Recursive 3D Visual Motion Estimation Using Subspace Constraints *

Stefano Soatto† and *Pietro Perona*‡

† *Control and Dynamical Systems*
California Institute of Technology 116-81, Pasadena – CA 91125
‡ *Electrical Engineering* and *Computation and Neural Systems*
California Institute of Technology 116-81, Pasadena – CA 91125
‡*Dipartimento di Elettronica ed Informatica*, Università di Padova, Padova –Italy
{soatto,perona}@systems.caltech.edu

Keywords: Dynamic vision, rigid motion estimation, nonlinear implicit identification.

Abstract

The problem of estimating rigid motion from projections may be characterized using a nonlinear dynamical system, composed of the rigid motion constraint and the perspective map. The time derivative of the output of such a system, which is called the "motion field" and approximated by the "optical flow", is bilinear in the motion parameters, and may be used to specify a subspace constraint on either the direction of translation or the inverse depth of the observed points. Estimating motion may then be formulated as an optimization task constrained on such a subspace. Heeger and Jepson [9], who first introduced this constraint, solve the optimization task using an extensive search over all possible directions of translation.

We pose the optimization problem in a system-theoretic framework as the the identification of a nonlinear implicit dynamical system with parameters on a differentiable manifold, and use techniques which pertain to nonlinear estimation and identification theory to perform the optimization task in a principled manner. The application of a general method presented in [20] results in a recursive and pseudo-optimal solution of the motion problem, which has robustness properties far superior to other existing techniques we have implemented.

Experiments on real and synthetic image sequences show very interesting results in terms of robustness and accuracy. By releasing the constraint that the visible points lie in front of the observer, we may explain some psychophysical effects on the nonrigid percept of rigidly moving shapes.

1 Introduction

"Visual motion estimation" is one of the oldest [8, 25] and at the same time one of the most crucial and challenging problems in computer vision. In this paper we study the recursive estimation of the three-dimensional motion of a scene viewed from a sequence of monocular perspective images.

*This work is registered as CDS Technical Report **CIT-CDS** 94-005, California Institute of Technology, January 1994 – revised February 1994. Reduced version in "Proc. of the 1^{st} IEEE Conf. on Image Processing", Austin–Texas Oct. 1994.

A number of schemes exist for estimating recursively structure for known motion [17, 18], motion for known structure [2, 6, 7] or both structure and motion simultaneously (for example [1, 12]).

We argue against simultaneous structure and motion estimation for two reasons: (a) complexity – including the structure of the scene into the state of the filter makes it computationally demanding and requires sophisticated heuristics for dealing with occlusions and appearance of new features; (b) convergence problems – the schemes proposed so far have poor model observability (see [19] for a thorough discussion of this issue). The recursive estimation of motion is a relatively unexplored subject: to our knowledge, the only recursive 3D motion estimation scheme that is independent on the structure of the scene (given some general position assumptions) and makes full use of the epipolar geometry is the "essential filter" [21].

We present here a novel recursive motion estimator, which we call the "subspace filter", that is based upon the differential version of the epipolar constraint introduced by Longuet-Higgins [16] along the lines proposed by Heeger and Jepson [9] in their two-frames algorithm.

2 Visual motion estimation from a dynamic model

Let a scene be represented by a set of N feature points in 3D space moving rigidly with respect to the viewer; the "visual motion estimation" problem is *defined* by the rigidity constraint and the perspective projection equations. If $\mathbf{X}_i \doteq [X_i \; Y_i \; Z_i]^T$ are the coordinates of the i^{th} point in the viewer's reference frame and $\mathbf{x}_i \doteq [x_i \; y_i]^T = [\frac{X_i}{Z_i} \; \frac{Y_i}{Z_i}]^T$ the corresponding projections, we may write

$$\begin{cases} \dot{\mathbf{X}}_i = \Omega \wedge \mathbf{X}_i + V & \mathbf{X}(0) = \mathbf{X}_0 \\ \mathbf{x}_i = \pi(\mathbf{X}_i) + n_i & \forall i = 1 : N \end{cases} \tag{1}$$

where n_i represents an error in measuring the position of the projection of the point i, and π represents an ideal perspective projection. Solving the visual motion estimation problem consists of reconstructing the ego-motion V, Ω from all the visible points, i.e. estimating the input of the above system from its noisy output. We show that it is possible to invert the above system using a technique which has been recently introduced in [20] for identifying nonlinear implicit systems with parameters on a topological manifold.

Our scheme is motivated by the work of Heeger and Jepson [9, 10], who formulate the task as an optimization problem and then solve it by extensive search over all possible directions of translational velocity. This procedure does not exploit the geometric structure of the problem and is computationally expensive. Furthermore, it does not take into account the measurement noise, which enters into the minimization in a highly structured fashion. Temporal coherence of motion is also not taken into account; on the contrary, we want to exploit all the processing performed at the previous time instants and update the motion estimates recursively and causally.

The scheme we present may be considered as a recursive solution to the task of Heeger and Jepson, using methods which pertain to the field of nonlinear estimation and identification theory. As a result, the minimization task which is the core of the subspace method for recovering rigid motion needs not to be performed by extensive search. Instead, an Implicit Extended Kalman Filter (IEKF) [4, 13, 14, 20] is in charge of estimating the motion parameters recursively according to nonlinear prediction error criteria (for an introductory treatment of Prediction Error Methods (PEM) in a linear context, see for example [24]). The method exploits in a pseudo-optimal manner the information coming from a long stream of images, making the scheme robust and efficient.

3 Motion reconstruction via (least squares) inversion constrained on subspaces

3.1 Recovery of the direction of translation

Consider the following expression of the first derivative of the output of the model (1), which is referred to in the literature as the "motion field" (or "real optical flow") and is approximated for computation purposes by the ("apparent") optical flow:

$$\dot{\mathbf{x}}_i(t) = \left[\frac{1}{Z_i}\mathcal{A}_i \mid \mathcal{B}_i\right] \begin{bmatrix} V(t) \\ \Omega(t) \end{bmatrix} \tag{2}$$

where

$$\mathcal{A}_i \doteq \begin{bmatrix} 1 & 0 & -x_i \\ 0 & 1 & -y_i \end{bmatrix} \qquad \mathcal{B}_i \doteq \begin{bmatrix} -x_i y_i & 1+x_i^2 & -y_i \\ -1-y_i^2 & x_i y_i & x_i \end{bmatrix}. \tag{3}$$

If we observe a sufficient number of points $\mathbf{x}_i \ \forall i = 1 \dots N$, we can write an overdetermined system which can be solved for the inverse depth and the rotational velocity in a least-squares fashion. To this end, we rearrange the previous equation as

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

Since the translational velocity V multiplies the inverse depth of each point, both can be recovered only up to an arbitrary scale factor. In order to get rid of this scale ambiguity, we choose V to be of unit norm, and represent it in local (spherical) coordinates as $V(\theta,\phi) \in \mathbf{S}^2$. If some scale information becomes available, as for example the size of a visible object, it is possible to rescale the depth and the translational velocity, as we will discuss in the experimental section. When we observe N points we can rearrange the equations above into a vector equality:

$$\dot{\mathbf{x}} = \tilde{\mathcal{C}}(\theta,\phi)[\frac{1}{Z_1}, \dots, \frac{1}{Z_N}, \Omega]^T, \tag{4}$$

where

$$\tilde{\mathcal{C}}(\theta,\phi) \doteq \begin{bmatrix} \mathcal{A}_1 V & & & \mathcal{B}_1 \\ & \ddots & & \vdots \\ & & \mathcal{A}_N V & \mathcal{B}_N \end{bmatrix}$$

and $\mathbf{x}$ is a $2N$ column vector obtained by stacking the $\mathbf{x}_i \ \forall i = 1 \dots N$ on top of each other. At this point we could solve the above equation (4) in a least squares fashion for the inverse depth and rotation:

$$\begin{bmatrix} \frac{1}{\hat{Z}_1} \\ \vdots \\ \frac{1}{\hat{Z}_N} \\ \hat{\Omega} \end{bmatrix} = \tilde{\mathcal{C}}^{\dagger}\dot{\mathbf{x}}$$

where the symbol $\dagger$ denotes the pseudo-inverse. We can then plug the result into equation (4),

$$\dot{\mathbf{x}} = \tilde{\mathcal{C}}(\theta,\phi)\tilde{\mathcal{C}}^{\dagger}\dot{\mathbf{x}},$$

ending up with an *implicit constraint* on the direction of translation θ, ϕ. By rearranging the terms and writing explicitly the pseudo-inverse we get the following subspace algebraic constraint [9]:

$$\left[I - \tilde{\mathcal{C}} \left(\tilde{\mathcal{C}}^T \tilde{\mathcal{C}} \right)^{-1} \tilde{\mathcal{C}}^T \right] \dot{\mathbf{x}} \doteq \tilde{\mathcal{C}}^{\perp} \dot{\mathbf{x}} = 0. \tag{5}$$

Note that, if $U \Sigma V^T$ denotes the Singular Value Decomposition (SVD) of $\tilde{\mathcal{C}}$, then we have $\tilde{\mathcal{C}}^{\perp} = (I - UU^T)$. We can now try to approximate this constraint by solving the following nonlinear optimization problem:

$$\hat{V} = \arg \min_{V \in \mathbf{S}^2} \| \tilde{\mathcal{C}}^{\perp} \dot{\mathbf{x}} \|. \tag{6}$$

In other words we are looking for the best vector in the two-dimensional sphere such that $\dot{\mathbf{x}}$ is the null space of the orthogonal complement of the range of $\tilde{\mathcal{C}}$. If the matrix $\tilde{\mathcal{C}}$ was invertible, the above constraint would be satisfied trivially for all directions of translation. However, when $2N > N + 3$, $\tilde{\mathcal{C}} \tilde{\mathcal{C}}^{\dagger}$ has rank at most $N + 3$, and therefore $\tilde{\mathcal{C}}^{\perp}$ is not identically zero.

Note that we are trying to "adapt" the orthogonal complement of $\tilde{\mathcal{C}}$, which is highly structured as a function of θ, ϕ, until a given vector $\dot{\mathbf{x}}$ is its null space. Heeger and Jepson [9] solve this problem by minimizing the two-norm of the above constraint using an extensive search over θ, ϕ, or a sampling of the sphere. This procedure does not exploit the geometric structure of the problem and does not take into account the measurement noise, which enters into the minimization in a highly structured fashion. Temporal coherence of motion is also not taken into account: at each step we want to exploit all the processing performed at the previous time instants and update the motion estimates recursively and causally.

In section 4 we rephrase the subspace constraints described in this section as a nonlinear and implicit dynamic model in exterior differential form [3]. Estimating motion corresponds to the identification of such a model with the parameters living on a sphere: we propose a principled solution for performing the optimization task. The method outputs motion estimates together with their reliability in the form of the second order statistics of the estimation error.

3.2 Recovery of rotation and depth

Once the direction of translation has been estimated, we may compute a least-squares estimate of the rotational velocity and inverse depth from

$$\begin{bmatrix} \frac{1}{Z_1} \\ \vdots \\ \frac{1}{Z_N} \\ \Omega \end{bmatrix} = \tilde{\mathcal{C}}^{\dagger}(\hat{\theta}, \hat{\phi}) \dot{\mathbf{x}}.$$

Note that from the variance/covariance of the estimation error of the direction of translation θ, ϕ, we can characterize the second order statistics of the estimate of the rotational velocity. We may therefore design a simple linear Kalman filter which uses the above estimates as "pseudo-measurements" and is based upon the linear model

$$\begin{cases} \Omega(t+1) = \Omega(t) + n_{rw} \\ \tilde{\mathcal{C}}^{\dagger}_{2N+1:2N+3}(\theta, \phi) \dot{\mathbf{x}} = \Omega(t) + n_{\Omega} \end{cases}$$

where the notation $\tilde{\mathcal{C}}^{\dagger}_{2N+1:2N+3}$ stands for the rows from $2N + 1$ to $2N + 3$ of the pseudoinverse of the matrix $\tilde{\mathcal{C}}$; n_{rw} is the noise driving the random walk model, which is to be intended as a tuning parameter, and n_{Ω} is an error whose variance is inferred from the variance of θ, ϕ.

The equations for the Kalman filter corresponding to the above (linear) model are standard, and can be found in textbooks, see for example [13]. Note that the filter that estimates the direction of translation is independent on the rotational velocity, that can therefore be estimated offline.

3.3 Recovery of structure

After the rotational and translational velocities have been recovered as described above, they may be fed, together with the variance of their estimation error, into a recursive structure from motion module which processes motion error, such as for example [18, 23]. The main focus of this paper is the estimation of motion, and in the experimental section we have estimated structure using the scheme presented in [23]. However, we point out in this section an alternative way of estimating structure, that comes from noting that the inverse depth of each point and the direction of translation play interchangeable roles, as it is evident from the motion field (2). We may therefore "pseudo-invert" the system (2) with respect to the direction of translation and the rotational velocity, and then perform a minimization similar to (6) with respect to the inverse depth of each point. Call $\mathcal{C}_i \doteq \left[\frac{1}{Z_i}\mathcal{A}_i \mid \mathcal{B}_i\right]$, we have

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \\ \vdots \\ \dot{\mathbf{x}}_N \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathcal{C}_i \\ \vdots \end{bmatrix} \begin{bmatrix} V(t) \\ \Omega(t) \end{bmatrix} \Rightarrow \begin{bmatrix} \hat{V}(t) \\ \hat{\Omega}(t) \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathcal{C}_i \\ \vdots \end{bmatrix}^{\dagger} \begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \\ \vdots \\ \dot{\mathbf{x}}_N \end{bmatrix} \doteq \mathcal{C}^{\dagger}\dot{\mathbf{x}}$$

where $\dagger$ denotes the pseudo-inverse. Note that $\mathcal{C}_i$ depends on the depth of the point Z_i, which we do not know. By substituting the above expression into eq. (2), we have an *implicit constraint* on $Z_i \ \forall \ i$ [9]:

$$\dot{\mathbf{x}} = \mathcal{C} \begin{bmatrix} \hat{V}(t) \\ \hat{\Omega}(t) \end{bmatrix} = \mathcal{C}\mathcal{C}^{\dagger}\dot{\mathbf{x}} \ \Rightarrow \ (I - \mathcal{C}\mathcal{C}^{\dagger})\dot{\mathbf{x}} \doteq \mathcal{C}^{\perp}\dot{\mathbf{x}} = 0.$$

We now approximate this constraint by solving w.r.t. Z_i the following optimization problem:

$$\hat{Z}_i = \arg\min_{Z_i} \|\mathcal{C}^{\perp}\dot{\mathbf{x}}\|. \tag{7}$$

If $\mathcal{C}$ was invertible, again the above constraint would be satisfied trivially for all motions. However, in general $(I - \mathcal{C}\mathcal{C}^{\dagger}) \neq 0$.

In many applications it is of interest to estimate the average distance of an object from the camera (depth of the centroid), in order to maintain the information about the scaling of a scene. For this case, it is sufficient to consider the minimization in eq. (7) when $Z_i = Z_c \ \forall i$; Z_c is the distance of the centroid.

4 Solving the subspace optimization

Let us define $\alpha \doteq [\theta, \phi]^T$; $\mathbf{x}_i$ are measured up to some error, $\mathbf{y}_i \doteq \mathbf{x}_i + n_i$, which is by no means white, Gaussian and zero-mean. However, these hypotheses are often satisfied within reasonable margins, so we will assume, as customary, $n_i \in \mathcal{N}(0, R_{n_i})$. The error in the location of the features induces an error in the derivative, $\mathbf{y}'_i = \dot{\mathbf{x}}_i + n'_i$, which is usually approximated by either the optical flow, or by first differences of feature positions between time t and $t+1$. Call $\mathbf{x}$ the column vector obtained by stacking the components of $\mathbf{x}_i$, similarly with $\dot{\mathbf{x}}$. Now define $\tilde{\mathcal{C}}^{\perp}(\mathbf{x}, \alpha) \doteq (I - UU^T)$, where

$U\Sigma V^T$ is the SVD of $\tilde{\mathcal{C}}(\mathbf{x},\alpha)$. Then the subspace constraint (5) may be written as $\tilde{\mathcal{C}}^{\perp}(\mathbf{x},\alpha)\dot{\mathbf{x}} = 0$. Now

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

represents a nonlinear implicit dynamical system of a particular class, called Exterior Differential Systems [3]. *Solving for the translational velocity is equivalent to identifying the above exterior differential system with parameters α on a differentiable manifold* (the sphere in this case) from the noisy data $\mathbf{y}$.

We have addressed this problem using a general method presented in [20], which is based upon an Implicit Extended Kalman Filter (IEKF) having the unknown parameters (the local coordinates α of the direction of translation in our case) as its state. The solution is given by the usual Kalman iteration with the vector $\epsilon(t) \doteq \tilde{\mathcal{C}}^{\perp}(\mathbf{y}(t), \hat{\alpha}(t+1|t))\mathbf{y}'$ playing the role of the "pseudo-innovation" process [13, 14]:

Prediction step

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

Update step

$$\begin{cases} \hat{\alpha}(t+1|t+1) = \hat{\alpha}(t+1|t) + \\ \quad + L(t+1)\tilde{\mathcal{C}}^{\perp}(\mathbf{y}(t), \hat{\alpha}(t+1|t))\mathbf{y}' \\ P(t+1|t+1) = \\ \quad = \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + \\ \quad + L(t+1)D_+(t)R_n(t+1)D_+^T(t)L^T(t+1) \end{cases}$$

where

$$\begin{cases} L(t+1) = P(t+1|t)C^T(t+1)\Lambda^{\dagger}(t+1) \\ \Lambda(t+1) = C(t+1)P(t+1|t)C^T(t+1) + D_+(t+1)R_n(t+1)D_+^T(t+1) \\ \Gamma(t+1) = I - L(t+1)C(t+1) \\ D_+(t+1) \doteq \left(\frac{\partial \tilde{\mathcal{C}}^{\perp}\dot{\mathbf{x}}}{\partial \mathbf{x}(t)}\right)_{|\mathbf{y}(t),\hat{\alpha}(t)} \\ C(t+1) \doteq \left(\frac{\partial \tilde{\mathcal{C}}^{\perp}\dot{\mathbf{x}}}{\partial \alpha(t)}\right)_{|\mathbf{y}(t),\hat{\alpha}(t)} \end{cases}$$

and R_n is the variance/covariance matrix of the measurement errors, while R_α is the noise driving the random walk model of the motion parameters, and is to be intended as a tuning parameter. A complete derivation of the scheme may be found in the appendix A of [20][1].

In order to be able to assess the convergence of the above scheme, one must prove its observability/identifiability. When translated into the language of dynamic estimation, the analysis of Heeger and Jepson [11] can be intended as the observability analysis of our method; in particular it shows that the scheme converges under general position conditions.

Enforcing rigid motion: the positive depth constraint

When estimating motion from visible points, we must enforce the fact that the measured points are *in front of the observer*. This may be easily done in the prediction step by computing the mean distance of the centroid and checking whether it is positive. If it is not, the prediction is reflected on the sphere (the antipodal point of the state space sphere is chosen as the prediction).

[1] This paper can be obtained via the Worldwide Web `Mosaic` (http://avalon.caltech.edu/cds/techreports/)

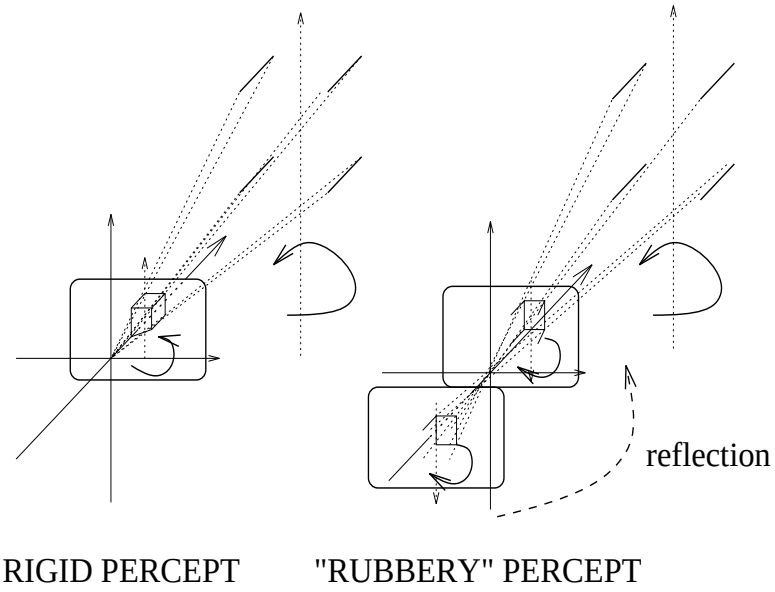

Figure 1: *Geometric interpretation of the "rubbery" perception: motion is estimated without imposing the positive depth constraint; this may result in a motion estimate which is compatible with a rigid structure interpretation behind the observer. Once such a structure is reflected in front of the observer, it gives rise to the perception of a rubbery structure rotating in the opposite direction.*

When we do not impose such a constraint, the filter may converge to a rigid motion which corresponds to points moving behind the observer, and is therefore not physically realizable. However, if we allow such a condition to happen by releasing the positive depth constraint, and then feed the estimate into a structure estimation, such as for example a simple Extended Kalman Filter [17, 18, 23] initialized with points at positive depth and a large model-error variance, the result is a *rubbery interpretation of structure* which has been observed also in psychophysical experiments [15].

The geometric interpretation of the rubbery percept is illustrated in figure 1. Note that both affine 3D motion and similarity transformations viewed under projection admit a geometric invariant, which is the absolute conic [5]. On the contrary, the orientation of a rigid motion is not invariant under projection.

Independence on structure estimation

It is worth noting that the state of the filter contains only the local coordinates of the direction of translation, and is therefore independent on the structure of the observed scene. In particular, we do not need to track a specific set of features; instead, we can at each step change set of features or locations where we compute the optical flow. This is a key property of the filter, since it allows us to deal easily with occlusion and appearance of new features.

Also note that the filter is able to work properly even when the number of visible features drops down to *one* (for slowly-varying velocity), since it integrates over time the information from each incoming frame. This, together with the robustness and noise-rejection properties, is a substantial advantage over two-views schemes, that need a minimum number of visible points in order to estimate ego-motion.

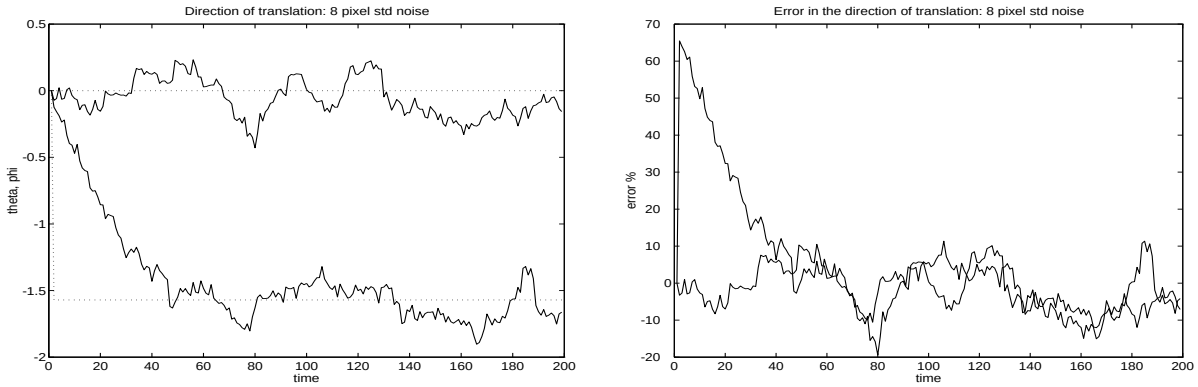

Figure 2: *(Left) Estimates of the two components of the direction of translation. The noise in the image plane measurements had 8 pixel standard deviation. The initial conditions were zero for both components. The ground truth is in dotted lines. (Right) Estimation error for the direction of translation. With noise of 8 pixel std in the data, the estimates are still within 20 % of the true value. The positive depth constraint has been enforced.*

5 Experimental assessment

We have experimented with the scheme on real and noisy synthetic image sequences. In this section we report a set of simulations and two experiments on real sequences.

5.1 Simulation setup

We have generated at random a set of 20 points in space, distributed uniformly in a cubic volume of side 1 m, with the centroid placed 1.5 m ahead of the image plane. The points are then projected onto an image plane of 512×512 pixels, and Gaussian noise is added to the position of the projected points with standard deviations ranging from 1 to 8 pixels. The cloud of points is moved in space with piecewise constant velocity, and is viewed under an angle of approximately 45^o.

5.2 Simulated experiment

We have considered the synthetic cloud of dots described above, rotating about its centroid with a velocity of circa 5^o/step, and added to the projections white, zero-mean Gaussian noise. The motion is rototraslatory in the viewer's reference frame, and is challenging since the effect of rotation and translation superimpose.

Convergence is reached from *arbitrary initial conditions* and noise in the image plane coordinates up to 8 pixel std (see figure 2). An estimate for more usual noise levels (1 pixel std) is reported in figure 3. In both cases the positive depth constraint has been enforced. The transient for converging from zero initial conditions ranges from 5 to 40 steps, depending on the noise level, the type of motion and the tuning of the filter.

The least-squares pseudo-measurements of the rotational velocity, computed as described in section 3.2, are plotted in figure 4 (dashed lines), and compared with the recursive estimates (solid line) using the linear Kalman Filter described in section 3.2.

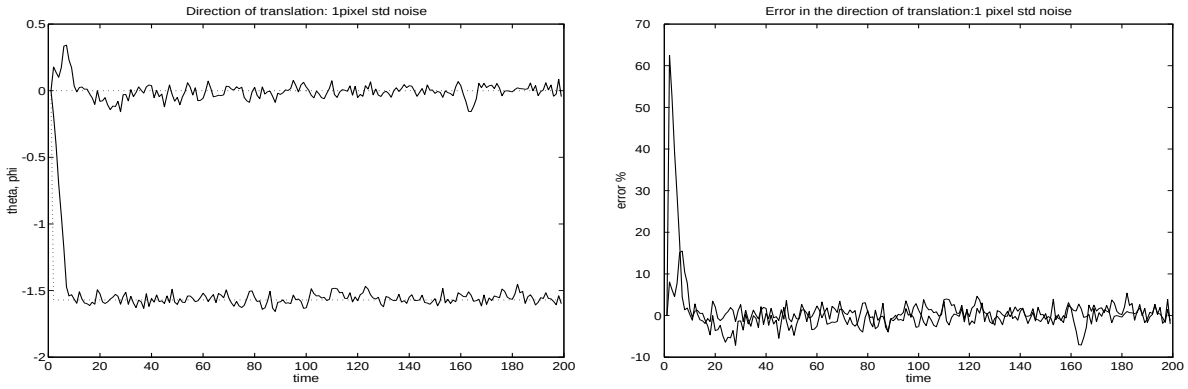

Figure 3: *Estimates and errors for the direction of translation when the noise in the image plane has a standard deviation of 1 pixel (according to the performance of common optical flow/feature tracking schemes). Ground truth is displayed in dotted lines. Note that convergence is reached from zero initial condition in about 10 steps.*

5.2.1 The residual plot in the state-space

A typical plot of the residual function, which is the value of the constraint (5) as a function of the state of the filter θ, ϕ, is shown in figure 5. The bright areas indicate a small residual value. The black asterisk indicates the position of the motion (in the local coordinates of the sphere of directions of translation) which generated the residual. It is noted that the minimum of the residual is displaced from the true motion when the norm of the rotational velocity is large. This is due to the fact that we approximate the velocity of the projected points (motion field) with first differences; the approximation is good as long as $R \doteq e^{\Omega\wedge} \cong I + \Omega\wedge$, i.e. as long as the norm of the rotational velocity is small.

5.2.2 Convergence and local minima

The reader may have noticed the presence of local minima in the plots of the residual function (figg. 5-9): if motion is estimated *instantaneously* from two frames, as in [9], the estimate can be trapped into a local minimum.

In our experiments, however, we have never experienced convergence to a local minimum, unless temporary. This is due to the recursive nature of the scheme, which integrates information over a large baseline. In figure 7 and 8 we show a typical example of the temporary convergence of the filter to a local minimum: after few iterations the observations are no longer compatible with the motion interpretation, and the norm of the pseudo-innovation process grows, forcing the filter out of the local minimum.

5.2.3 Rubbery motion

A qualitatively different local minimum is the one corresponding to the "rubbery motion". When the positive depth constraint is not enforced the filter may converge either to the rigid or to the rubbery interpretation (figure 6). In figures 7 and 8 (left) we show the convergence to the "rubbery motion interpretation" when the positive depth constraint is released. In figures 7 and 8 (right) we show the convergence of the filter to the rubbery interpretation. Note that, when the positive

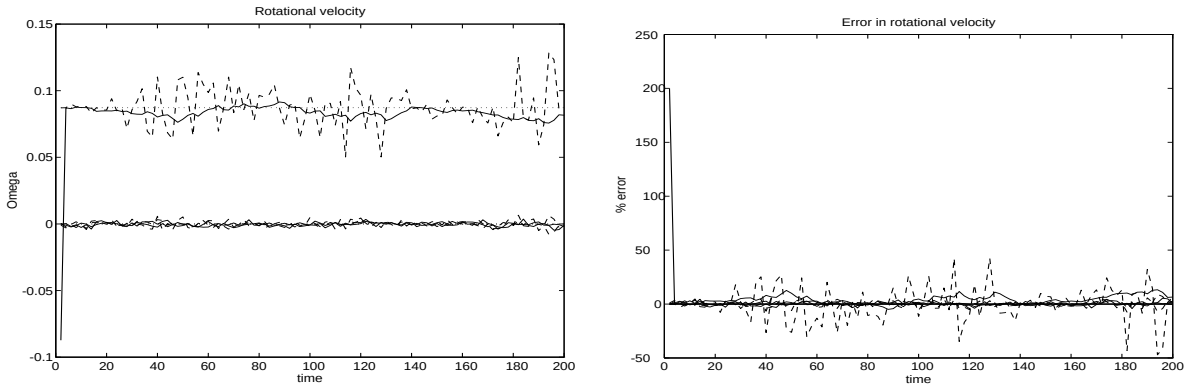

Figure 4: *Estimates for the components of rotational velocity (left) and corresponding error (right). Ground truth is displayed in dotted lines; the filtered estimates are in solid lines. The least-squares computation of the rotational velocity is in dashed lines. The linear Kalman filter was initialized with one step of a two-views algorithm using the first and the third frames.*

depth constraint is enforced, the estimate is reflected onto the correct rigid interpretation (figure 9).

5.2.4 Structure estimation

When we feed the motion estimates into a structure from motion module initialized with points at positive depth and a large model-error variance [23], we may observe either a rigid set of points which move according to the correct motion (a top view of the points is shown in figure 10 (left)) or to a "rubbery" percept (figure 10 (right)). This is in accordance with the experience in psychophysical experiments [15]. Note that the rubbery solution disappears as soon as we impose the positive depth constraint.

5.2.5 Comparison with the essential filter

The filter proposed in this paper proves far less sensitive to noise in the measurements and to the initial conditions than the essential filter [21].

In particular, for 20 observed points, the essential filter converges for initial conditions within 30 %, while the subspace filter converges from any initial condition. Furthermore, the subspace filter is less sensitive to noise, and may tolerate up to 5 times more noise on the measured image plane coordinates than the essential filter. This is due to the simple structure of the state-space of the filter as well as its low dimensionality.

Once properly initialized, however, the essential filter proves more accurate, achieving easily less than 1 % error in the components of velocity for one pixel std error or less, while the subspace filter is more robust but less accurate, achieving accuracies in the order of 2-5 % under the same conditions.

The essential filter has, in our current implementation, an advantage in terms of complexity as the number of points increases. In fact the linearization of the measurement equation C in the subspace filter has dimensions $2N \times N + 3$, where N is the number of visible feature-points, while

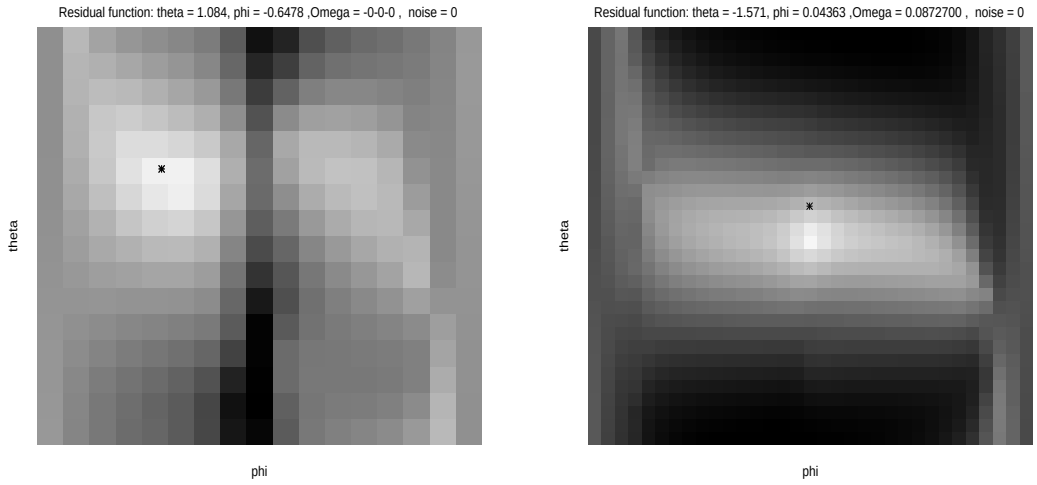

Figure 5: *Brightness plots of the residual function. The value of the residual is plot on the state-space of the filter, which are the local coordinates of the sphere of directions of translation. Bright regions denote small residuals. The black asterisk is the "true" motion which generated the residual. Note that for small rotation (left) the minimum of the residual coincides with the true motion. When translation is large (right) the Euler step approximation is no longer valid, and the minimum moves from the true location.*

in the essential filter it is $2N \times 9$. However, the linearization of the subspace filter has a sparse structure that could be exploited.

In the essential filter the positive depth constraint is encoded directly in the definition of the state space manifold (the essential manifold). The convergence of the essential filter is illustrated in fig. 11: on the left the convergence is shown when starting from the rubbery motion interpretation and imposing positive depth. On the right the positive depth constraint has been released (equivalently, reflections are allowed in the essential manifold), and therefore we may observe occasionally convergence to the local minimum corresponding to the rubbery interpretation.

5.3 Experiments with real image sequences

We have tested the scheme on real image sequences: the noise level achieved by the most common feature tracking/optical flow techniques is easily handled by the filter. As a first example we report here the filter estimates for the rocket scene, for comparison with [21]. Due to the fact that the filter takes about 10 frames to converge, we have doubled the sequence, which is displayed in figure 12.

In a second experiment we estimated the motion of a box rotating on top of a chair (see figure 13). The box has a side of approximately 25 cm and its centroid is placed at a distance of about 45 cm from the camera. The features are detected and tracked using a multiscale Sum of Square Difference (SSD) criterion. Two features are chosen as reference in order to evaluate the scale factor. The estimate of their distance in space is kept constant and, if the true distance is known at any time (even a posteriori), it is possible to rescale the translational velocity and the depth of each point.

In order to get rid of the features belonging to the background, the scene is first segmented using

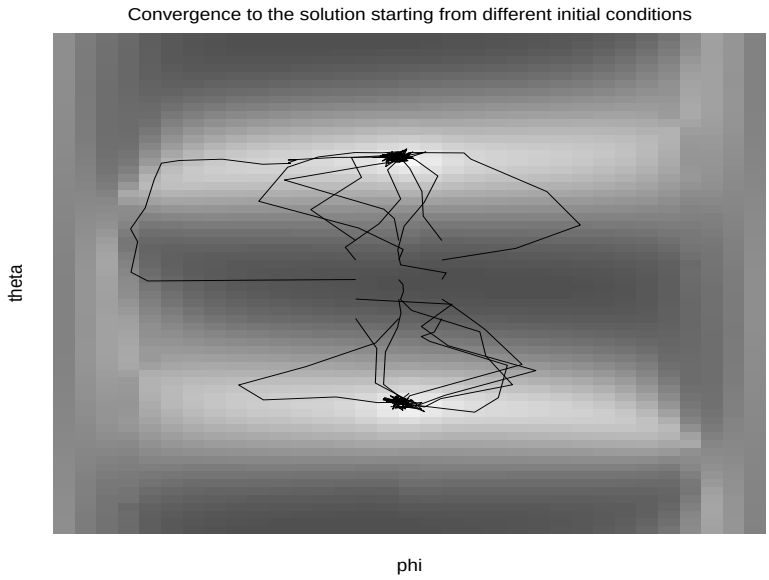

Figure 6: *Convergence when the positive depth constraint is not imposed and the initial condition is chosen at random around the origin (which appears in the center of the plot): a number of trajectories is shown in black solid lines superimposed on the brightness plot of the residual function. The filter may converge to either the correct rigid interpretation (bright spot on the top half of the plot) or to the local minimum corresponding the "rubbery" interpretation (bright area on the bottom half of the plot).*

an algorithm described in [22]. A set of clusters of points is formed based on the coherence of 3D motion between the first two frames, then a filter is initialized for each cluster, and filters converging to similar motions are merged. In a regime situation there is a filter running for each independently moving object, constantly checking the rigidity assumption and rejecting the outliers [22].

The estimates of the direction of translation, with the error-bars corresponding to the variance of the prediction error, are plotted in figure 14 (left), and similarly for the rotational velocity, which is estimated using the pseudo-measurements $\tilde{\Omega} = \tilde{\mathcal{C}}^{\dagger}_{2N+1:2N+3}\dot{\mathbf{x}}$ as input to a linear Kalman filter as described in section 3.2 (see figure 14 right).

Once motion is estimated – together with the appropriate variance of the estimation error – it is fed into a "Structure From Motion" module that processes motion error [23] in order to estimate the structure of the scene. A slice of the scene viewed from the top is plotted in figure 15 (left), and the corresponding image-plane view is depicted in figure 15 (right).

5.4 Implementation

We have implemented the filter using `Matlab` and a tensor algebra toolbox written by G. Brunthaler. Each update step consists essentially in 15 products of matrices of size varying from 2×2 to $2N\times2N$, one inversion of the $2N \times 2N$ variance of the pseudo-innovation, 5 sums and the computation of the SVD of $\tilde{\mathcal{C}}$, for a total of circa 1.066 Mflops for 20 points. However, the computation can be cut to 512Kflops by taking into account the sparse structure of the matrices involved in the computation (block-diagonal structure of R_n and $\tilde{\mathcal{C}}$). A time-consuming part of the algorithm is

also the linearization of the system with respect to the measurements, $D_+(t+1) \doteq \left(\frac{\partial \tilde{\mathcal{C}}^\perp \dot{\mathbf{x}}}{\partial \mathbf{x}(t)}\right)_{|\mathbf{y}(t),\hat{\alpha}(t)}$.

Since the Extended Kalman Filter is based upon the assumption that the linearization error is negligible, which is not often the case, we have added to the variance $D_+ R_n D_+^T$ a small symmetric random matrix in order to account for the linearization error. This practice typically improves the performance of the Extended Kalman Filter for models which are strongly nonlinear.

A more drastic measure, which speeds the filter up considerably, consists in generating the matrix D_+ at random. Despite the roughness of the operation, the filter preserves good convergence properties. This is due to the fact that, for some motion configurations, the linearization error is comparable if not more significant than the measurement error and the model error. However, when the noise level increases, it is necessary to compute the variances using the appropriate linearization matrices.

A crucial part of the design of an EKF consist in "tuning" it, i.e. in assigning a value to the elements of the variance/covariance matrices of the model errors: $R_\alpha, R_{n_{rw}}$. A custom procedure is to assume that these matrices are diagonal, and then play with their values until the prediction error is as white as possible. Standard tests are available for this procedure, such as the "cumulative periodogram". In our experiments we have performed a coarse tuning by changing the variances of the model errors by orders of magnitude. We did not perform any ad-hoc or fine tuning, and the setting was the same throughout the different experiments.

6 Conclusions

We have formulated a new recursive scheme for estimating rigid motion under perspective via identifying a nonlinear implicit dynamic model with parameters on a manifold.

The motivation comes from the work of Heeger and Jepson [9], who propose to view motion estimation as an optimization problem constrained on a subspace. However, they solve this problem by an extensive search over the unit sphere. This procedure does not exploit the geometric structure of the problem and is computationally expensive. Furthermore, it does not take into account the measurement noise, which enters into the minimization in a highly structured fashion. Temporal coherence of motion is also not taken into account; on the contrary, we want to exploit all the processing performed at the previous time instants and update the motion estimates recursively and causally.

Using results from nonlinear estimation and identification theory, we formulate a motion estimator which is fast, efficient, accurate and more robust than any recursive motion estimation scheme we have implemented. Extensive experiments have been performed that highlight such features.

Acknowledgements

We wish to thank Prof. Ruggero Frezza and Prof. Giorgio Picci for their constant support and advice. This research has been funded by the California Institute of Technology, a scholarship from the University of Padova, a fellowship from the "A. Gini" Foundation, an AT&T Foundation Special Purpose grant, ONR grant N0014-93-1-0990, grant ASI-RS-103 from the Italian Space Agency and the National Young Investigator Award (P. P.).

References

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

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

[3] Bryant, Chern, Goldberg, and Goldsmith. *Exterior Differential Systems.* Mathematical Research Institute. Springer Verlag, 1992.

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

[5] O. Faugeras. *Three dimensional vision, a geometric viewpoint.* MIT Press, 1993.

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

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

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

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

[10] D. Heeger and A. Jepson. Subspace methods for recovering rigid motion i: algorithm and implementation. RBCV TR-90-35, University of Toronto – CS dept., November 1990. Revised July 1991.

[11] D. Heeger and A. Jepson. Subspace methods for recovering rigid motion ii: theory. RBCV TR-90-35, University of Toronto – CS dept., November 1990. Revised July 1991.

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

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

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

[15] C. Kolb, J. Braun, and P. Perona. Object segmentation and 3d structure from motion. In *Invest. Ophtalmol. Vis. Sci. (Supplement)*, 1994.

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

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

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

[19] S. Soatto. Observability/identifiability of rigid motion under perspective. *Submitted to Automatica. Registered as technical Report CIT-CDS 94-001, California Institute of Technology. Reduced version to appear in the proceeding of the 33rd Conf. on Decision and Control. Also available through the Worldwide Web Mosaic* (`http://avalon.caltech.edu/cds/techreports/`), 1994.

[20] S. Soatto, R. Frezza, and P. Perona. Motion estimation via dynamic vision. *Submitted to the IEEE Trans. on Automatic Control. Registered as Technical Report CIT-CDS-94-004, California Institute of Technology. Reduced version to appear in the proc. of the 33 IEEE Conference on Decision and Control. Available through the Worldwide Web Mosaic* (`http://avalon.caltech.edu/cds/techreports/`) , 1994.

[21] S. Soatto, R. Frezza, and P. Perona. Motion estimation on the essential manifold. *In "Computer Vision ECCV 94, Lecture Notes in Computer Sciences vol. 801", Springer Verlag*, May 1994.

[22] S. Soatto and P. Perona. 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. Also in Proc. of the 1994 IEEE Workshop on Motion of Nonrigid and Articulated Objects*, 1993.

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

[25] Hermann von Helmholtz. *Treatise on Physiological Optics*. 1910.

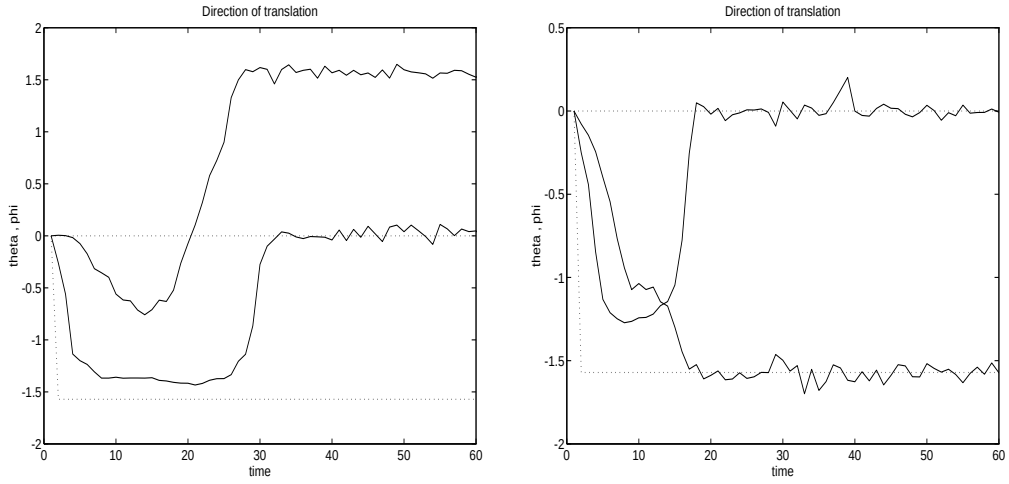

Figure 7: *(Left) convergence to a shallow local minimum and then the local minimum corresponding to the rubbery interpretation when the positive depth constraint is not enforced. (Right) convergence to a shallow local minimum and then to the correct rigid motion when the positive depth constraint is enforced (see also figure 8).*

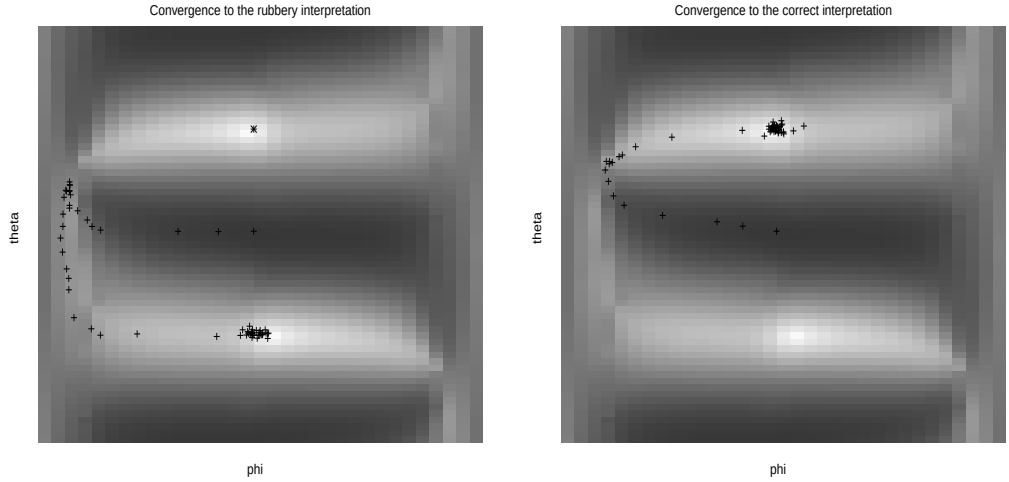

Figure 8: *Convergence to the "rubbery interpretation" (left) versus convergence to the rigid motion interpretation (right). The state of the filter at each step is represented as a black + and superimposed to the average residual function (darker tones for larger residuals). After the transient the states accumulate either around the local minimum corresponding to the rubbery interpretation (the one on the bottom half of the plot) or to the one corresponding to the true motion, on the upper half of the plot. The trajectory is also plotted componentwise in figure 7.*

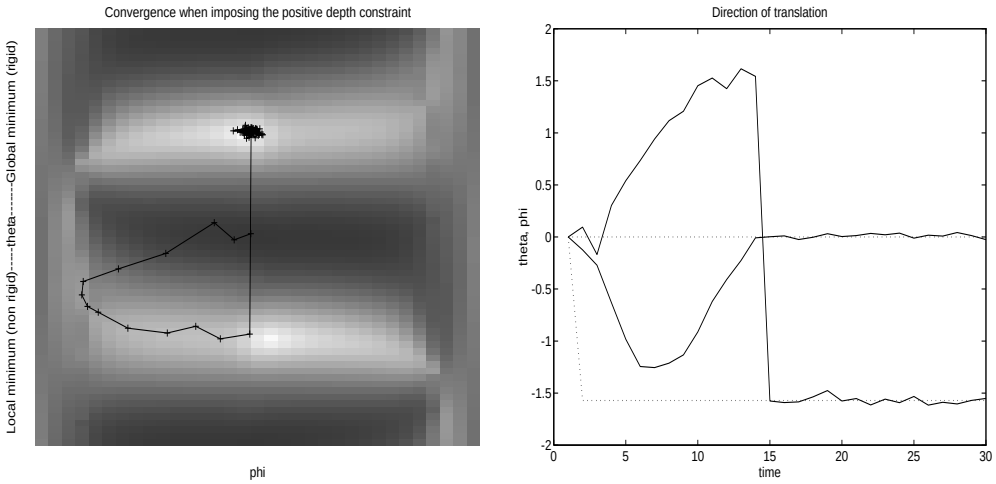

Figure 9: *Convergence when the positive depth constraint is enforced: (left) trajectory of the filter on top of the brightness plot of the residual function, (right) corresponding motion components. Initial conditions* are *zero.*

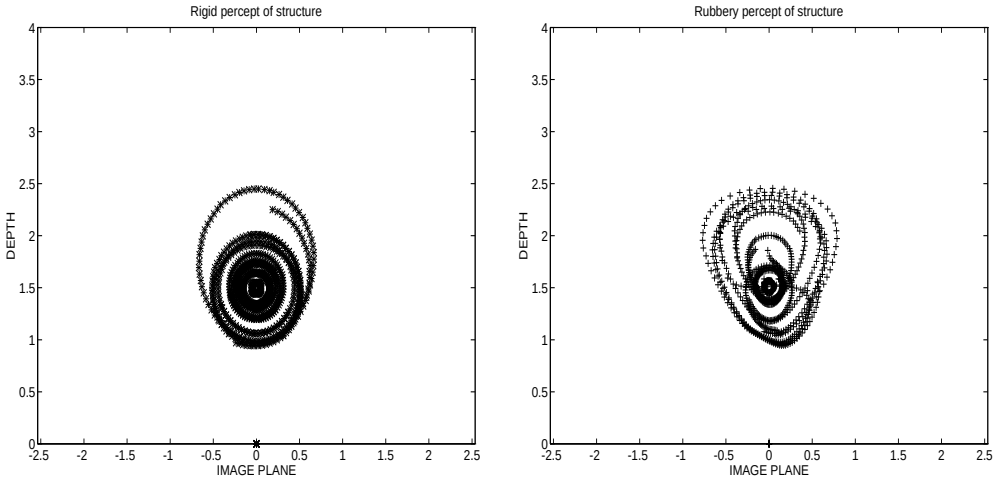

Figure 10: *Convergence of* a *structure from motion module to* a *rigid interpretation of structure (left) or to* a *rubbery object rotating in the opposite direction (right). The plots show* a *top view of the points, with the image plane on the lower end.*

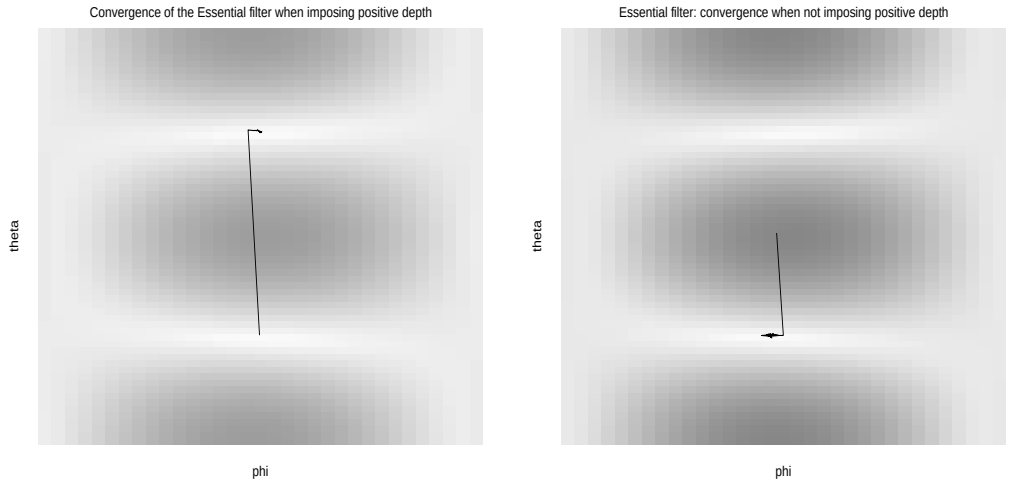

Figure 11: *Convergence of the essential filter: the residual function is plotted on a two-dimensional slice of the five-dimensional state space. The remaining states that are not represented (the ones corresponding to the rotational velocity) are set to the ground truth. On the left plot the filter is initialized with a motion close to the minimum corresponding to the rubbery interpretation. The filter, however, imposes automatically the positive depth constraint and the estimate switches fast to the correct motion interpretation. (Right) By releasing the positive depth constraint, it is possible for the filter to converge to the rubbery interpretation. The initial condition is assigned with the rotational velocity corresponding exactly to the rubbery interpretation, and the remaining two states, corresponding to the direction of translation, biased towards the local minimum of the rubbery motion.*

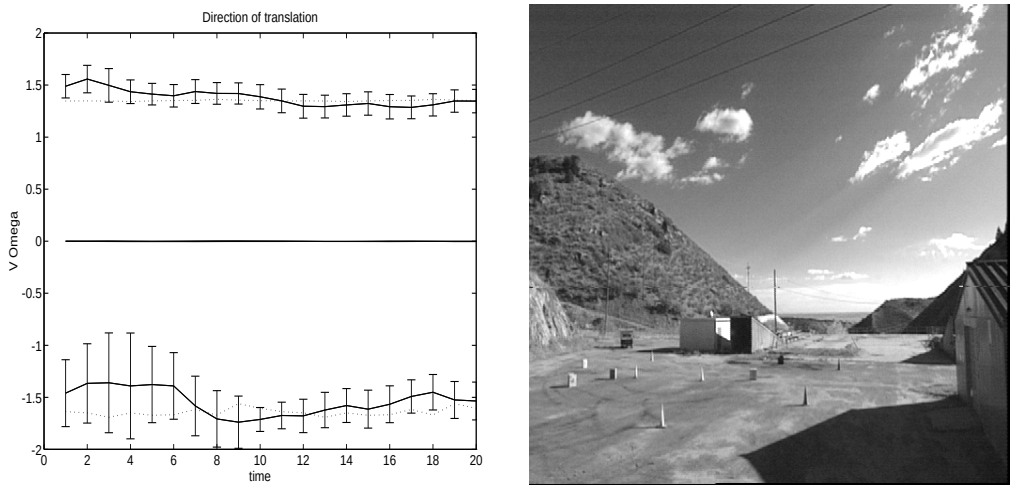

Figure 12: *(Left) Estimate of the direction of translation for the rocket scene. (Right) One image of the rocket scene. The ground truth is shown in dotted lines, while the filter estimates are in solid lines. The error-bars are three times the variance of the estimation error.*

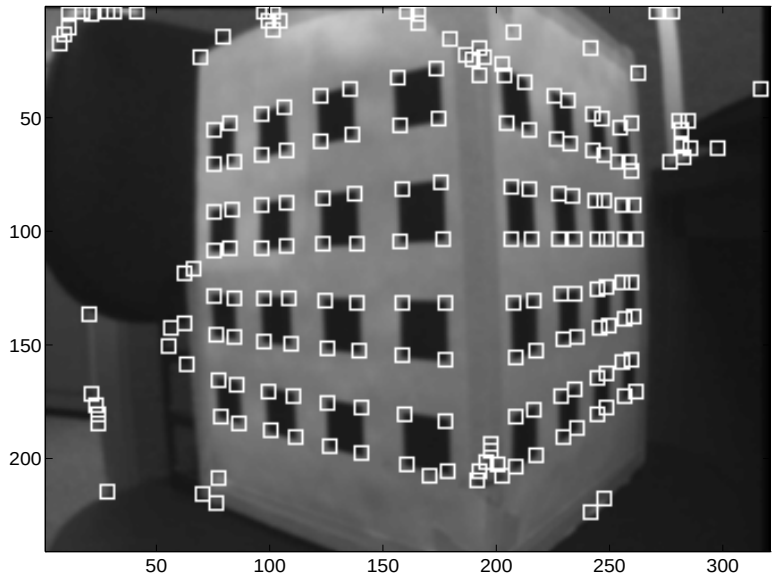

Figure 13: *One image of the box sequence. Features (marked as white boxes) are selected using the Sum of Square Difference (SSD) criterion and then clustered according to their rigid motion as estimated between the first two time instants. Two features are chosen as reference in order to update the scale factor.*

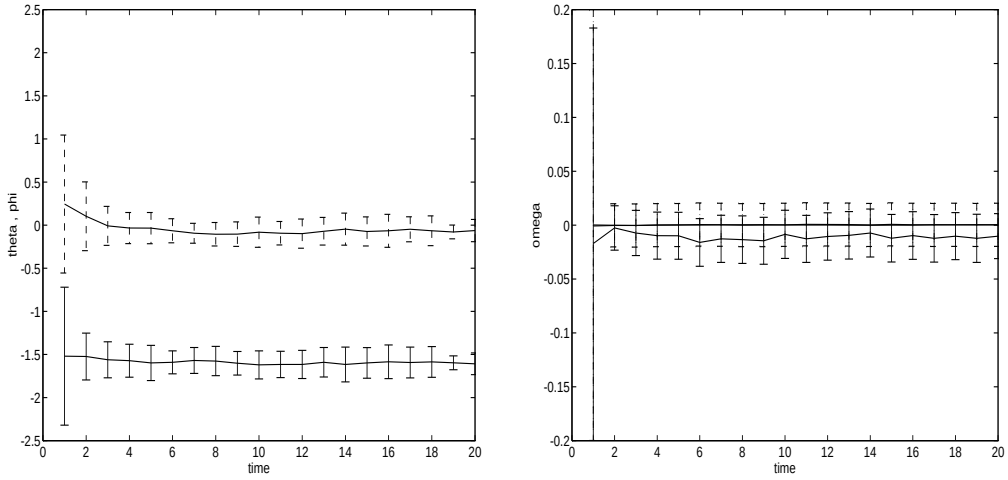

Figure 14: *(Left) Estimate of the direction of translation for the rotating box. The error-bars are three times the variance of the estimation error (diagonal of the P matrix of the filter). (Right) Estimates of the components of rotational velocity, estimated using a linear Kalman filter that processes the pseudo-measurements derived from the direction of translation, as described in section 3.2.*

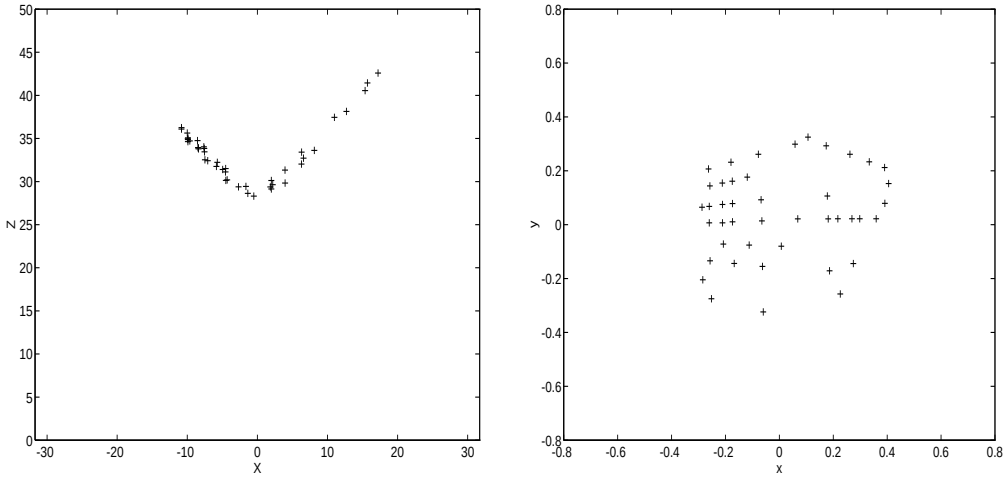

Figure 15: *(Left) Top view of the estimated scene. Note that some features have been lost during the tracking procedure. The camera was not perfectly aligned with the box, which causes the walls of the reconstructed box to appear "thick". The structure was estimated using a simple Extended Kalman Filter having as input the feature points and the motion estimates together with their variance/covariance matrices. (Right) Image-plane view of the scene.*