Motion and Structure

Motion Estimation on the Essential Manifold*

Stefano Soatto[1], Ruggero Frezza[2], Pietro Perona[1,2]

[1] California Institute of Technology 116-81, Pasadena–CA 91125, soatto@caltech.edu
[2] Università di Padova, Dipartimento di Elettronica ed Informatica, Padova–Italy

Abstract. We introduce a novel perspective for viewing the "ego-motion reconstruction" problem as the estimation of the state of a dynamical system having an implicit measurement constraint and unknown inputs. Such a system happens to be "linear", but it is defined on a space (the "Essential Manifold") which is not a linear (vector) space.
We propose two recursive schemes for performing the estimation task: the first consists in *"flattening the space"* and solving a nonlinear estimation problem on the flat (euclidean) space. The second consists in viewing the system as embedded in a larger euclidean space, and solving at each step a *linear estimation* problem on a *linear* space, followed by a *"projection"* onto the Essential Manifold.
Both schemes output 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 [18, 22] in order to estimate 3D scene structure.
Experiments are presented with real and synthetic image sequences.

1 Introduction

A camera (or a human eye) is moving inside a static 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 of reconstructing the motion of the camera ("ego-motion") and the "structure" of the scene from its time-varying projection.
A simple representation of the "structure" of a scene is obtained from the position of a (finite) set of salient "feature" points in 3D space with respect to some reference frame, for example the one moving with the viewer. 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 in a cartesian frame, and we let $i = 1 \dots N$. As the camera moves between two discrete time instants, with rotation R and translation T, the coordinates change according to the rigid motion constraint:

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

* Research funded by the California Institute of Technology, ONR grant N00014-93-1-0990, an AT&T Foundation Special Purpose grant and the ASI-RS-103 grant from the Italian Space Agency. P.P. gratefully acknowledges the Newton Institute for Mathematical Sciences of Cambridge, UK, where he conducted part of this research.

Lecture Notes in Computer Science, Vol. 801
Jan-Olof Eklundh (Ed.)
Computer Vision – ECCV '94

where $T \in \mathbb{R}^3$ and $R \in SO(3)$ —the group of Special Orthogonal (rotation) matrices. We model the camera as an ideal perspective projection of the euclidean space onto the real-projective plane [2, 16] (pinhole camera):

$$\pi : \mathbb{R}^3 \to \mathbb{R}P^2 \qquad \mathbf{X} \mapsto \pi(\mathbf{X}) \doteq \begin{bmatrix} x & y & 1 \end{bmatrix}^T = \mathbf{x} \doteq \begin{bmatrix} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{bmatrix}^T. \quad (2)$$

In [20] we show how visual motion can be formulated as a combined "inversion-estimation" or "identification-estimation" task for the dynamical model (1)-(2). However, due to the "driftless" structure of the model, any inverse system is essentially *instantaneous*, and hence it does not exploit the benefits of recursiveness in terms of noise rejection and computational efficiency. Using the trick of "dynamic extension" [9] we show how the visual motion task can be transformed into the estimation of the state of a nonlinear system with unknown inputs, which in the estimation process are viewed as disturbances. A fundamental issue in deriving a state estimator (observer) is of course *observability*, which for linear systems is a necessary and sufficient condition for having an estimation error with spectrally assignable dynamics. For nonlinear systems the issue is more subtle [9, 12]; however, at least "local weak observability" is required in order to be able to state sufficient conditions for the existence of an observer with linear and spectrally assignable error dynamics. Other traditional state estimation techniques, such as the Extended Kalman Filter (EKF) [11, 10] are based upon the linearization of the model about the current trajectory.
The model which derives from (1)-(2) has the peculiarity of not only having a linearization which is not observable, but of also being non-"locally weakly observable". Hence, for the local linearization-based methods, it is not possible to derive sufficient conditions for convergence [19]. However, we show that, *once motion is estimated*, structure is linearly observable in the model (1)-(2), and therefore standard techniques, such as the EKF, can be used effectively for *structure estimation* [14, 18, 22]. Therefore the representation described by (1) and (2), though being the very simplest one can imagine, is not the most appropriate for motion estimation.

The recent literature proposes a variety of techniques for recovering structure and/or motion recursively [3, 14, 8, 7, 1, 18, 22], all of them based essentially on the same basic model (1)-(2), which in fact *defines* the visual motion problem for feature-points in the euclidean 3D space[3]. In particular, among those dealing with both structure and motion estimation, [1] is based on an extended model with motion added to the state space, the structure referred to the observer's reference at time 0 and a more general camera model. In [18] motion is recovered from 2 frames and fed to a model similar to (1)-(2), hence at each step motion is considered known and it does not exploit a dynamical model, as in [14]. In [22] motion is computed instantaneously as in [18], and then inserted it into the state dynamics with a model similar to the one used in [1].

[3] We have described a "viewer-centered" representation of the visual motion problem. "Object-centered" representations are essentially equivalent to the previous up to a diffeomorphism, therefore we will not make a distinction between the two.

This work is motivated by the fundamental limitations of the model (1)-(2), and presents a new dynamic model for motion estimation which is globally observable [4]. In section 2 we introduce and describe the Essential Manifold, in sections 3 and 4 we show how motion can be represented and estimated on the Essential Manifold. We introduce then the two approaches for performing the estimation task, which are unified within the new representation. In section 5 we address some special cases and possible generalizations. Finally in section 6 we show some experiments on real and synthetic image sequences.

2 The Essential Space

2.1 Rigid motion and the Essential Constraint

Suppose the correspondence of N feature points is given between time t and $t+1$, while the viewer has moved of (T, R). It is immediate to see (fig. 1 left) that the vector $\mathbf{X}$, describing the coordinates of the generic point at time t, the corresponding vector $\mathbf{X}'$ at time $t+1$, and T are coplanar, and therefore their triple product is zero [13]. This is also true with $\mathbf{x}$ in place of $\mathbf{X}$, since the two represent the same projective point. When expressed with respect to a common reference, for example that at time t, the coplanarity condition is written as $\mathbf{x}_i'^T R(T \wedge \mathbf{x}_i) = 0 \ \forall \ i \ = \ 1 \ldots N$. Once more than 8 correspondent points in general position are given [13, 15, 6, 16], the above constraint is also sufficient to characterize rigid motions up to a finite number of solutions. The operator $T\wedge$ belongs to $so(3)$ —the algebra of skew symmetric matrices; following the notation of Longuet-Higgins [13] we define $\mathbf{Q} \doteq RS \doteq RT\wedge$ so that the above coplanarity condition, which we call the "Essential Constraint", becomes

$$\mathbf{x}_i'^T \mathbf{Q} \mathbf{x}_i = 0 \ ; \ \forall i = 1 \ldots N. \tag{3}$$

Since the constraint is linear in $\mathbf{Q}$, it can be written as $\chi(\mathbf{x}'(t), \mathbf{x}(t))\mathbf{q}(t) = 0$; χ is an $N \times 9$ matrix whose generic row is $[x_1 x_1' \ x_2 x_1' \ x_1' \ x_1 x_2' \ x_2 x_2' \ x_2' \ x_1 \ x_2 \ 1\,]$, and $\mathbf{q}$ is a nine-vector obtained by stacking the columns of $\mathbf{Q}$. We will occasionally use the (improper) notation $\chi\mathbf{Q} \doteq \chi\mathbf{q}$, confusing $\mathbf{Q}$ and $\mathbf{q}$.

2.2 The Essential Manifold

We have seen that a rigid motion can be encoded using the Essential Constraint (3) based on the 3×3 matrix $\mathbf{Q} \doteq R(T\wedge) \subset \mathbb{R}^9$. Since we can reconstruct translation only up to a scale factor, we can restrict $\mathbf{Q}$ to belong to $\mathbb{R}P^8$ — the real projective space of dimension 8— or impose the norm of translation to be unitary. We will address later the case $T = 0$. The matrix $\mathbf{Q}$ belongs to the set $\tilde{E} \doteq \{RS | R \in SO(3), S \in so(3)\} \cap \mathbb{R}P^8$ which we call the *Essential*

[4] The maximal dimension of the observability codistribution of the basic model is reached after four levels of Lie-differentiation. Therefore in order to recover the observable components of the state-space it is necessary to perform a number of error-prone operations. The model that we will introduce has the advantage of being globally observable with only one level of differentiation [19].

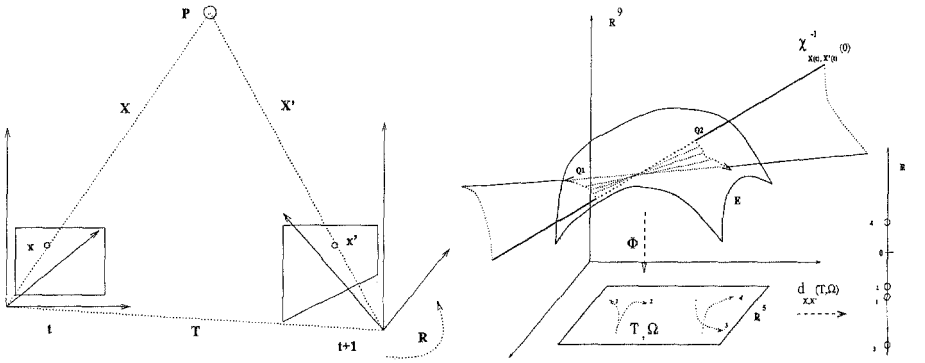

Fig. 1. *(Left) The Essential Constraint. (Right) Structure of the motion problem on the Essential Space*

Space; it encodes rigid motion in a more compact way than $SE(3)$ —the Special Euclidean group of rigid motions— the price being that we loose the smooth group structure. However, a slight modification of the Essential Space proves to have the structure of a topological manifold of class at least C_0 [20]. Consider the map

$$\Phi : E \to \mathbf{S}^2 \times \mathbb{R}^3 \sim \mathbb{R}^5 \qquad \mathbf{Q} \mapsto \begin{bmatrix} T \\ \Omega \end{bmatrix} = \begin{bmatrix} \pm V_3 \\ U R_Z(\pm\frac{\pi}{2}) V^T \end{bmatrix} \qquad (4)$$

where U, V are defined by the Singular Value Decomposition (SVD) of $\mathbf{Q} \doteq U\Sigma V^T$, 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 denote the local coordinates[5] of $\mathbf{Q}$; T is represented in spherical coordinates and Ω is the rotation 3-vector corresponding to the 3×3 rotation matrix $UR_Z(\frac{\pi}{2})V^T$ via the Rodrigues' formulae [17]. The map Φ defines the local coordinates of the Essential Manifold modulo a sign in the direction of translation and in the rotation angle of R_Z. This ambiguity can be resolved by imposing that the observed points are in front of the viewer [13]. Consider one of the four local counterparts of $\mathbf{Q} \in E$, and the triangulation map $d_{x,x'} : E \to \mathbb{R}^{1+1}$, $d_{x,x'}(\mathbf{Q}) = [Z, Z']^T$ which gives depth of each point as a function of its projections and the motion parameters. We redefine the Essential Space as $E \doteq \tilde{E} \cap d^{-1}_{x,x'}(\mathbb{R}^2_+)^N$, or

$$E \doteq \{RS | R \in SO(3), S \doteq T\wedge \in so(3), \|T\| = 1, d_{x^i,x'^i}(RS) > 0 \ \forall i = 1 \dots N\}.$$

Now it is easy to see that Φ, restricted to E, locally qualifies as a homeomorphism. The inverse map is simply $\Phi^{-1}(\Omega, T) = e^{(\Omega\wedge)}(T\wedge)$, which is smooth. E also has the structure of an algebraic variety [15], which we will not discuss in this paper.

[5] There is an abuse of notation: T indicates both the translation between two frames and the translation part of the canonical (screw) coordinates of motion. We allow such an ambiguity since the two are equivalent up to a diffeomorphism [17].

3 Motion representation on the Essential Manifold

We observe N points moving in space under some rigid motion through their noisy projections onto the image plane: $\mathbf{x}_i(t) + n_i(t)$; $i = 1 \dots N$. At each time instant we have a constraint in the form $\chi\mathbf{q}(t) \doteq -\tilde{n} \cong 0$, and hence $\mathbf{q}$ lies at the intersection between the Essential Manifold and the linear variety $\chi^{-1}(0)$ (see fig. 1 right). $\tilde{n}$ is a noise process which can be characterized in terms of the noise in the image-plane measurements n_i [20]. As time goes by, the point $\mathbf{Q}(t)$, corresponding to the actual motion, describes a trajectory on E satisfying

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

The last equation is in fact just a *definition* of the right-hand side, since we do not know $n_{\mathbf{Q}}(t)$. If we want to make use of such a model for estimating $\mathbf{Q}$ we have to make some assumptions. For now we will consider it as a discrete time dynamical model for $\mathbf{Q}$ on the Essential Manifold, having *unknown* inputs. If we accompany it with the Essential Constraint, we get

$$\begin{cases} \mathbf{Q}(t+1) = \mathbf{Q}(t) + n_{\mathbf{Q}}(t) \ ; \ \mathbf{Q} \in E \\ 0 = \chi\mathbf{Q}(t) + \tilde{n}(t). \end{cases} \qquad (*)$$

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 inputs. 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 appropriate in this context, since E is not a linear space.

4 Recursive estimation on the Essential Space

The first approach for performing the estimation task consists in composing $(*)$ with the local coordinates chart Φ defined in (4), ending up with a *nonlinear* dynamical model for motion in the *linear* space $\mathbb{R}^5$. At this point we have to make some assumptions about motion: if we do not have any dynamical model available, we may assume a statistical model. In particular we will assume that motion is a *first order random walk in* $\mathbb{R}^5$ *lifted to the Essential Manifold* (see fig. 2 left). The task is now to estimate the state of a nonlinear system driven by white, zero-mean gaussian noise. This will be done using a variation of the traditional EKF for systems with implicit measurement constraints, which we call the Implicit Extended Kalman Filter (IEKF).

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. 2 left). 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 $\mathbb{R}^9$ and then "projecting" the estimate onto the Essential Manifold (see fig. 2 right). The notion of projection onto the Essential Manifold will be made clear later.

It is very important to understand that these are *modeling assumptions* and can be validated only a posteriori. In general we observe that the first method

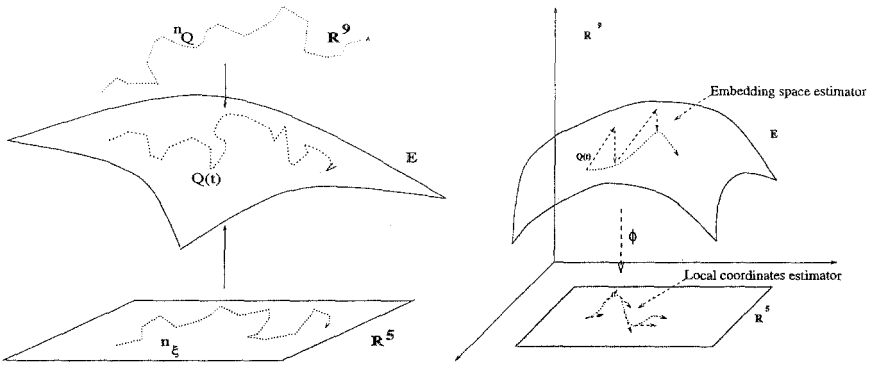

Fig. 2. *(Left) Model of motion as a random walk in $\mathbb{R}^5$ lifted to the manifold or as a random walk in $\mathbb{R}^9$ projected onto the manifold. (Right) Estimation on the Essential Space*

solves a strongly nonlinear problem with techniques based upon linearizing the system about the current reference trajectory. The update of the second method does not involve linearization, while it imposes the constraint of belonging to the Essential Manifold in a weaker way. The next two sections are devoted to describing these two techniques which produce, 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 [22].

4.1 Local coordinates estimator

Compose the model $(*)$ with the map Φ defined in (4). Call $\xi \doteq [T, \Omega]^T \in \mathbb{R}^5$ the local coordinates of $\mathbf{Q}$. Then the system becomes

$$\begin{cases} \xi(t+1) = \xi(t) + n_\xi(t) \ ; \ \xi(t_0) = \xi_0 \\ 0 = \chi \mathbf{Q}(\xi(t)) + \tilde{n}(t). \end{cases} \qquad (**)$$

We model motion as a first order random walk, i.e. $n_\xi(t) \in \mathcal{N}(0, R_{n_\xi})$ for some R_{n_ξ} which is referred to as variance of the model error. While the above assumption is arbitrary and can be validated only a posteriori, it is often safe to assume that the noise in the measurements $n_i(t)$ is a white zero-mean gaussian process. The second order statistics of $\tilde{n}$ can be inferred from n_i, as it has been done in [20]. Now $(**)$ is in a form suitable for using an IEKF. A derivation of the IEKF is reported in [20]: 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 call $C \doteq \left(\frac{\partial \chi \mathbf{Q}}{\partial \xi}\right)$ and $D \doteq \left(\frac{\partial \chi \mathbf{Q}}{\partial \mathbf{X}}\right)$, we have

Prediction step :

$$\begin{cases} \hat{\xi}(t+1|t) = \hat{\xi}(t|t) \ ; \ \hat{\xi}(0|0) = \xi_0 \\ P(t+1|t) = P(t|t) + R_{n_\xi} \ ; \ P(0|0) = P_0 \end{cases}$$

Update step :

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

Gain :

$$\begin{cases} L(t+1) = P(t+1|t)C^T(t+1)A^{-1}(t+1) \\ A(t+1) = C(t+1)P(t+1|t)C^T(t+1) + R_{\tilde{n}}(t+1) \\ \Gamma(t+1) = I - L(t+1)C(t+1) \end{cases}$$

Variance of $\tilde{n}$:

$$\left\{ R_{\tilde{n}}(t+1) = D(t+1)R_{\mathbf{x}}D^T(t+1) \right.$$

Note that $P(t|t)$ is the variance of the motion estimation error which is modeled as variance of measurement error by the subsequent modules of the motion and structure estimation scheme [22]. Also note that $\mathbf{Q}(\xi)$ is a strongly nonlinear function. This model was first introduced by Di Bernardo et al. [5] in a slightly different formulation. The Implicit Kalman Filter was used in the past by Darmon [4] and in later works.

4.2 The Essential Estimator in the embedding space

Suppose that motion, instead of being a random walk in $\mathbb{R}^5$, is represented on the Essential Manifold as the "projection" of a random walk through $\mathbb{R}^9$ (see fig. 2 left). The "projection" onto E is defined as follows:

$$\begin{aligned} pr_{<E>} : \mathbb{R}^{3\times 3} &\to E \\ M &\mapsto U\mathrm{diag}\{1,1,0\}V^T \end{aligned}$$

where $U, V \in \mathbb{R}^{3\times 3}$ are defined by the SVD of $M \doteq U\Sigma V^T$. The fact that this operator maps onto the Essential Manifold is a standard result [15] and is proved in [20]. Note that the projection minimizes the Frobenius norm and the 2-norm of the distance of a point in $\mathbb{R}^{3\times 3}$ from the Essential Manifold. Now define the operator $\oplus$ that takes two elements in $\mathbb{R}^{3\times 3}$, sums them and then projects the result onto the Essential Manifold:

$$\begin{aligned} \oplus : \mathbb{R}^{3\times 3} \times \mathbb{R}^{3\times 3} &\to E \\ M_1, M_2 &\mapsto \mathbf{Q} = pr_{<E>}(M_1 + M_2) \end{aligned}$$

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

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

where $n_{\mathbf{Q}}$ is modeled as a white zero-mean gaussian noise in $\mathbb{R}^9$ with variance $R_{n_{\mathbf{Q}}}$. If we couple the above equation with the lower part of $(*)$, we have again a dynamical model on an euclidean space driven by white gaussian noise. Note that the final model is precisely $(*)$ with $\oplus$ in place of $+$ and the constraint $\mathbf{Q} \in E$ released. The Essential Estimator is the least variance filter for such a model, and corresponds to a linear Kalman filter update in the embedding space, followed by a projection onto the Essential Manifold (see fig. 2 right). Note that in principle the gain could be precomputed offline, for each possible configuration of motion and feature positions.

Prediction step :

$$\begin{cases} \hat{\mathbf{q}}(t+1|t) = \hat{\mathbf{q}}(t|t) \ ; \ \hat{\mathbf{q}}(0|0) = \mathbf{q}_0 \\ P(t+1|t) = P(t|t) + R_{n_{\mathbf{Q}}} \ ; \ P(0|0) = P_0 \end{cases}$$

Update step :

$$\begin{cases} \hat{\mathbf{q}}(t+1|t+1) = \hat{\mathbf{q}}(t+1|t) \oplus L(t+1)\chi(t)\hat{\mathbf{q}}(t+1|t) \\ P(t+1|t+1) = \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + L(t+1)R_{\tilde{n}}(t+1)L^T(t+1) \end{cases}$$

Gain :

$$\begin{cases} L(t+1) = -P(t+1|t)\chi^T(t)\Lambda^{-1}(t+1) \\ \Lambda(t+1) = \chi(t)P(t+1|t)\chi^T(t) + R_{\tilde{n}}(t+1) \\ \Gamma(t+1) = I - L(t+1)\chi(t) \\ R_{\tilde{n}}(t+1) = D(t+1)R_{\mathbf{x}}D^T(t+1) \end{cases}$$

5 Special cases and generalizations

Singular case: what if we observe less than 8 points? – Suppose we are in the situation $N(t) < 8$ for some (possibly all) t. Then the Essential Constraint will have a preimage which is a whole subspace, and its intersection with the Essential Manifold (see fig. 1 right) will no longer be two points on E. However, suppose we move under constant (or "slowly varying") velocity; at each time instant we get 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 on the Essential Manifold, when the viewer does not move on a quadric containing all the visible points [19]. It is interesting to note that extended observations of *one only point* are sufficient to determine ego-motion.

Zero-translation case – The above schemes were described under the standing assumption of non-zero translation. When translation is zero there is no parallax, and we are not able to perceive depth. The Essential Constraint is undetermined, however we can still recover rotation and hence update the previous estimate of structure correctly. In fact, due to noise in the measurements of $\mathbf{x}_i, \mathbf{x}_i'$, 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 us to recover the correct rotation and scales depth by the inverse norm of the true translation. If we keep track of the scale factor, as described below, we can update the current estimate of structure and recover translation within the correct scale. This procedure has proved successful, as we show in the experimental section.

Recovery of the scale factor – The Essential filters recover translation only up to a scale factor. However, once some scale information is available *at one step* it can be propagated across time allowing recovery of motion and structure within the correct scale. This has been tested in the simulations by adding the scale factor in the filter dynamics with a random walk model.

On-line camera calibration – In introducing our algorithms we have described the camera as a simple static map from $\mathbb{R}^3$ to $\mathbb{R}P^2$. The model for the camera may be made more general [6, 16], time-varying, and inserted into the state dynamics with a statistical model, as we have done for motion. As long as the resulting model preserves observability properties, this will allow us to recover camera calibration together with relative orientation. Azarbayejani et al. [1] include the camera focal length in the standard formulation (1)-(2).

Segmentation and detection of outliers – The Essential models are peculiar in that they do not represent structure explicitly in the state, which allows varying the feature set at each time [21]. However, the innovation process of the filters is a measure of how far *each point* is from the current rigid-motion interpretation. At each time instant it is possible to compare each component of the innovation with the variance of the prediction at the previous time and reject all points that do not fall within a threshold. The Essential filters have proved useful in building a scheme for 3D transparent motion-based segmentation, which is reported in [21].

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 [22]. These consist of views of a cloud of points under a discontinuous motion with singular regions (zero-translation and non-zero rotation). 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. Tuning has been performed within an order of magnitude. See [20] for details.

The local coordinates estimator – In fig. 3 we show the three components of translational and rotational velocity as estimated by the local coordinates estimator. Convergence is reached in less than 20 steps. Initialization is performed with one step of the traditional Longuet-Higgins algorithm [13]. The computational cost of one iteration os of about 300 Kflops for 20 points. Note that if we have some dynamical model available for motion, we can easily insert it into the state model.

The Essential Estimator in the embedding space – When the estimated state is brought to local coordinates we have estimates for rotation and translation (see fig. 4). 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 step for 20 points. We report the mean of the estimation error of the two schemes, in order to show the absence of estimation biases, and the standard deviation to compare the performance. The results are summarized in the following table:

Scheme	T_X **(m, std)** 10^{-3}	T_Y	T_Z	Ω_X	Ω_Y	Ω_Z	Flops	Conv.
Local	(.2,.4)	(-1.5, 4.8)	(.2, .4)	(.8, 2.2)	(.2, .2)	(-.2, .8)	300K	15 steps
Embedding	(.0397, .1)	(1.7, 1.3)	(.2, .1)	(-.8, .4)	(.004, .2)	(-.0016, .4)	41K	50 steps

Experiments on real image sequences – We have tested our schemes on a sequence of 10 images of the rocket scene (see fig. 5). There are 22 feature points

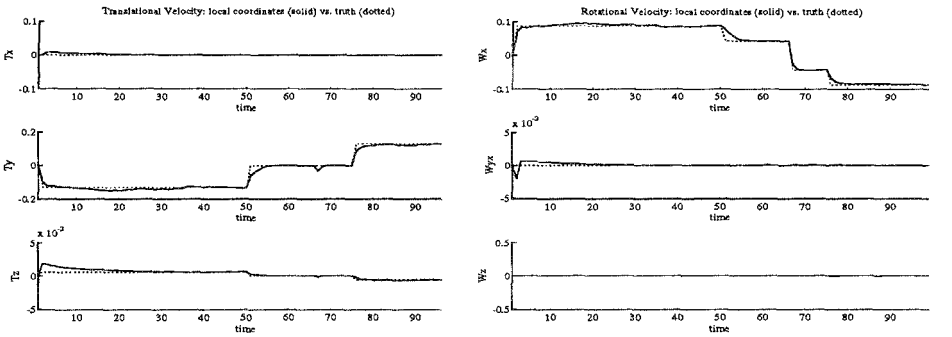

Fig. 3. *(Left) Components of translational velocity as estimated by the local coordinates estimator. The ground truth is shown in dotted lines. (Right) Rotational velocity.*

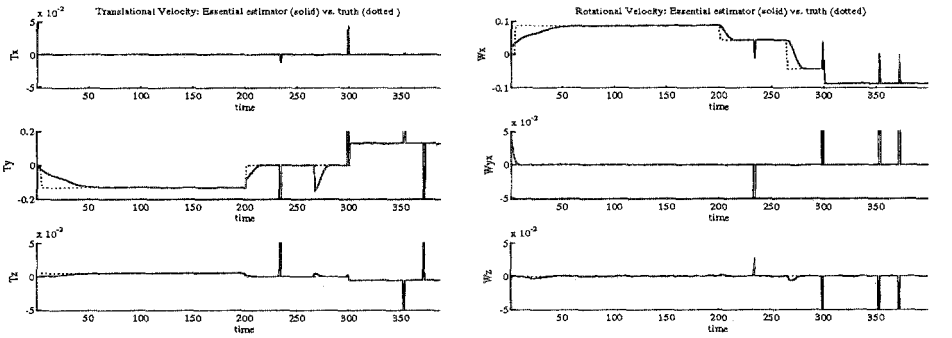

Fig. 4. *(Left) Components of translational velocity as estimated by the Essential estimator. Note the spikes due to the local coordinates transformation. Note also that they do not affect convergence since they do not occur in the estimation process, but while transferring to local coordinates. (Right) Rotational velocity.*

visible, and the standard deviation of the error on the image plane is about one pixel. The local coordinates estimator has a transient of about 20 steps to converge from any initial condition. Hence we have run it starting from zero, and used the final estimate as initial condition for a new run, the results of which are reported in figure 5. We did not perform any ad hoc tuning, and the setting was the same as that used in the simulation experiments. As it can be seen, the estimates are within 5% error, and the final estimate is less than 1% off the true motion. In this experiment we have used the true norm of translation as scaling factor. We have also run experiments in which the scale factor was calculated by updating the estimate of the distance between the two closest features, as in the simulation experiments. In this case convergence is slower, and the innovation norm reaches regime in about 20-25 steps (three runs over the sequence).

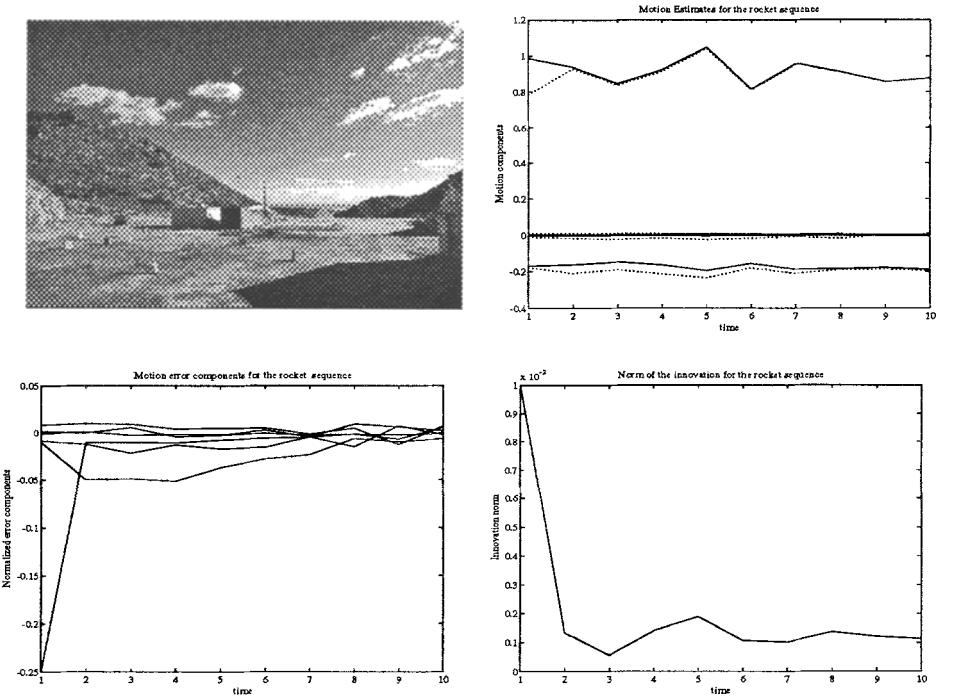

Fig. 5. *(Top-Left) One image of the rocket scene. (Top-Right) Motion estimates for the rocket sequence: The six components of motion as estimated by the local coordinates estimator are showed in solid lines. The corresponding ground truth is in dotted lines. (Bottom-Left) Error in the motion estimates for the rocket sequence. All components are within 5% of the true motion. (Bottom-Right) Norm of the pseudo-innovation process of the local estimator for the rocket scene. Convergence is reached in less than 10+5 steps.*

7 Conclusions

We have presented a novel perspective for viewing motion estimation. This has resulted in two different approaches for solving the motion problem which are cast in a common framework. Each scheme has its own personality, the filter in the embedding space being faster and more geometrically appealing, the local coordinates estimator being more flexible and robust. The schemes are based on a globally observable model and enjoy common features such as recursiveness, allowing us to exploit at each time all previous calculations, and noise rejection from exploiting redundancy. They all benefit from independence from structure estimation, which allows us to deal easily with a variable number of points and feature sets. Hence we do not need to track specific features through time, and we can deal easily with occlusion and presence of outliers.

Both schemes produce, together with an estimate of motion, the second order statistics of the estimation error.

The approaches can be interpreted as an extension of the Longuet-Higgins' algorithm [13] to infinite baseline, and the observability analysis as a generaliza-

tion of N-points M-frames theorems. The schemes work for any number of points provided that enough frames are viewed. Possible extensions include on-line estimation of the camera model.

Acknowledgements

We wish to thank Prof. Giorgio Picci for his constant support and advice, Prof. J.K. Åström for discussions on implicit Kalman filtering, Prof. Richard Murray, Prof. Shankar Sastry and Andrea Mennucci for many useful suggestions. We also thank John Oliensis and J. Inigo Thomas for providing the rocket sequence.

References

1. Azarbayejani, A., Horowitz, B., and Pentland, A. Recursive estimation of structure and motion using relative orientation constraints. *Proc. CVPR* (New York, 1993).
2. Boothby, W. *Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press, 1986.
3. Broida, T., and Chellappa, R. Estimating the kinematics and structure of a rigid object from a sequence of monocular frames. *IEEE Trans. Pattern Anal. Mach. Intell.* (1991).
4. Darmon. A recursive method to apply the hough transform to a set of moving objects. *Proc. IEEE, CH 1746 7/82* (1982).
5. Di-Bernardo, E., Toniutti, L., Frezza, R., and Picci, G. Stima del moto dell'osservatore e della struttura della scena mediante visione monoculare. *Tesi di Laurea–Università di Padova* (1993).
6. Faugeras, O. *Three dimensional vision, a geometric viewpoint.* MIT Press, 1993.
7. Gennery, D. Tracking known 3-dimensional object. In *Proc. AAAI 2nd Natl. Conf. Artif. Intell.* (Pittsburg, PA, 1982), pp. 13–17.
8. Heel, J. Direct estimation of structure and motion from multiple frames. *AI Memo 1190, MIT AI Lab* (March 1990).
9. Isidori, A. *Nonlinear Control Systems.* Springer Verlag, 1989.
10. Jazwinski, A. *Stochastic Processes and Filtering Theory.* Academic Press, 1970.
11. Kalman, R. A new approach to linear filtering and prediction problems. *Trans. of the ASME-Journal of basic engineering. 35-45* (1960).
12. Krener, A. J., and Respondek, W. Nonlinear observers with linearizable error dynamics. *SIAM J. Control Optim. vol. 23 (2)* (1985).
13. Longuet-Higgins, H. C. A computer algorithm for reconstructing a scene from two projections. *Nature 293* (1981), 133–135.
14. Matthies, L., Szelisky, R., and Kanade, T. Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. of computer vision* (1989).
15. Maybank, S. *Theory of reconstruction from image motion*, vol. 28 of *Information Sciences.* Springer-Verlag, 1992.
16. Mundy, J., and Zisserman, A., Eds. *Geometric invariance in computer vision.* MIT Press, Cambridge, Mass., 1992.
17. Murray, R., Li, Z., and Sastry, S. *A Mathematical Introduction to Robotic Manipulation.* Preprint, 1993.
18. Oliensis, J., and Inigo-Thomas, J. Recursive multi-frame structure from motion incorporating motion error. *Proc. DARPA Image Understanding Workshop* (1992).
19. Soatto, S. Observability of rigid motion under perspective projection with application to visual motion estimation. *Technical Report CIT-CDS 94-001, California Institute of Technology* (1994).
20. Soatto, S., Frezza, R., and Perona, P. Recursive motion estimation on the essential manifold. *Technical Report CIT-CDS 93-021 and CIT-CNS 32/93, California Institute of Technology* (1993).
21. Soatto, S., and Perona, P. Three dimensional transparent structure segmentation and multiple 3d motion estimation from monocular perspective image sequences. *Technical Report CIT-CDS 93-022, California Institute of Technology* (1993).
22. Soatto, S., Perona, P., Frezza, R., and Picci, G. Recursive motion and structure estimation with complete error characterization. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.* (New York, June 1993), pp. 428–433.