Q}\|_2 = \|T\| = 1$. The matrix $\mathbf{Q}$ belongs to the space

$$E \doteq \{RS | R \in SO(3), S \doteq T\wedge \in so(3), \|T\| = 1\} \subset \mathbb{R}P^8$$

which is called the *essential space.*

The essential space encodes rigid motion in a more compact way than $SE(3)$, the price being that we loose the group structure. Given that we lost the group structure, we want to see if we still preserve some topological properties ($SE(3)$, as a Lie group, is also a smooth manifold [3]). We have in fact the following

Theorem 2.1 *E is a topological manifold of class at least C_0.*

Proof:
E inherits the topology from $GL(3)$. Consider the map

$$\begin{array}{rcl} \Phi : E & \rightarrow & \mathbb{R}P^2 \times \mathbb{R}^3 \sim \mathbb{R}^5 \\ \mathbf{Q} & \mapsto & \begin{bmatrix} T \\ \Omega \end{bmatrix} = \begin{bmatrix} V_{.3} \\ U R_Z(\frac{\pi}{2}) V^T \end{bmatrix} \end{array} \tag{7}$$

where U, V are defined by the Singular Value Decomposition (SVD) [6] of $\mathbf{Q} = U \Sigma V^T$, $V_{.3}$ denotes the third column of V and $R_Z(\frac{\pi}{2})$ is a rotation of $\frac{\pi}{2}$ about the Z axis. T, Ω are the local coordinates of $\mathbf{Q}$. Note that Ω is the rotation 3-vector corresponding to the 3×3 rotation matrix $U R_Z(\frac{\pi}{2}) V^T$ and is obtained using the Rodrigues' formulae [13], which are in fact a local coordinate parametrization of $SO(3)$. It follows from the properties of the SVD that Φ is continuous, and furthermore it is bijective. It will be shown in appendix B that $\Sigma = diag\{1\ 1\ 0\}$ and hence the subspaces $< V_{.1}, V_{.2} >$ and $< U_{.1}, U_{.2} >$ can switch. This happens however without affecting continuity of T and Ω. The inverse map is simply

$$\begin{array}{rcl} \Phi^{-1} : \mathbb{R}P^2 \times \mathbb{R}^3 & \rightarrow & E \\ \begin{bmatrix} T \\ \Omega \end{bmatrix} & \mapsto & e^{(\Omega\wedge)}(T\wedge). \end{array}$$

which is smooth. Hence Φ is a homeomorphism, and E is a topological manifold of class at least C_0. **Q.E.D.**[1]
E also has the structure of an algebraic variety [17], which we will not discuss in this paper.

Remark 2.1 *E is an "essential" representation of $SE(3)$, which has no group structure, but still has interesting topological properties, first of which that of being naturally immersed in an euclidean space of minimal dimensions. Note also that elements of E are composed by a symmetric part, pertaining to R, and a skew-symmetric part, pertaining to T. The homeomorphism Φ is doing nothing but separating these two parts.*

[1] For a purist, the above representation of motion in local coordinates is not strictly correct, since V represents the (discrete-instantaneous) translation vector, while Ω is a true velocity vector. We allow to confuse V with its velocity representation, since the two are uniquely related via a diffeomorphism.

3 Motion representation on the essential space

A rigid motion with unit norm translation can be represented as an element of the essential manifold E. For non-unit translations (but still positive norm), it is sufficient to scale $\mathbf{Q}$ to $\|T\|\mathbf{Q}$, since $\|RT\wedge\| = \|T\|$. This can be done easily since, as shown in appendix B, the singular values of the scaled $\mathbf{Q}$ are $\{\|T\|, \|T\|, 0\}$.

Suppose we observe N points moving in space under some rigid motion $(R(t), T(t))$, through their projection onto the image plane: $\mathbf{x}_i(t)$; $i = 1 : N$. At each time instant we have a set of N constraints in the form

$$F_{x'(t),x(t)}(\mathbf{Q}(t)) \doteq \chi q = 0,$$

and hence q lies at the intersection between the essential manifold and the linear variety $F^{-1}_{x'(t),x(t)}(0)$, i.e. the null space of χ intersected with the unit ball in $\mathbb{R}^9$ (see fig. 2).

Note that even imposing unit norm there is still a sign indeterminacy on $\mathbf{Q}$, which accounts for the two solutions $\mathbf{Q}_1$ and $\mathbf{Q}_2$. These solutions become four when transformed to local coordinates, due to the arbitrary sign of the rotation $R_Z(\pm\frac{\pi}{2})$ (see [7, 21]). These ambiguities can be overcome by imposing the positive depth constraint: in fact out of the four different combinations of R and T, only one corresponds to points which are in front of the observer [24, 7, 21].

As time goes by, the point $\mathbf{Q}(t)$, corresponding to the actual motion, describes a trajectory on E and one in local coordinates. By definition we have:

$$\dot{\mathbf{Q}}(t) \doteq \nu(t) \in T_{\mathbf{Q}}E$$

where $T_{\mathbf{Q}}E$ denotes the tangent space to E at $\mathbf{Q}$ [3], or

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

The last two equations are in fact just a *definition* of their right-hand side, since we do not know $n_{\mathbf{Q}}(t)$ or $\nu(t)$. If we want to make use of a model for estimating $\mathbf{Q}$ we have to make assumptions about ν or $n_{\mathbf{Q}}$. This will be done in the next sections. For now we will consider the previous equations as either a continuous time or a discrete time dynamical model for $\mathbf{Q}$ on the essential manifold, having ν or $n_{\mathbf{Q}}$ as *unknown* inputs. If we accompany one of them with the essential constraint, we get

$$\begin{aligned} \mathbf{Q}(t+1) &= \mathbf{Q}(t) + n_{\mathbf{Q}}(t) \ ; \ \mathbf{Q} \in E && (8) \\ 0 &= F_{x'(t),x(t)}(\mathbf{Q}(t)) + m(t) && (9) \end{aligned}$$

where $m(t)$ is a noise process which will be characterized in appendix A.

This shows that motion estimation can be viewed as state estimation of a dynamical system defined on a topological manifold and having an implicit measurement constraint and unknown input.

As it can be seen the system is "linear" (both the state equation and the essential constraint are linear in $\mathbf{Q}$), but the word "linear" is not proper in this context, since E is not a linear space.

4 Recursive estimation on the Essential Space

We have seen in the previous section that motion estimation can be regarded as estimation of the state of a system of a difference equations on the essential manifold having unknown inputs.

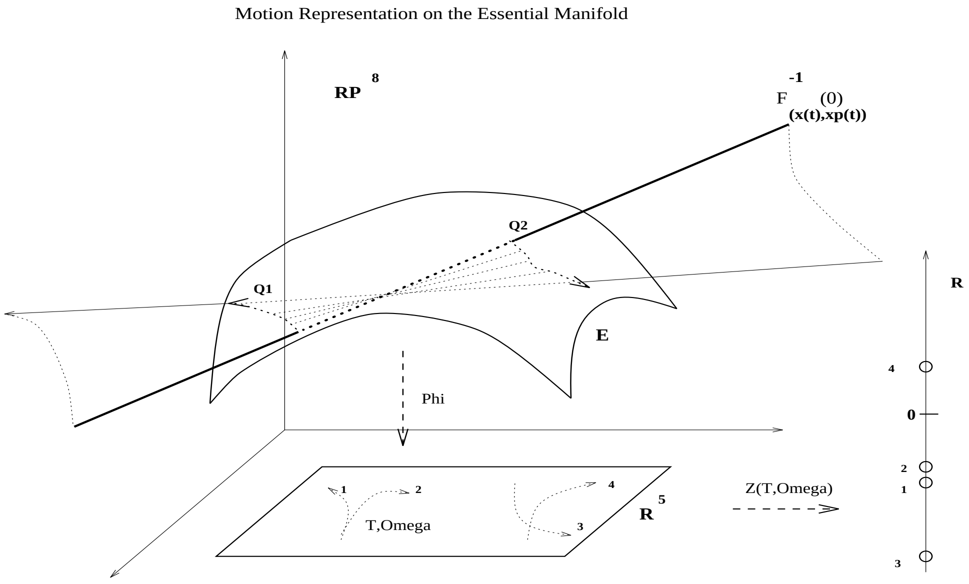

Figure 2: *Structure of the motion problem on the Essential Space*

The first approach we describe consists in composing equations (8) and (9) with the local coordinate chart Φ defined in eq. (7), 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* (brownian motion) *in* $\mathbb{R}^5$ (see fig. 3). The problem then becomes that of estimating the state of a nonlinear system driven by white, zero-mean gaussian noise (see fig. 5). This will be done using a variation of the traditional Extended Kalman Filter (EKF) [10] for systems with implicit measurement constraints, which is derived in appendix A.

In the second approach we change the model for motion: in particular we assume motion to be a *first order random walk in* $\mathbb{R}P^8$ *projected onto the essential manifold* (see fig. 4). We will see that this leads to a method for estimating motion via solving at each step a *linear estimation* problem in the linear embedding space and then "projecting" the estimate onto the essential manifold (see fig. 5). The notion of projection onto the essential manifold will be made clear later.

It is very important to understand that these assumptions about motion can be validated only a posteriori. In general we can only observe that the first method solves a strongly nonlinear problem with techniques which are based upon linearization of the system about the current reference trajectory, so that the linearization error can be relevant. The second method does not involve any linearization, while it imposes the constraint of belonging to the essential manifold in a weaker way. This approach has indeed a very transparent structure which can be studied in full detail.

The third method is based upon splitting the iteration in a nonlinear fixed-point iteration at each fixed time, for which local results of convergence are available, and a propagation of information across time which is linear and has all of the desirable asymptotic properties.

The next three sections are devoted to describing these three techniques. We will also show 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 and motion estimation scheme.

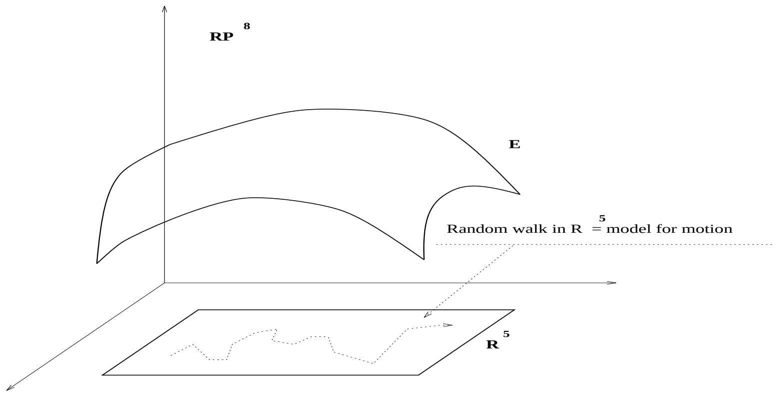

Figure 3: *Model of motion as a random walk in* $\mathbb{R}^5$

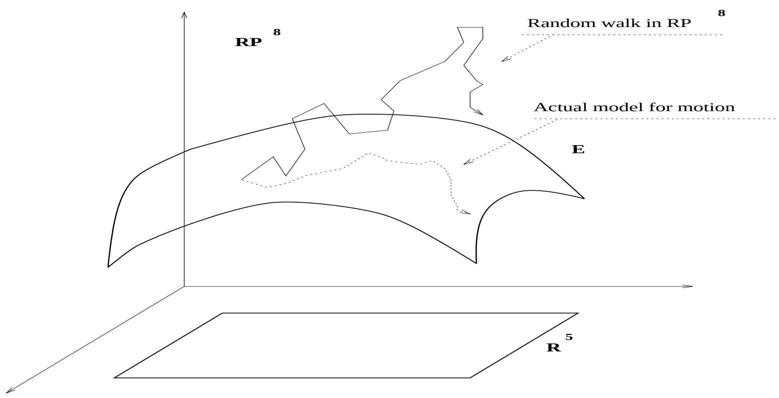

Figure 4: *Model for motion as projection of a random walk in* $\mathbb{R}P^8$ *onto the essential manifold.*

4.1 Local coordinates estimator

Consider composing the system (8,9) with the map Φ defined in (7):

$$\begin{array}{rcl} \Phi : E & \rightarrow & S^2 \times \mathbb{R}^3 \sim \mathbb{R}^5 \\ \mathbf{Q} & \mapsto & \xi \doteq \left[\begin{array}{c} T \\ \Omega \end{array} \right] \end{array}$$

where T is expressed in spherical coordinates for radius one, for convenience of representation. Then the system in local coordinate becomes

$$\begin{aligned} \xi(t+1) &= \xi(t) + n_\xi(t) \; ; \; \xi(t_0) = \xi_0 && (10) \\ 0 &= F_{x(t),x'(t)}(\mathbf{Q}(\xi(t))) + m(t) \doteq \tilde{F}(\xi(t), t) + m(t). && (11) \end{aligned}$$

As we said we model motion (ξ) as a first order random walk, which is zero-mean gaussian white noise integrated once. Hence $n_\xi(t) \in \mathcal{N}(0, R_n)$ for some R_n which is referred to as variance of the model error. While the above assumption is rather arbitrary and can be validated only a posteriori,

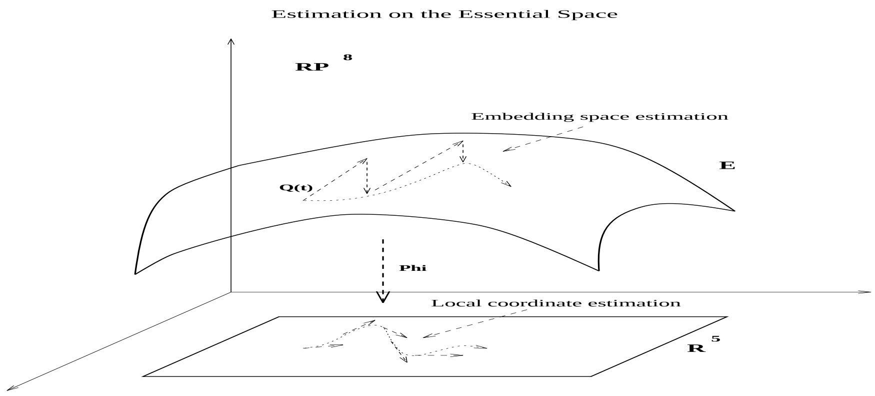

Figure 5: *Estimation on the Essential Space*

it is often safe to assume that the noise in the measurements $x(t), x'(t)$ is a white zero-mean gaussian process with variance $R_{\mathbf{x}}$.

The system above is now in a form suitable for using an Extended Kalman Filter (EKF) [10]. We have however an implicit measurement constraint, for we have to make some slight variation to adapt the usual EKF: we will call this adaptation Implicit Kalman Filter (IEKF). A derivation of the IEKF is reported in appendix A: it is based upon the fact that the variational model about the best current trajectory is linear and *explicit*, so that a linear update equation can be derived and a pseudo-innovation process can be defined.

Finally the equations of the estimator can be summarized: call $C \doteq \left(\frac{\partial \tilde{F}}{\partial \xi}\right)$ and $D \doteq \left(\frac{\partial \tilde{F}}{\partial \mathbf{x}}\right)$, where $\mathbf{x}$ are the measurements of the feature positions on the image plane.

Prediction step:

$$\begin{aligned} \hat{\xi}(t+1|t) &= \hat{\xi}(t|t) \; ; \; \hat{\xi}(0|0) = \xi_0 && (12) \\ P(t+1|t) &= P(t|t) + R_n \; ; \; P(0|0) = P_0 && (13) \end{aligned}$$

Update step:

$$\begin{aligned} \hat{\xi}(t+1|t+1) &= \hat{\xi}(t+1|t) - L(t+1)\tilde{F}(\hat{\xi}(t+1|t), t) && (14) \\ P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)R_m(t+1)L^T(t+1) && (15) \end{aligned}$$

Gain:

$$\begin{aligned} L(t+1) &= P(t+1|t)C^T(t+1)\Lambda^{-1}(t+1) && (16) \\ \Lambda(t+1) &= C(t+1)P(t+1|t)C^T(t+1) + R_m(t+1) && (17) \\ \Gamma(t+1) &= I - L(t+1)C(t+1) && (18) \end{aligned}$$

Innovation variance:

$$R_m(t+1) = D(t+1)R_{\mathbf{x}}D^T(t+1) \qquad (19)$$

Note that $P(t|t)$ is the variance of the motion estimation error which is used as variance of measurement error by the subsequent modules of the motion and structure estimation scheme. This formulation was first introduced by Di Bernardo et al. [2] in a slightly different formulation. The implicit Kalman filter was used by other researchers such as Darmon [5], Faugeras [14, 25] and Heel [8].

4.2 The Essential estimator

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 $\mathbb{R}P^8$ (see fig. 4). The "projection" operator onto the space E is denoted by $pr_{<E>}(\cdot)$ and is defined as follows:

$$\begin{array}{rcl} pr_{<E>} : GL(3) & \rightarrow & E \\ M & \mapsto & U diag\{1,1,0\}V^T \end{array} \tag{20}$$

where $U, V \in GL(3)$ are defined by the Singular Value Decomposition of $M = U\Sigma V^T$. The fact that this operator maps onto the essential manifold is proved in appendix B. Note that the projection minimizes the Frobenius norm and the 2-norm of the distance from a point in $GL(3)$ to the essential manifold [7, 17, 25].

Now we define the operator $\oplus$ that takes two elements in $GL(3)$, sums them and then projects the result onto the essential manifold:

$$\begin{array}{rcl} \oplus : GL(3) \times GL(3) & \rightarrow & E \\ M1, M2 & \mapsto & Q = pr_{<E>}(M1 + M2) \end{array}$$

where the symbol $+$ is the usual sum in $GL(3)$. With the above definitions our model for motion becomes simply

$$Q(t+1) = Q(t) \oplus n_Q(t) \tag{21}$$

where $n_Q(t) \in \mathcal{N}(0, R_{n_Q})$ is represented by a white zero-mean gaussian noise in $\mathbb{R}P^8$. If we couple the above equation with (9) we have again a dynamical model on an euclidean space (in our case $\mathbb{R}^9$) driven by white noise. The Essential Estimator is the least variance filter built for the above model, and corresponds to a linear Kalman filter update in the embedding space, followed by a projection onto the essential manifold. Note that in principle the gain could be precomputed offline, for each possible configuration of motion and feature positions.

The equations of the essential estimator are written for $q(t)$ rather than for $\mathbf{Q}$. The two forms are equivalent, but for the latter the gain would be obtained by multiplying $3 \times 3 \times 3$ real tensors, which is not easily implemented.

Prediction step:

$$\hat{q}(t+1|t) = \hat{q}(t|t) \;;\; \hat{q}(0|0) = q_0 \tag{22}$$
$$P(t+1|t) = P(t|t) + R_n \;;\; P(0|0) = P_0 \tag{23}$$

Update step:

$$\hat{q}(t+1|t+1) = \hat{q}(t+1|t) \oplus L(t+1)\chi(t)\hat{q}(t+1|t) \tag{24}$$
$$P(t+1|t+1) = \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)R_m(t+1)L^T(t+1) \tag{25}$$

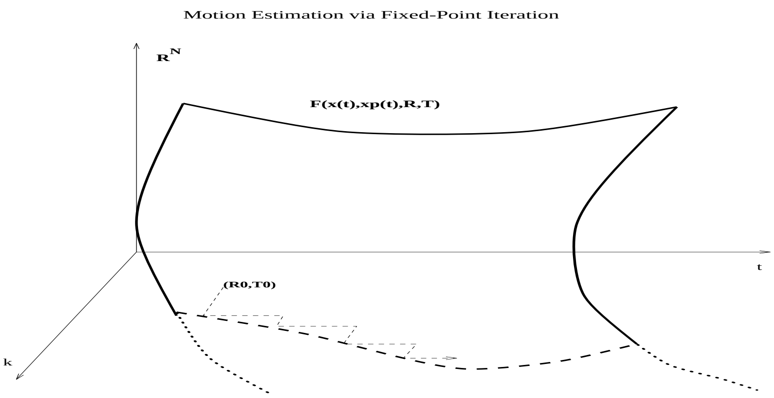

Figure 6: *2-D iteration for solving a "moving fixed point" problem*

Gain:

$$
\begin{aligned}
L(t+1) &= -P(t+1|t)\chi(t)\Lambda^{-1}(t+1) && (26)\\
\Lambda(t+1) &= \chi(t)P(t+1|t)\chi(t) + R_m(t+1) && (27)\\
\Gamma(t+1) &= I - L(t+1)\chi(t) && (28)\\
R_m(t+1) &= D(t+1)R_{\mathbf{X}}D^T(t+1) && (29)
\end{aligned}
$$

4.3 2-D fixed-point estimator

The IEKF update seen in the previous section resembles closely the Newton-Raphson iteration for solving a square problem $f(x) = 0$:

$$\hat{x}(k+1) = \hat{x}(k) - L_{NR}(k)f(\hat{x}(k))$$

where $L_{NR} = J_f^{-1}(\hat{x}(k))$ and J_f is the jacobian of f.

The IEKF is in fact solving a problem of finding $\xi(t)$ such that $\chi(t)q(\xi(t)) \doteq \tilde{F}(\xi(t), t) = 0$. The function $\tilde{F}$ varies with time, as a new measurement becomes available. In this sense the IEKF is a sort of "adaptive" Newton-Raphson iteration. If $\tilde{F}$ was not a function of time, we could perform a true Newton-Raphson iteration, for which local convergence results are known as well as bounds on the convergence rate. This suggests to fix t and perform a newton iteration along the k coordinate. Once this is done we define an iteration in time, which now is *linear*, and has all the desirable asymptotic properties (see fig. 6).

4.3.1 Iteration at each fixed time

At each time instant a new set of measurements becomes available in the form of position of projected points onto the image plane, encoded in $\chi(t)$. The essential constraint imposes

$$\chi(t)q(\xi(t)) \doteq \tilde{F}(\xi(t), t) = 0 \quad \forall t$$

Define $T_\xi \tilde{F} : \mathbb{R}^5 \rightarrow \mathbb{R}^N$ to be the derivative of the map $\tilde{F}$ and $J_{\tilde{F}}(\xi)$ the Jacobian matrix of $\tilde{F}$ calculated at the point ξ. Note that $\tilde{F}$ is differentiable as it is a composition of differentiable functions. Suppose that there exists some ξ^* such that $\tilde{F}(\xi^*, t) = 0$ for our particular (fixed) t.

Then we can write a first order expansion around the point ξ^*, starting from some point ξ_0 (we neglect time indices for the remainder of this section):

$$\tilde{F}(\xi^*) = \tilde{F}(\xi_0) + T_{\xi_0}\tilde{F}(\xi^* - \xi_0) + \|\xi^* - \xi_0\| E(\xi^*, \xi_0) = 0.$$

The usual Newton-Raphson method is based upon neglecting the higher order term E, and approximating ξ^* iteratively with ξ_k, with the iteration defined by

$$\tilde{F}(\xi_k) \doteq J_{\tilde{F}}(\xi_k)[\xi_{k+1} - \xi_k].$$

At each iteration we solve for Y the linear problem

$$J_{\tilde{F}}(\xi_k)Y = \tilde{F}(\xi_k)$$

and then define$\xi_{k+1} \doteq \xi_k + Y$. In the case $N(t) \geq 5$, we can assume without loss of generality that $J_{\tilde{F}}$ has full column rank 5, i.e. $Null(J_F) = \emptyset$ (this will be true for points in general position, and even for singular configurations, when noise is present). In general, also due to noise, we can expect $\tilde{F}$ not to be in the range space of $J_{\tilde{F}}$: $\tilde{F}(\xi_k) \notin Ra(J_{\tilde{F}}(\xi_k))$, so that we will be seeking for Y such that $J_{\tilde{F}}(\xi_k)Y$ is the projection of $\tilde{F}(\xi_k)$ onto the range space of $J_{\tilde{F}}(\xi_k)$. At the next iteration such a space will be modified and we will now be seeking for the projection of $\tilde{F}(\xi_{k+1})$ onto the range space of $J_{\tilde{F}}(\xi_{k+1})$. The Newton-Raphson iteration is therefore defined as

$$\xi_{k+1} \doteq \xi_k - L_{NR}(k)\tilde{F}(\xi_k).$$

where $L_{NR}(k) \doteq \left(J^T_{\tilde{F}}(\xi_k)J_{\tilde{F}}(\xi_k)\right)^{-1} J^T_{\tilde{F}}(\xi_k)$. The map defined by the right-hand side of the above equation is contractive as long as $J_{\tilde{F}}(\xi_k)$ 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 ξ^*, which best explains the noisy measurements x_i, x'_i; hence we have $\xi^* = \xi + n_\xi$ where n_ξ is a noise term whose variance can be inferred from the variance of x_i, x'_i and a linearization of the scheme about zero-noise. The measurement obtained at each fixed time, together with its variance, is fed to a time-integration step, which we describe next.

4.3.2 Propagation along time: disambiguation of local minima

At each fixed time the iteration along k converges to a fixed point $\xi^*(t)$, then we can propagate the information across time with a similar iteration:

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

which implements a linear Kalman filter based upon the model

$$\begin{aligned} \xi(t+1) &= \xi(t) + n(t) && (30)\\ \xi^*(t) &= \xi(t) + n_\xi(t) && (31) \end{aligned}$$

where n is the error of the random walk model for motion, which we assume to be white zero-mean and gaussian, and n_ξ is the error made by the fixed-time iteration. L is the usual linear Kalman gain [12, 10]. The above model has all the desirable properties, as it satisfies the conditions of the fundamental theorem of the asymptotic theory of Kalman Filtering, which guarantees that there exists a unique and positive definite solution to the riccati equation defining the variance of the estimation error.

Suppose now that the $k-$iteration has converged to a local minimum, which is a motion compatible with the current observation. At the next step the $t-$iteration will predict a motion which is no longer compatible with the current observations, and the $k-$iteration will switch to a different minimum. We have observed that after some switches the algorithm converges to a global minimum.

5 Various issues in motion estimation

5.1 Singular case: what if we observe less than 5 points?

Suppose now we are in the situation $N(t) < 5$. Then the essential constraint will have a preimage which is a whole subspace, and its intersection with the essential manifold (see fig. 2) will no longer be two points on E. However suppose we move under constant velocity motion; at each time instant we get a new measurement set and a new essential constraint, whose preimage intersects the essential manifold in a new variety. The intersection of these varieties eventually comes to a single point in the essential manifold. This point will be described following the derivation of the 2-D filter. Before getting into the details, observe that if we just let any of the three described algorithms work with less than 5 points, this will try to "average" values which belong to different subspaces. Those belonging to the intersection have higher probability to appear and will eventually pop out. In essence the filters would be "intersecting via averaging".

In order to see this point, consider the 2-D algorithm for less than 5 points and suppose, again without loss of generality, that $J_{\tilde{F}}(\xi_k)$ has full row rank $Ra(J_{\tilde{F}})^{\perp} = \emptyset$. Then the problem $J_{\tilde{F}}(\xi_k)Y = \tilde{F}(\xi_k)$ introduced in the derivation of the 2-D estimator can be solved for Y only up to a subspace of $Ra(J_{\tilde{F}})$. However we can exploit the constant velocity model, i.e. $\xi(t+1) = \xi(t)$. Such a model could serve also when the velocity is varying slowly compared with the sampling rate. At time $t+1$ a new set of measurements becomes available, which is subject to its own essential constraint $\tilde{F}(\xi(t+1), t+1) = 0$. Due to the constant velocity model, the above constraint becomes $\tilde{F}(\xi(t), t+1) = 0$. This can be appended to the constraints considered previously, until we reach some integer r for which the jacobian of

$$\mathcal{F}(\xi(t), t, r) \doteq \begin{bmatrix} \tilde{F}(\xi(t), t) \\ \tilde{F}(\xi(t), t+1) \\ \vdots \\ \tilde{F}(\xi(t), t+r) \end{bmatrix}$$

has full rank. We can then consider the problem of finding Y such that

$$J_{\mathcal{F}}(\xi(t), t, r)Y = \mathcal{F}(\xi(t), t, r)$$

Hence the case $N(t) < 5$ reduces to the previous case when the velocity is kept constant.

It is interesting to note that extended observations of the motion of *one only point* are sufficient to determine the observer motion. This seems counterintuitive, as there are many possible *instantaneous* 3-D motions corresponding to the same projected motion. This is on fact true *locally*, but the different motions are identifiable as time increases.

5.2 Zero-translation case

The above schemes were described under the standing assumption of non-zero translation; we claim that there is no loss of generality in this assumption.

When translation is zero there is no parallax, and we are not able to perceive structure (depth). The essential constraint leaves rotation undetermined, however we realize that we can still recover rotation and hence update the previous estimate of structure correctly. In fact, due to noise in the measurements of x_i, x_i', there will be always a small translation compatible (in least squares sense) with the observed points. This translation is automatically scaled to norm one by the algorithm. This allows to recover the correct rotation and scales depth by the inverse norm of the true translation. If we keep track over time of the scale factor, as described in the next section, we can scale the norm of translation and hence update depth and rotation within the correct scale.

Hence zero-translation is, thanks to noise, a "zero-measure set", and the algorithms we present are able to recover structure and motion correctly even when the observer is undergoing a purely rotational motion, as shown in the simulations.

5.3 Recovery of the scale factor

Translation and depth can be recovered only up to a scale factor. However once some scale information is available *at one time* it can be propagated across time allowing to recover motion and structure within the correct scale. This has been tested and discussed in the experimental section.

5.4 A remark on camera calibration

In introducing our algorithms we have described the camera as a simple static map from $\mathbb{R}^3$ to $\mathbb{R}P^2$. This map can be made more general, allowing also a time-varying focal length and inserting it into the state dynamics, as we have done for motion, with a statistical model. As long as the resulting model preserves observability properties this will allow to recover camera focal length together with motion. This has been implemented by Azarbayejani et al. [1] for the standard formulation (1,2).

5.5 Implementation and tuning

The state models we have used are in essence first order random walks whose variances do not have a physical meaning. These parameters have to be set using a custom procedure which is called "tuning" and consists in choosing them so that the estimation error is as uncorrelated as possible. Standard tests as for example the "cumulative periodogram" are available for the purpose.

6 Experimental assessment

We have tested the described algorithms on a variety of motion and structure configurations. We report the simulations performed on the same data sets of [23]. These consist of views of a cloud of points under a discontinuous motion with singular regions (zero-translation and non-zero rotation) and are described in detail in [22]. Gaussian noise with 1 pixel Std has been added to the measurements. Simulations have been performed with a variable number of points down to 1 point for constant velocity motion, and show consistent performance.

The local coordinates estimator

In fig. 7,8 we show the three components of translational and rotational velocity as estimated by the local coordinate estimator. Convergence is reached in less than 20 steps. Tuning has been performed, as with the other schemes, within an order of magnitude, and the Std of the estimation error are reported in the tables below. 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 linearization error is negligible, while in this case it is not. Initialization is performed with one step of the traditional Longuet-Higgins algorithm. The computational cost of one iteration os of about 300 Kflops for 20 points.

Note that if we have available some dynamical model for motion we can easily insert it into the state model. This is not true for the essential estimator, which is hence less flexible than the local coordinates one.

The Essential estimator

In fig. 11 we show the 9 components of the essential matrix as estimated by the essential estimator. convergence is 4 times slower than the local coordinate version, but each step is 10 times faster. Note that in principle the gains can be precomputed offline, for each possible configuration of points in the image plane. We have noted step-like convergence with plateaus followed by switching regions. These correspond to switching of the first two eigenspaces of the SVD of $\mathbf{Q}$. When brought to local coordinates we have estimates for rotation and translation 9,10. It is noted that the homeomorphism Φ can have singularities due to noise when the last eigenspace is changed with one of the other two. This causes the spikes observed in the estimates of motion. However note that there is no transient to recover, since the errors do not occur in the estimation step, but in trasferring to local coordinates. The switching can be avoided by a higher level control on the continuity of the singular values. The computational cost amounts to circa 41 Kflops per each step for 20 points.

The 2-D iteration

The performace of the 2-D iteration can be viewed in fig. 12,13. This scheme proved very accurate after proper initialization, even though the error analysis used for calculating the variance of the estimates at each fixed time was approximate. Speed can be adjusted by varying the number of iterations at each fixed time. We have noticed that this converges after a number of steps between 3 and 7. The cost of the scheme for 7 iterations and 20 points is 100 Kflops. The simulations reported were done using a constant variance of the error of the k-iteration, and hence show a higher variance than the other schemes.

We now summarize the perfomance of the three schemes: mean and Std are computed between time 30 and 50 for the local coordinate scheme and the 2-D iteration, while between time 150 and 200 for the essential estimator.

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

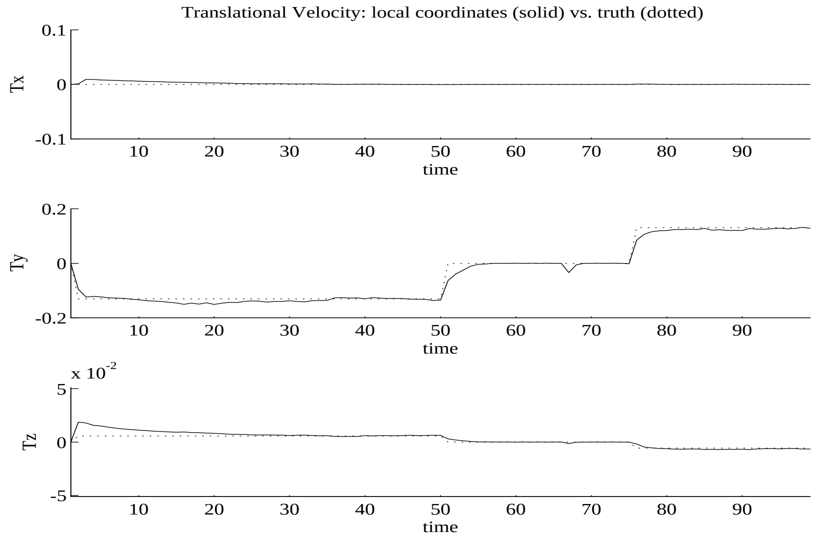

Figure 7: Components of translational velocity as estimated by the local coordinates estimator. The ground truth is shown in dotted lines.

Experiments on real image sequences

We have tested our schemes on a sequence of 10 images of the rocket scene (see fig 14). There are 22 feature points visible, and the standard deviation of the location error on the image plane is about one pixel.

The local coordinates estimator has a transient of about 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 15-17. We did not perform any ad hoc tuning, and the setting was the same used in the simulations described at the previous paragraphs. In fig. 15 we report the 6 motion components as estimated from the local coordinates estimator and the corresponding ground truth (in dotted lines); the estimation error is plotted in figure 16. 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.17 we report the norm of the pseudo-innovation of the filter, which converges to a value of about 10^{-3} in less than 5 steps.

In this experiment we have used the true norm of translation as the scale factor. We have also run simulations in which the scale factor was calculated by updating the estimate of the distance between the two closest features, as in the experiments described in the previous paragraphs. In this case however convergence is slower, as the innovation norm reaches regime in about 20-25 steps.

7 Conclusions

We have presented a novel perspective for viewing motion estimation. This has resulted in three different approaches for solving the motion problem which are cast in a common framework and correspond to three different ways of tackling the same problem. Each scheme has its own person-

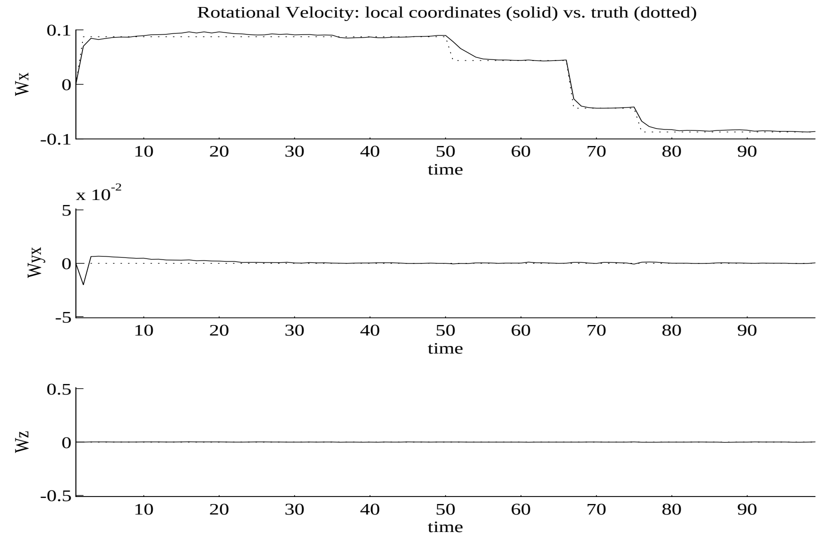

Figure 8: Components of rotational velocity as estimated by the local coordinates estimator.

ality, the essential filter being faster, the 2-D iteration more accurate and the IEKF more flexible. All the schemes enjoy common features such as recursiveness, allowing to exploit at each time all previous calculations, and noise rejection from exploiting redundancy. They all benefit independence on structure estimation, which makes the model observable and allows to deal easily with a variable number of points and feature set. Hence we do not need to track specific features through time, and we can deal easily with occlusion.

All the schemes produce, together with an estimate of motion, complete information about the reliability of such estimates, in the form of second order statistics of the estimation error.

The approaches can be interpreted as an extension of the Longuet-Higgins algorithm to infinite baseline, and a generalization of P-points N-frames theorems: they all work for any number of points provided that enough frames are viewed. Partial camera calibration can be inserted in the state model and hence estimated on line.

Simulations are presented for motions with discontiuities and singular regions.

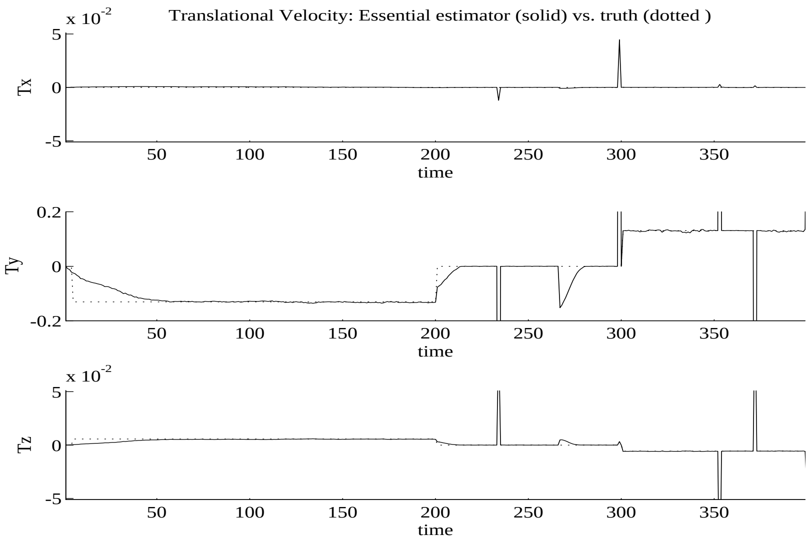

Figure 9: Components of translational velocity as estimated by the essential estimator. Note the spikes due to the local coordinates transformation. Note also that they do not affect convergence since they do not occur in the estimation process, but while transfering to local coordinates.

References

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

[2] 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.

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

[4] T. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE Trans. Pattern Anal. Mach. Intell.*, Jan. 1986.

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

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

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

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

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

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

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

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

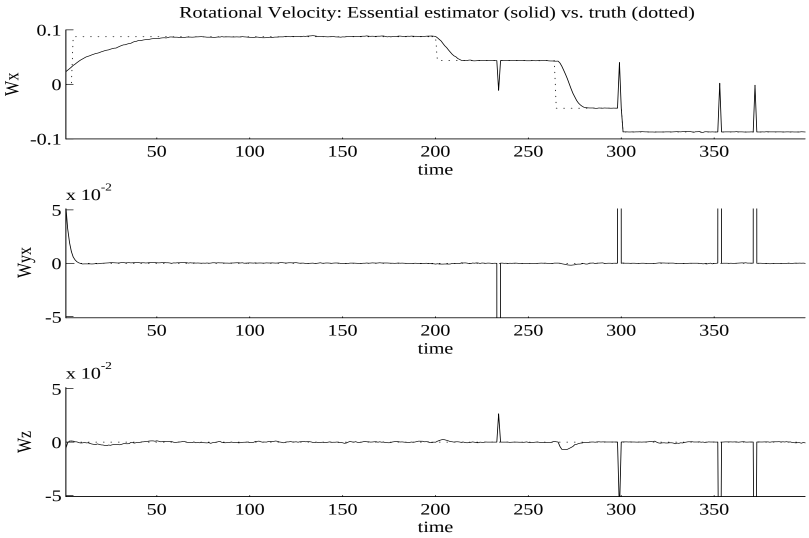

Figure 10: Components of rotational velocity as estimated by the local coordinates estimator. The ground truth is shown in dotted lines. Note the spikes due to the local coordinates transformation. Note also that there is no transient to recover since they do not occur in the estimation process.

[13] Z. Li, R.M. Murray, and S.S. Sastry. *A Mathematical Introduction to Robotic Manipulation.* Preprint, 1993.

[14] 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.

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

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

[17] S. Maybank. Springer Verlag, 1993.

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

[19] H. Nijimeijer and A.J.Van Der Shaft. *Nonlinear Dynamical Control Systems.* Springer Verlag, 1990.

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

[21] P. Perona and S. Soatto. Motion and structure from 2 perspective views of p points – algorithm and error analysis. *Technical Report CIT/CNS 23-93 – California Institute of Technology*, Oct. 1992.

[22] S. Soatto. and P. Perona. Time integration of visual information for the robust recovery of motion and structure. *Technical Report CIT/CNS 24-93 – California Institute of Technology*, Oct. 1992.

[23] S. Soatto, P. Perona, R. Frezza, and G. Picci. Recursive motion and structure estimation with complete error characterization. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 428–433, New York, June 1993.

[24] 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.

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

A Extended Kalman filter for implicit measurements

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

$$x(t+1) = f(x(t)) + v(t) \ ; \ x(t_0) = x_0$$

where $v(t) \in \mathcal{N}(0, Q_v)$. Suppose there is a measurable quantity $y(t)$ which is linked to x by the constraint

$$h(x(t), y(t)) = 0 \ \forall t. \tag{32}$$

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

$$y(t) = y_m(t) + w(t) \ : \ w(t) \in \mathcal{N}(0, R_w) \tag{33}$$

where the variance/covariance matrix R_w is known from knowledge of the measurement device.[2] The model we consider is hence of the form

$$\begin{aligned} x(t+1) &= f(x(t)) + v(t) \ ; \ x(t_0) = x_0 & (34) \\ h(x(t), y_m(t) + w(t)) &= 0. & (35) \end{aligned}$$

Construction of the variational model about the reference trajectory

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

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

and the jacobian matrix

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

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

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

where

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

Exploiting the fact that $h(x, y) = 0$, calling $\delta x(t) \doteq x(t) - \bar{x}(t)$ and neglecting the arguments in C and D, we have, up to second order terms

$$h(\bar{x}(t), y_m(t)) = -C\delta x(t) - Dw(t).$$

[2] WARNING! do not confuse R and Q, variance of the measurement and model errors, with the rotation matrix R and the essential matrix $\mathbf{Q}$, as they appear throughout the paper.

Prediction Step

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

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

For small displacements we can write

$$\delta x(t+1) = F(\bar{x}(t))\delta x(t) + \tilde{v}(t) \tag{36}$$

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

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

$$\hat{x}(t+1|t) = \bar{x}(t+1) = f(\bar{x}(t)) = f(\hat{x}(t|t)). \tag{37}$$

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

$$P(t+1|t) = FP(t|t)F^T + \tilde{Q} \tag{38}$$

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.

Update Step

At time $t+1$ a new measurement becomes available $y_m(t+1)$, together with the prediction $\hat{x}(t+1|t)$ and its error variance $P(t+1|t)$. Exploiting the linearization of the measurement equation about $\bar{x}(t+1) = \hat{x}(t+1|t)$, we obtain, letting $\hat{x} \doteq \hat{x}(t+1|t)$ and $y_m \doteq y_m(t+1)$,

$$h(\hat{x}, y_m) = -C(\hat{x}, y_m)\delta x(t+1) - n(t+1) \tag{39}$$

where we have defined $n \doteq D(\hat{x}, y_m)w(t+1)$. This, together with the equation (36) defines a linear and *explicit* variational model, for which we can finally write the update equation based on the traditional linear Kalman filter:

$$\delta\hat{x}(t+1|t+1) = \delta\hat{x}(t+1|t) + L(t+1)\left[h(\hat{x}, y_m) + C\delta\hat{x}(t+1|t)\right] \tag{40}$$

where

$$\begin{aligned}
L(t+1) &= -P(t+1|t)C^T\Lambda^{-1}(t+1) \\
\Lambda(t+1) &= CP(t|t)C^T + R_n(t+1) \\
P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + LR_n(t+1)L^T \\
\Gamma &= (I - LC).
\end{aligned}$$

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

$$\hat{x}(t+1|t+1) = \hat{x}(t+1|t) + L(t+1)h\left(\hat{x}(t+1|t), y_m(t+1)\right). \tag{41}$$

In this formulation the quantity $h\left(\hat{x}(t+1|t), y_m(t+1)\right)$ plays the role of the pseudo-innovation. The noise n defined in (39) has a variance which is calculated from its definition:

$$R_n(t) = D(\hat{x}, y_m)R_w(t)D^T(\hat{x}, y_m).$$

B Projection onto the essential space

We have defined the projection operator onto the essential manifold without proving that the result is in fact an element of the essential manifold. In fact the following theorem, which was apparently first stated by Faugeras in 1990 [7, 17], shows that a characterizing property of the essential manifold is that its elements have two non-zero equal singular values and a zero singular value.

Theorem B.1 .
Let $Q = U\Sigma V^T$ be the SVD of an element of $GL(3)$. Then

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

Proof:

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

($\Leftarrow$) let $Q = U\Sigma_0 V^T$ for some orthonormal U, V and for some λ. Let furthermore $R_Z(\frac{\pi}{2})$ be a rotation of $\frac{\pi}{2}$ about the Z axis, then

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

Now call $R \doteq UR_Z(\frac{\pi}{2})^T V^T$ and $S \doteq VR_Z(\frac{\pi}{2})\Sigma_0 V^T$; it is immediate to see that $RR^T = R^TR = I_3$ and $S^T = -S$, hence the claim. **Q.E.D.**

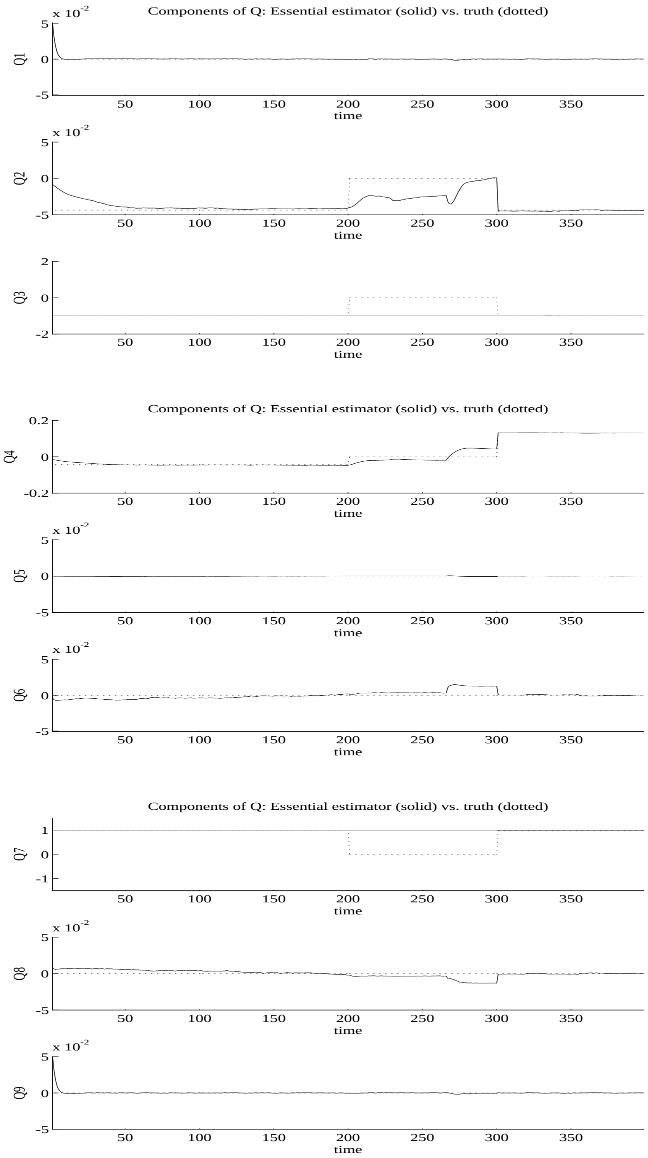

Figure 11: Components of the essential matrix as estimated by the essential estimator. Note that there are no spikes and the estimate is smooth. Note that the estimates between time 200 and 300 are not significant, as the ground truth (dotted line) is scaled to zero.

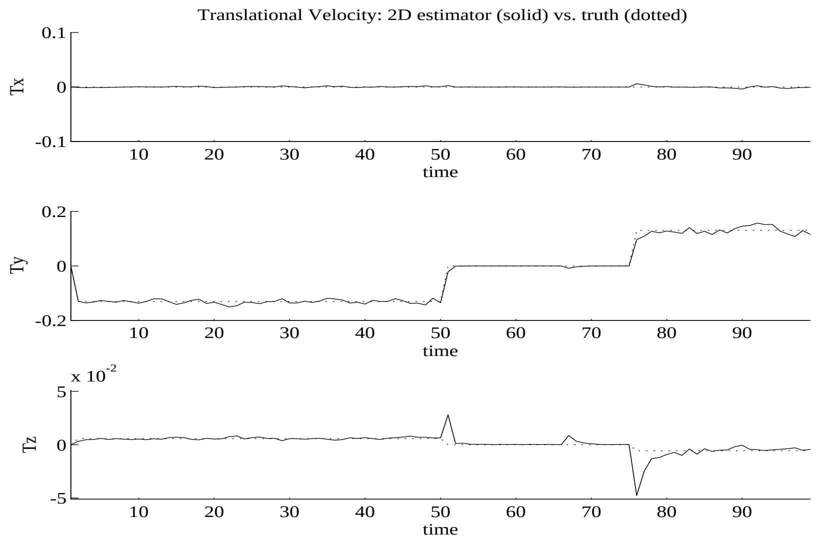

Figure 12: Components of translational velocity as estimated by the double iteration estimator.

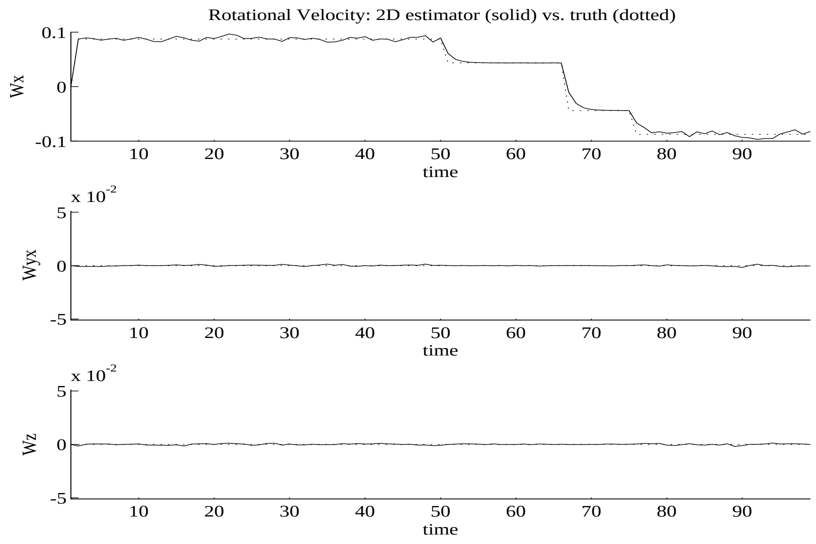

Figure 13: Components of rotational velocity as estimated by the double iteration estimator.

Figure 14: *One image of the rocket scene .*

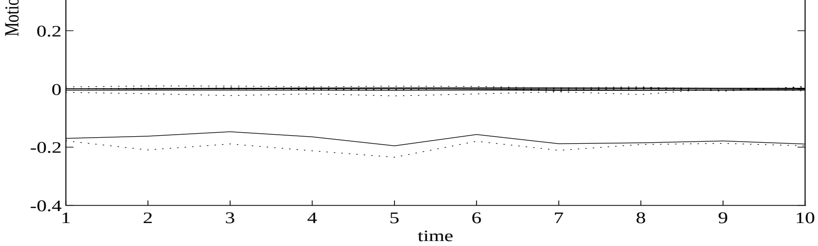

Figure 15: Motion estimates for the rocket sequence: The six components of motion as estimated by the local coordinates estimator are showed in solid lines. The corresponding ground truth is in dotted lines.

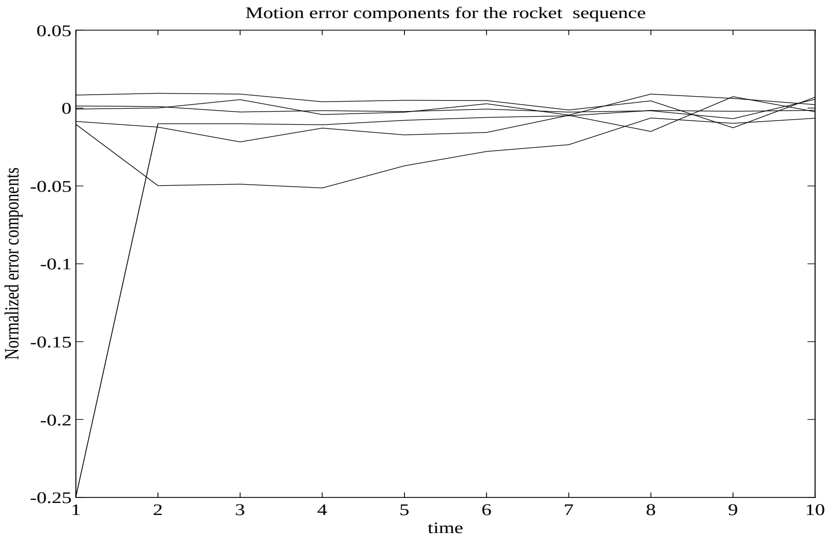

Figure 16: Error in the motion estimates for the rocket sequence. All components are within 5% of the true motion.

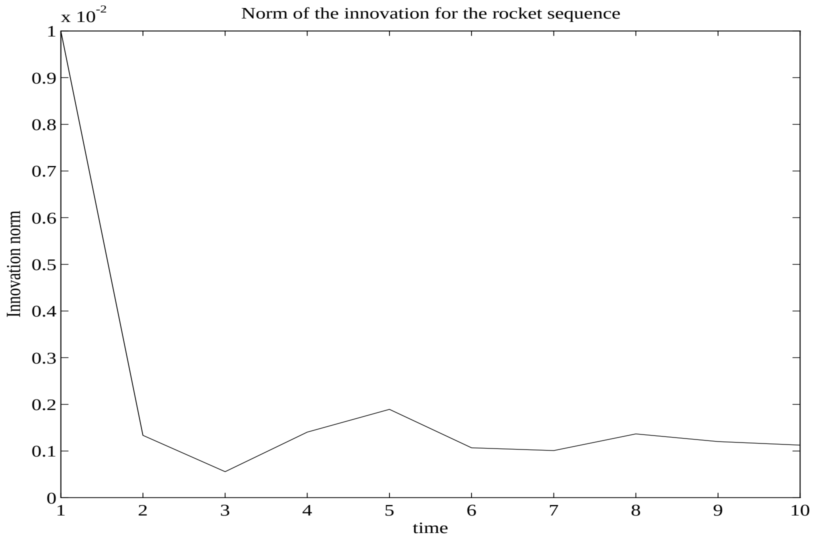

Figure 17: Norm of the pseudo-innovation process of the local estimator for the rocket scene. Convergence is reached in less than 5 steps.