Dynamic Rigid Motion Estimation From Weak Perspective

Stefano Soatto† and *Pietro Perona*†‡

† Control and Dynamical Systems - CDS
California Institute of Technology 116-81, Pasadena - CA 91125
‡ Università degli Studi di Padova, Padova - Italy
{soatto,perona}@caltech.edu

Abstract

"Weak-perspective" represents a simplified projection model that approximates the imaging process when the scene is viewed under a small viewing angle and its depth relief is small relative to its distance from the viewer. We study how to generate dynamic models for estimating rigid 3-D motion from weak-perspective. A crucial feature in dynamic visual motion estimation is to decouple *structure from motion in the estimation model. The reasons are both geometric - to achieve global observability of the model - and practical, for a structure-independent motion estimator allows us to deal with occlusions and appearance of new features in a principled way. It is also possible to push the decoupling even further, and isolate the motion parameters that are affected by the so-called "bas-relief ambiguity" from the ones that are not. We present a novel method for reducing the order of the estimator by decoupling portions of the state-space from the time-evolution of the measurement constraint. We use this method to construct an estimator of full rigid motion (modulo a scaling factor) on a six-dimensional state-space, an approximate estimator for a four-dimensional subset of the motion-space, and a reduced filter with only two states. The latter two are immune to the bas-relief ambiguity. We compare strengths and weaknesses of each of the schemes on real and synthetic image sequences.*

1 Introduction

Recovering 3-D rigid motion from a sequence of perspective images is a remarkably difficult estimation problem, which becomes ill-conditioned when the field of view gets smaller, up to the point where the noise in the image makes perspective-based algorithms fail. One possible way of approaching this problem consists of *approximating* the perspective projection by some simpler projection model, such as orthographic projection or "weak-perspective". Weak-perspective is a scaled orthographic projection that can be used for approximating the imaging model when the field of view is small as well as the scene "flat" relative to its distance from the viewer. It has the advantage of being simpler than full perspective (it is in fact a scaled linear transformation). The caveat is that, while the analysis is done for one measurement model (the weak perspective), the actual measurements obey a different process (perspective projection). Therefore, the practical feasibility of motion and structure estimation from weak perspective, once tested analytically, has to be verified in practice with algorithms working on realistic sequences. Since weak-perspective is a mere approximation of the imaging formation process, the hope is that, by using a simpler model, we are able to estimate better *whatever we can estimate.*

Such an approach was chosen, for example, by Tomasi and Kanade for the case of orthographic projection [21], and extended to "para-perspective" in [15], using a batch algorithm based upon the solution of fixed-rank approximation problems with the Singular Value Decomposition. A different philosophy consists in trying to retrieve *partial* information about motion and/or structure compatible with the measured images, for example structure modulo an arbitrary affine transformation of $\mathbb{R}^3$. There is a vast literature on affine as well as partial Euclidean motion and structure reconstruction from weak-perspective; see for example [1, 3, 4, 5, 6, 7, 8, 13, 16, 12, 22] and references therein. Koenderink and Van Doorn proposed an analytic discussion about "what" can be estimated from two and three weak-perspective views of a number of feature points. In [11], they present a geometric stratification of structure and motion estimation from weak perspective, obtained by imposing subsequently the projective, affine and Euclidean structure of the problem. In this paper we are mainly concerned with real-time estimation of rigid motion, and therefore we restrict our attention to recursive and causal motion estimators.

There are a number of recursive motion and structure estimation schemes from perspective projection; however, only in rare cases has the weak-perspective approximation been considered. For instance [14] or [2], which admits it as a degenerations of the full perspective model. In all cases, however, structure is coupled to motion in the estimation model, which causes complications when dealing with occlusions or appearance of new features. In fact, one of the main obstacles encountered in implementing recursive structure and motion estimators is the short lifespan of the point-features on the image plane. Features disappear due to occlusion, or degrade when the brightness constancy assumption is violated, or change shape due to

0-8186-7042-8/95 $4.00 © 1995 IEEE

the reflectance properties of the scene, or simply exit the field of view, while new feature-point candidates enter the field of view. Representing the structure of the scene inside the dynamic model does not allow one to cope efficiently with realistic situations, since one would have to deal with a *variable number of* states on a large state-space [14].

A crucial target which we aim at in this work is the *decoupling* of structure from motion estimation, which is important both from the geometric point of view (because it makes the models observable) and from a practical side, since it allows us to deal easily with occlusion and appearance of new features. One may push such a decoupling even further, in order to isolate the motion components which are affected by the so-called "bas-relief ambiguity" [11].

In this paper we present a novel method for reducing the order of the estimator by decoupling a portion of the state-space from the time-evolution of the measurement constraint. Such a method is derived directly from what in the literature of linear dynamical systems is called the "Luenberger's reduced-order observer" [10]. We then apply such a method to construct an estimator of full rigid motion (modulo a scaling factor) on a six-dimensional state-space, an approximate estimator for a four-dimensional subset of the motion-space, and a reduced filter with only two states. The latter two are not affected by the bas-relief ambiguity.

1.1 Organization of the paper

In section 1.2 we introduce the notation and state the problem of estimating structure and motion from a dynamic model having both structure and motion in its state, which is therefore $3N(t) + 5$-dimensional, where $N(t)$ is the number of visible feature-points at time t. Section 2 describes the core of the method for reducing the order of the motion estimator. First, the idea underlying the classical reduced-order observer is reviewed, which allows us to reduce the state-space from $3N(t) + 5$ to $N(t) + 6$ (section 2.1). Then, the idea is extended in section 2.2 in order to decouple the $N(t)$ states corresponding to the structure parameters, leaving us with a 6-dimensional state-space.

The same concept is then pushed over in section 3 in order to further decouple the motion-states into the ones that are affected by the bas-relief ambiguity and the ones that are not. To this end, section 3.1 describes a choice of local coordinates of the motion space which is adapted from [11]. Section 3.2 describes the approximate filter with four states, and section 3.3 the one reduced to two-states; both are immune from the bas-relief ambiguity, for they only estimate portions of the motion-space which are not sensitive to it. Finally, in section 4 we test each filter both on real and synthetic noisy image sequences.

1.2 Notation and statement of the problem

We consider a number N of point-features $\mathbf{P}^i$ of coordinates $\mathbf{X}^i \in \mathbb{R}^3$ $\forall i = 1 : N$, that project perspectively onto the image plane in the points $\bar{\mathbf{p}}^i \in \mathbb{R}P^2$. The weak-perspective points

$$\mathbf{p}^i \doteq \pi(\mathbf{P}^i), \text{ of coordinates } \mathbf{x}_i \doteq \frac{\begin{bmatrix} \mathbf{X}_1^i \\ \mathbf{X}_2^i \end{bmatrix}}{\bar{d}}, \tag{1}$$

where $\bar{d}$ is the average distance between the scene and the center or projection, can be considered an approximation to the true projection under a small visual angle and a negligible relief. If the scene undergoes a *rigid motion* $g \in SE(3)$ – which can be represented as an instantaneous translation $T \in \mathbb{R}^3$ and a rotation matrix $R \in SO(3)$ – then the motion of the points in 3-D and the weak-perspective measurements describe a nonlinear dynamical system of the form

$$\begin{cases} \mathbf{P}^i(t+1) = g_t \circ \mathbf{P}^i(t) \in \mathbb{R}^3 \quad \textit{rigid motion} \\ g_{t+1} = g_t \oplus n_g(t) \in SE(3) \; \textit{small acceleration} \\ \mathbf{p}^i(t) = \pi(\mathbf{P}^i(t)) \in \mathbb{R}^2 \quad weak - perspective \end{cases} \tag{2}$$

where $\oplus$ represents a first order random walk in the local coordinates of the motion parameters (as a simple mean of modeling some inertia).

In principle, a state observer for the above dynamical model could be employed for estimating jointly the structure and rigid motion of the scene from weak-perspective. Such an observer might, however, exhibit poor performance due to the observability properties of the model [17] and to the presence of structure in the state, which results in a state-space that has high and changeable dimension, as points get occluded or move out of the visual field. One possible strategy consists in trying to reduce the dimensions of the state-space as much as possible, ending up with a small-dimensional highly-constrained state-space model. In the next section we are going to explore the principles of reduced-order observers, which are the basis for constructing dynamic models for estimating motion independent of the scene's structure.

2 The general principle: pushing the reduced order observer

In this section we will outline the main idea underlying this work, which consists in pushing the process that leads to the so-called "reduced-order observer" to upper levels of time-delays (or Lie-derivatives in the continuos-time case). We will illustrate the principle using the symbolic model (2) in order to keep the discussion as free as possible from lengthy notation; only at a later stage in section 4 will we report the actual expression of the model in its local coordinates.

2.1 Reducing the order of the model

We start from the model (2) and operate a change of coordinates (into observable canonical form) in order to linearize the measurement equation, which leads us to a model in the form

$$\begin{cases} \mathbf{p}^i(t+1) = f_1(\mathbf{p}^i, g_t, \bar{d}) + f_2(g_t, \bar{d})s^i \in \mathbb{R}^2 \\ s^i(t+1) = h_1(\mathbf{p}^i, g_t, \bar{d}) + h_2(g_t, \bar{d})s^i \in \mathbb{R} \\ g_{t+1} = g_t \oplus n_g(t) \in SE(3) \\ \mathbf{y}^i = \mathbf{p}^i + n^i \in \mathbb{R}^2 \qquad \forall\, i \;=\; 1 \ldots N \end{cases} \tag{3}$$

where $s^i \doteq \mathbf{X}_3^i/\bar{d}$ is the relative depth of each point and n^i is a measurement noise which is assumed to be zero-mean, white and Gaussian. We omit the time argument when it is t.

The first step towards reducing the order of the model consists in eliminating from the state the variables that are directly measured. Using a technique extrapolated from the so-called *reduced-order observer* [10], one can "solve" the measurement equation for the states one wishes to eliminate:

$$\mathbf{p}^i = \mathbf{y}^i - n^i \tag{4}$$

and then substitute them into the dynamics of the remaining states in equation (3):

$$\begin{cases} s^i(t+1) = h_1(\mathbf{y}^i, g_t, \bar{d}) + h_2(g_t, \bar{d})s^i + n_{s_i} \\ \bar{d}(t+1) = l(\mathbf{y}^i, g_t, s^i, \bar{d}) + n_d \\ g_{t+1} = g_t \oplus n_g(t) \qquad \in \{SE(3) \bmod \mathbb{R}\}. \end{cases} \tag{5}$$

The dynamics of the average depth $\bar{d}$ has been isolated from the other scaled depths s^i, since it will play a role in the coordinatization of the motion parameters in the presence of the scale-factor ambiguity (section 3.1). The original measurement equation becomes now trivial; however, the dynamics of the variable being eliminated becomes the new measurement constraint, which involves one time-delay:

$$\mathbf{y}^i(t+1) - f_1(\mathbf{y}^i, g_t, \bar{d}) - f_2(g_t, \bar{d})s^i = \tilde{n}^i. \tag{6}$$

The noise terms n_{s_i}, n_d and $\tilde{n}^i$ are induced by the measurement noise n^i. In principle, the time-delay could be eliminated from the measurement equation (6) using an output-dependent change of coordinates [10]. Here we do not pursue this approach, and we are content with keeping two images in memory at each time. The notation $\{SE(3) \bmod \mathbb{R}\}$ reminds us that there is an overall scale ambiguity in recovering the motion parameters; as a result, we represent g_t as a normalized translation $T \in \mathbb{S}^2$ and a rotation matrix $R \in SO(3)$.

2.2 Decoupling structure from motion

With the simple procedure described above, we have reduced the state-space from $3N + 5$ down to $N + 6$, while adding a time-delay to the $2N$ measurements. One could push this idea even further, and eliminate the N parameters s^i from the $2N$ new measurement equations (6):

$$s^i = f_2^\dagger(g_t, \bar{d}) \left(\mathbf{y}^i(t+1) - f_1(\mathbf{y}^i, g_t, \bar{d})\right), \tag{7}$$

where † denotes the subspace inverse, and then substitute them into the dynamics of the remaining states in (5). The measurement equation no longer becomes trivial, for there are still N independent constraints that have not been employed for "eliminating" s^i:

$$f_2^\perp(g_t, \bar{d}) \left(\mathbf{y}^i(t+1) - f_1\right) = f_2(g_t, \bar{d})^\perp \tilde{n}^i \tag{8}$$

where ⊥ denotes the subspace orthogonal complement. Again, the dynamics of the variable being eliminated becomes a measurement constraint, with now two delays. The final expression of the model becomes therefore of the form

$$\begin{cases} \bar{d}(t+1) = l(\mathbf{y}^i, g_t, f_2^\dagger \mathbf{y}^i(t+1) - f_2^\dagger f_1, \bar{d}) + n_d \\ g_{t+1} = g_t \oplus n_g(t) \qquad \in \{SE(3) \bmod \mathbb{R}\} \\ f_2^\perp(g_t, \bar{d}) \left(\mathbf{y}^i(t+1) - f_1\right) = f_2(g_t, \bar{d})^\perp \tilde{n}^i \\ f_2^\dagger(g_t, l) \left(\mathbf{y}^i(t+2) - f_1(\mathbf{y}^i(t+1), g_t, l)\right) + \\ \qquad -h_1 - h_2 f_2^\dagger \left(\mathbf{y}^i(t+1) - f_1\right) = \tilde{n}_{s_i} \end{cases} \tag{9}$$

where the arguments in f_1, h_1, h_2 and l have been omitted and $\tilde{n}_{s_i}$ is, as usual, a noise induced by substituting the measurements into the dynamics of s^i. We can write the above model in a more synthetic form as

$$\begin{cases} \xi(t+1) = m(\xi(t)) + n_\xi(t) \qquad \in M \sim \mathbb{R}^6 \\ \chi\left(\xi(t), \mathbf{y}(t), \mathbf{y}(t-1), \mathbf{y}(t-2)\right) = n_\chi(t) \ \in \mathbb{R}^{3N} \end{cases} \tag{10}$$

where ξ belongs to a six-dimensional state-space manifold that encodes the motion parameters and the average depth of the scene, and only $2N$ of the measurement constraints are independent.

In the experimental section 4 we will show an actual expression of the above model with an appropriate choice of coordinates, along with experiments of the performance of the filter derived from such a model on real and synthetic image sequences.

One could play the game just described over and over, and successively eliminate each state by solving from the time-evolution of the measurement equation and substituting into the state equations. This process is guaranteed to succeed as long as the dynamic model in question is locally-weakly observable [9]. In fact, the above process could be regarded as analogous to a level-wise inversion of the Observability Grammian in the linear case. Of course, the more levels are involved, the higher the number of delays that appear in the measurement equations (or the higher the number of Lie-derivatives of the measurements in the continuous-time case). In our case, two delays are sufficient, for the model is observable with two levels of bracketing. Complete proofs of these statements are beyond the scope of this paper.

3 Isolating the bas-relief ambiguity: motion decoupling and choice of coordinates

It is well-known that, under small visual angles and negligible relief, it is difficult to resolve some of the motion parameters. This effect, which is known as the "bas-relief ambiguity", can be observed for example by taking two flat surfaces, connecting them rigidly at a right angle, and rotating them about an axis (a rotating billboard, see figure 1). The perceived motion from a distance is strikingly non-rigid, as the surface which is more slanted seems to move faster. In the presence of the bas-relief ambiguity, it is important to parametrize the state-space manifold M so that the states that are affected are "isolated". Failure to do that may result in poor estimates of *all the states*, including the ones that are not affected by the ambiguity.

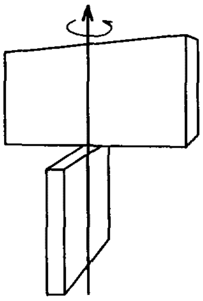

Figure 1: *One of the manifestations of the "bas-relief ambiguity" is evident from watching a rotating billboard. From a distance, the more slanted the surface, the faster it seem to move, while the two surfaces appear to move disjointly.*

3.1 Choosing the motion coordinates

At this point, one may ask if it is possible to push the reasoning outlined in section 2.1, and formulate filters that estimates *only the motion parameters that are not affected by the bas-relief ambiguity.* The procedure outlined in the previous section relies on the fact that one is able to "solve" for the states to be eliminated from the time-evolution of the measurement equation. Eliminating $\mathbf{p}^i$ and s^i was easy because they appeared *linearly* in the measurement equations (3) and (6) respectively. However, it is not so for the motion parameters, which are encoded into $\xi(t)$. In this section we will see that it is possible to decouple the motion parameters and formulate two filters, one with four states, and one with two states only, which are not affected by the bas-relief ambiguity. To this end, we need to make a choice for the local-coordinate parametrization of the motion parameters $\xi \in M \sim \mathbb{R}^6$.

We choose to represent motion using $\xi \doteq [\tilde{T}^T, \theta, \phi, \rho]^T$, defined such that

$$\tilde{T} \doteq \frac{T}{\bar{d}} \in \mathbb{R}^3 \qquad R = e^{\mathbf{e}_3\wedge\theta} e^{\left[e^{\mathbf{e}_3\wedge\phi}\mathbf{e}_1\right]\wedge\rho} \in SO(3) \tag{11}$$

where $(\mathbf{e}_3\wedge)$ is a 3×3 skew-symmetric matrix having all zeros but -1 in position $(1,2)$ and 1 in position $(2,1)$. The Euler-angle representation of rotation, which was introduced by Koenderink and Van Doorn [11], corresponds to rotating by ρ radiants about an axis on the image-plane, forming an angle ϕ with the horizontal axis, and then rotating about the optical axis by an angle θ. It has the advantage that the bas-relief ambiguity is isolated in the parameter ρ, while cyclo-rotation θ and the angle ϕ are always easy to estimate. The disadvantage is that, like all Euler-angles, it is only a local representation, and a filter based upon such a representation may run into singularities.

3.2 Approximate filter with four states

Under the choice of coordinates described in (11), eq. (8) may be written as

$$\left[\begin{array}{ccc} & & \vdots \\ \mathbf{y}(t+1)^T & \mathbf{y}(t)^T & 1 \\ & & \vdots \end{array}\right] \left[\begin{array}{c} cos(\phi)v(t) \\ sin(\phi)v(t) \\ -cos(\phi-\theta) \\ -sin(\phi-\theta) \\ -w(t) \end{array}\right] = \tilde{n} \tag{12}$$

which is a rank-four homogeneous equation up to zero-mean noise. In the above equation, v and w are approximated by [22]

$$\left\{\begin{array}{l} v(t) \cong sin(\rho)[-sin(\phi)\ cos(\phi)]\bar{\mathbf{y}}(t) + cos(\rho) + \tilde{T}_3 \\ w(t) = sin(\rho)\left(-sin(\phi)\tilde{T}_1 + cos(\phi)\tilde{T}_2\right) \end{array}\right. \tag{13}$$

under the condition $\rho \cong 0$. Eliminating ρ from this constraint, even though simplified, is not a trivial matter. A naïf approach consists in writing a filter for the 4 variables $\psi(t) \doteq [v(t), w(t), \theta(t), \phi(t)]^T$, having (12) as an implicit measurement equation. The problem is that the dynamics of ψ involves *all the states* ξ, and therefore we cannot hope to eliminate some of them and use their dynamics as a measurement equation. In fact, note that equation (12) comes from the residual measurement equations that were not used for eliminating s^i, but then it is necessary to integrate the measurement equations with the dynamics of the variables s^i being eliminated.

An approximate filter can be obtained, however, by modeling the dynamics of v and w as a random walk, and *neglecting the dynamics of the variables* $\tilde{T}, \rho$. Such a filter will have a reduced measurement constraint, and an approximate dynamics:

$$\left\{\begin{array}{l} \psi(t+1) = \psi(t) + n_\psi \ \in \mathbb{R}^4 \\ \text{eq. (12)} \in \mathbb{R}^N. \end{array}\right. \tag{14}$$

Note that, once the filter has estimated v and w, there is no way of unfolding the motion parameters $\tilde{T}$ and ρ out of them; we must be content with the four parameters ψ, which are only a partial representation of motion. Such an approach has been pursued, for instance, in [22], although derived differently.

3.3 Reduced filter with two states

The redundancy in the measurements may be exploited to the point in which we define a filter with only two states. To this end, consider eq. (12), which is obtained by eliminating the relative depth parameters s^i. The motion parameters $\tilde{T}$ and ρ appear through v and w, defined in eq. (13). Therefore, we may eliminate these four variables and be left with a filter that has only θ and ϕ in its state. To this end, rewrite eq. (12) as

$$\begin{bmatrix} & \vdots \\ \mathbf{y}^T_{(t+1)} \begin{bmatrix} c_\phi \\ s_\phi \end{bmatrix} & -1 \\ & \vdots \end{bmatrix} \begin{bmatrix} v(t) \\ w(t) \end{bmatrix} = \mathbf{y}^T_{(t)} \begin{bmatrix} c_{\phi-\theta} \\ s_{\phi-\theta} \end{bmatrix} + \tilde{n}(t) \tag{15}$$

or, in a more condensed form,

$$\mathcal{A}(t+1,\phi) \begin{bmatrix} v(t) \\ w(t) \end{bmatrix} = \mathcal{B}(t,\theta,\phi) + \tilde{n}(t). \tag{16}$$

Then, eliminating $\tilde{T}$ and ρ can be easily done my eliminating v and w, which appear *linearly* in the above equation. The final expression of the model involving only ϕ and θ is therefore

$$\begin{cases} \theta(t+1) = \theta(t) + n_\theta(t) \\ \phi(t+1) = \phi(t) + n_\phi(t) \\ \mathcal{A}^\perp(t+1,\phi)\mathcal{B}(t,\theta,\phi) = n_r(t) \end{cases} \tag{17}$$

where n_r is the noise of the reduced constraint, which is induced by the measurement noise n, and n_θ and n_ϕ are noise models driving the random walk, whose variances are to be regarded as tuning parameters.

4 Experimental Assessment

We have implemented three recursive filters for the models of eq. (9), (14) and (17), using a local observer based upon the Implicit Extended Kalman Filter, which is derived in [18, 19]. The equations of the filter can be derived directly from that reference; the only thing needed is the model and an expression of the local linearization of the model. Space limitations do not allow us to report all of the computation; we restrict ourselves here to writing the actual equations in the local coordinates for the model (9) (the other two are already in local coordinates and ready for use)

$$\begin{cases} \tilde{T}(t+1) = \frac{\tilde{T}(t)}{v(t)} \\ \theta(t+1) = \theta(t) + n_\theta(t) \\ \phi(t+1) = \phi(t) + n_\phi(t) \\ \rho(t+1) = \rho(t) + n_\rho(t) \\ eq.\ (12) \in \mathbb{R}^N \\ \Psi \begin{bmatrix} s_\phi v(t+1)v(t) \\ -c_\phi v(t+1)v(t) \\ -s_{\phi-\theta}v(t) - s_\phi c_\rho v(t) \\ c_{\phi-\theta}v(t) + c_\phi c_\rho v(t) \\ s_\phi s_\rho^2 + s_{\phi-\theta}c_\rho \\ -c_\phi s_\rho^2 - c_{\phi-\theta}c_\rho \\ (-s_\phi \tilde{T}_1 + c_\phi \tilde{T}_2)(1-c_\rho) + \tilde{T}_3 s_\rho \end{bmatrix} = n_\Psi \end{cases} \tag{18}$$

where $\Psi = [\mathbf{y}^T(t+2)\ \mathbf{y}^T(t+1)\ \mathbf{y}^T(t)\ 1] \in \mathbb{R}^{N\times 7}$ and v and its dynamics are defined from equation (13).

All the filters have been implemented in `Matlab` and tuned with the same parameters. Tuning the filters has proven to be a rather non-trivial matter. Since the measurements are actually generated by a full-perspective projection, while the dynamic model regards them as the outcome of a scaled orthography, the variance of the measurement noise must be increased so as to account for the perspective distortion. As a consequence, the variance of the model error must be also increased in order to avoid over-smoothing. Since perspective distortion is very poorly modeled by a white and zero-mean Gaussian noise, all the filters based upon the models described above (as well as all other models based upon the weak-perspective model) perform poorly in the presence of significant perspective effects.

The three schemes have similar computational complexity and they run, in the current `Matlab` implementation, at about 2 to 10 Hertz on a Sun Sparc 20, depending upon the number of visible points (usually on the order of 10 to 100).

4.1 Simulation experiment

A cloud of 20 dots, 1 meter in diameter, was generated at a distance of 10 meters from a viewer and rotated about a vertical axis with a speed of about five degrees per frame. Its projection onto a virtual image plane of 500 × 500 pixels was corrupted with noise whose level was varying between one tenth of a pixel to one pixel std. The three filters exhibit similar performance, for the states which they have in common, in the presence of low noise levels (see figure 2). As the noise level increases, the full filter with 6 states is affected by the bas-relief ambiguity, so that two of its states are estimated poorly (figure 3). However, notice that the remaining states converge with estimation errors very similar to that exhibited by the approximate filter and by the reduced filter, which are not affected by the bas-relief ambiguity.

4.2 The arm experiment

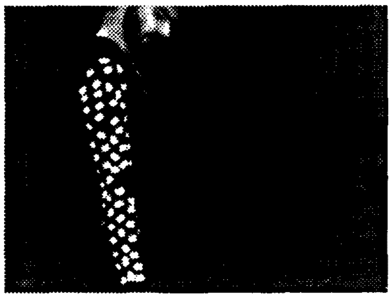

Figure 4: *L. Goncalves in his mimetic attire. The "arm sequence" is 250 frames long and the motion is rotatory on a plane parallel to the image plane. The arm was rotating upwards for half of the sequence, and then downwards for the rest of it.*

The "arm" experiment consists of a sequence of about 250 frames kindly provided to us by L. Goncalves. An arm with high contrast texture was rotating with a velocity of about half a degree per frame (figure 4). Features were selected and tracked automatically using simple gradient methods. The estimates of the full relative motion between the arm

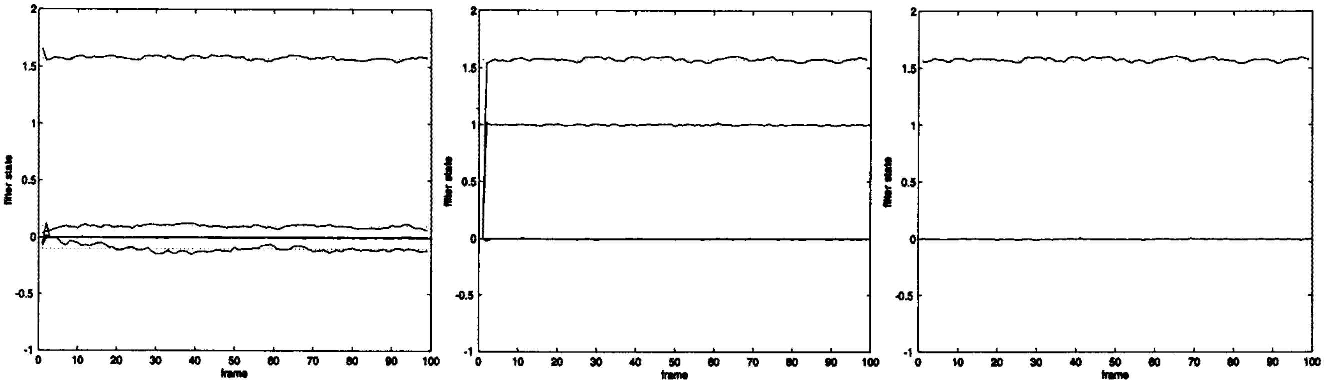

Figure 2: *Simulation experiment. Estimates of each filter (solid lines) along with ground truth (dotted lines) for a noise level of one tenth of a pixel std. The left plot shows the estimates of the state of the full filter with six states, the middle plot is the approximate filter with four states, and the right plot is the reduced filter with two states. Units are radiants/frame for the rotational velocity. Translation is adimensional since it is scaled to the average depth.*

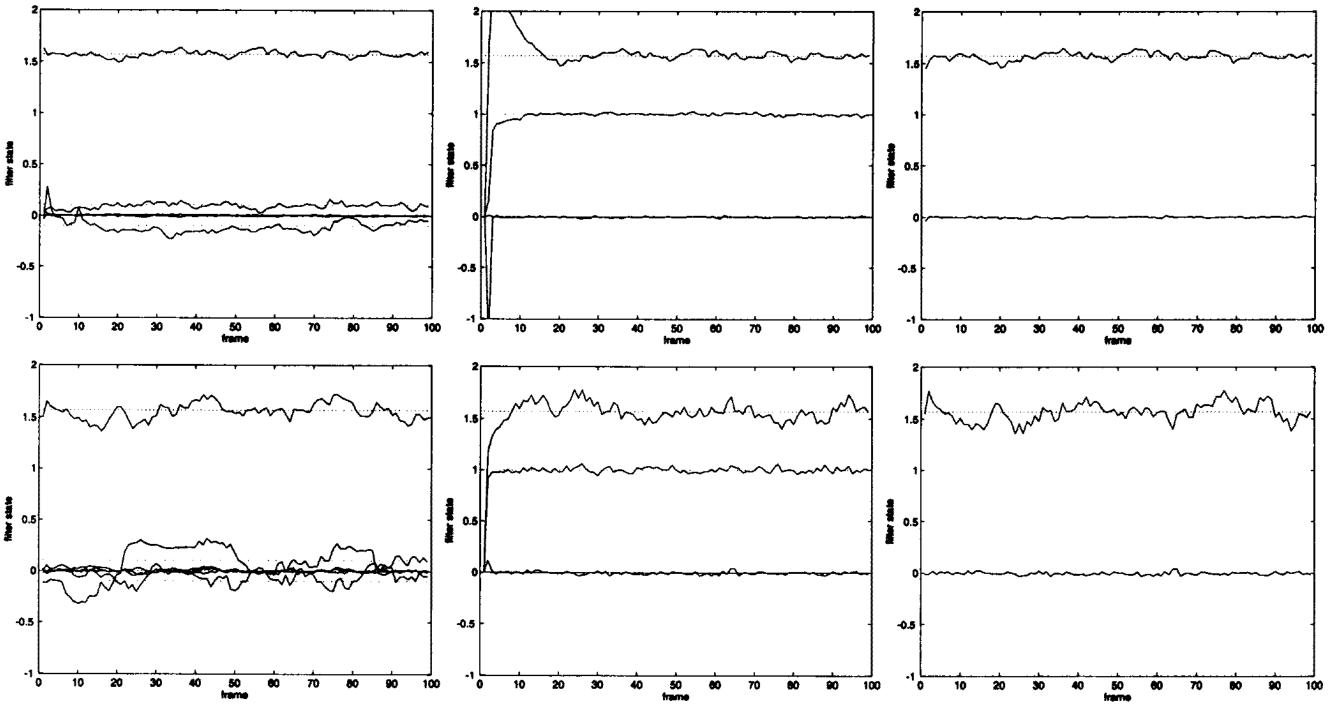

Figure 3: *Degradation of the estimates with increasing measurement noise. In the top row we report the behavior of the filters for a noise level of half a pixel std, and in the bottom row for one pixel std. We plot the estimates of each filter (solid lines) along with ground truth (dotted lines). The full-filter with 6 states (left column) degrades unevenly, for two of its states are subject to the bas-relief ambiguity. However, the particular choice of coordinates still allows estimating correctly the remaining 4 states which are not subject to the bas-relief ambiguity. The affine filter (central column) and reduced filter (right column) are not affected by the bas-relief ambiguity, and their estimation error increases gracefully with the increasing level of measurement noise. Units are rad/frame for the components of rotational velocity.*

and the camera are estimated by the full filter with 6 states, as reported in figure 5. The estimates correspond to the qualitative ground-truth provided with the sequence. In figure 6, we plot the variance of each estimate represented using error-bars. Since motion is mainly cyclo-rotational, any estimate of the angle ϕ is correct. Indeed, we are in a singularity of the coordinate representation. The filter estimates ϕ as being approximately $\frac{\pi}{2}$, and correctly assigns a large variance to the estimate. The estimates of the only significant state in common among all filters are compared in figure 7. There we also report the cyclo-rotation as estimated by the "Susbspace filter" [20], which is based upon a full-perspective model. The estimates of the filters are consistent. The ones based upon the weak-perspective models are more jittery, since the variance of the measurement error has to be increased in the tuning in order to account for the perspective distortion.

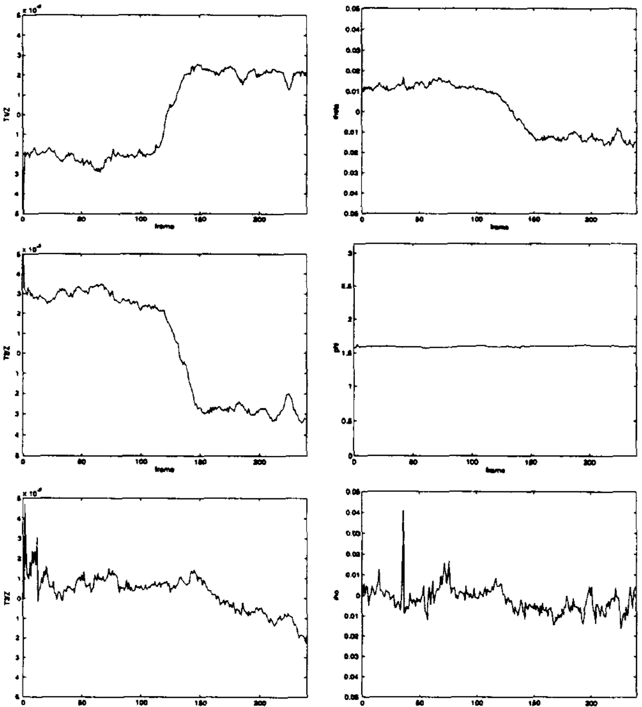

Figure 5: *The "arm experiment". In the left column we plot the three components of the estimated direction of translation normalized to the average depth of the scene; in the right column we display, respectively from top to bottom, the local coordinates of rotation: θ, ϕ and ρ. The algorithm was using on average 10 feature-points per frame. Units are rad/frame for the components of rotational velocity. Translation is adimensional since it is scaled to the average depth.*

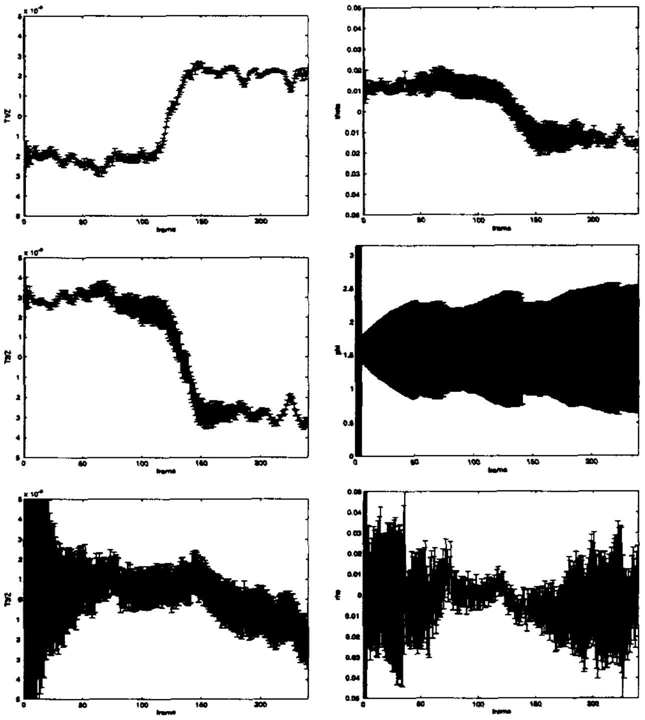

Figure 6: *The same estimates reported in figure 5 are now plotted along with their variance, represented using error-bars. It can be seen that, since rotation occurs only about the optical axis, the direction of the rotation axis on the image-plane, ϕ is arbitrary, and is indeed estimated with a very large variance (middle-right plot).*

Acknowledgements

This work was supported by a scholarship from the "Fondazione A. Gini" and a scholarship from the University of Padova - Italy (S. S.), the California Institute of Technology, the NSF ERC Center for Neuromorphic Systems Engineering, the NSF FYI Award (P. P.), and the Italian Space Agency ASI-RS-103.

References

[1] J. Aloimonos. Perspective approximations. *Image and vision computing vol. 8 nO. 3*, 1990.

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

[3] J. Nicola, B. Bennett, D. Hoffman and C. Prakash. Structure from two orthographic view of rigid motion. *J. Opt. Soc. of Am. vol. 6 no. 7*, 1989.

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

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

[6] C. Harris. Structure from motion under orthographic projection. *Proc. of the 1st ECCV*, 1990.

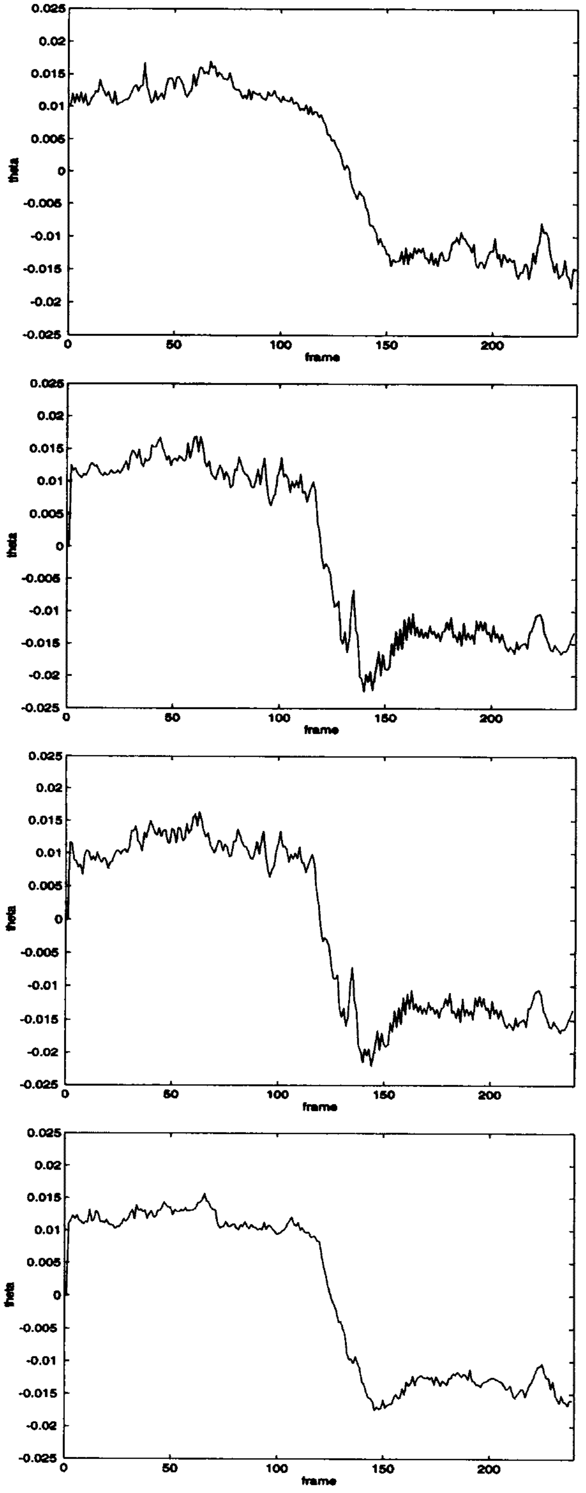

Figure 7: *Comparison of the estimates of the angle θ for, respectively from top to bottom, the full filter (six states), the approximate filter (four states), the reduced filter (two states), and the subspace filter based upon full-perspective.*

[6] C. Harris. Structure from motion under orthographic projection. *Proc. of the 1st ECCV,* 1990.

[7] X. Hu and N. Ahuja. Motion estimation under orthographic projection. *IEEE Trans. Rob. and Aut. vol 7 no 6,* 1991.

[8] T. Huang and C. Lee. Motion and structure from orthographic projections. *IEEE Trans. Pattern Anal. Mach. Intell.,* 1989.

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

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

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

[12] A. Zisserman L. Shapiro and M. Brady. Motion from point matches using affine epipolar geometry. *Proc. of the ECCV94, Vol. 800 of LNCS, Springer Verlag,* 1994.

[13] J. Lawn and R. Cipolla. Robust ego-motion estimation from affine motion parallax. *Proc. of the ECCV94, Vol. 800 of LNCS, Springer Verlag,* 1994.

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

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

[16] A. Sashua. Projective structure estimation. *MIT AI memo 1183,* 1993.

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

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

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

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

[21] C. Tomasi and T. Kanade. Shape and motion from image streams, a factorization method 1-3. Technical Report CMU-CS-90-166, Carnegie Mellon University, Sept. 1990.

[22] J. Weber and J. Malik. Rigid body segmentation and shape description from optical flow. *Proc. of the 5 IEEE Int. Conf. Comp. Vision,* 1995.