Motion from Fixation*

Stefano Soatto and *Pietro Perona*

Control and Dynamical Systems
California Institute of Technology 116-81
Pasadena – CA 91125, USA
soatto@benissimo.caltech.edu

February 20, 1995

Abstract

We study the problem of estimating rigid motion from a sequence of monocular perspective images obtained by navigating around an object while fixating a particular feature point. The motivation comes from the mechanics of the human eye, which either pursuits smoothly some fixation point in the scene, or "saccades" between different fixation points. In particular, we are interested in understanding whether fixation helps the process of estimating motion in the sense that it makes it more robust, better conditioned or simpler to solve.

We cast the problem in the framework of "dynamic epipolar geometry", and propose an implicit dynamical model for recursively estimating motion from fixation. This allows us to compare directly the quality of the estimates of motion obtained by imposing the fixation constraint, or by assuming a general rigid motion, simply by changing the geometry of the parameter space while maintaining the same structure of the recursive estimator. We also present a closed-form static solution from two views, and a recursive estimator of the absolute attitude between the viewer and the scene.

One important issue is how do the estimates degrade in presence of disturbances in the tracking procedure. We describe a simple fixation control that converges exponentially, which is complemented by a image shift-registration for achieving sub-pixel accuracy, and assess how small deviations from perfect tracking affect the estimates of motion.

1 Introduction

When a rigid object is moving in front of us (or we are moving relative to it), the information coming from the time-varying projection of the object onto *one* of our eyes suffices to estimate its motion, even when its shape is unknown.

*Research sponsored by NSF NYI Award, NSF ERC in Neuromorphic Systems Engineering at Caltech, ONR grant N00014-93-1-0990. This work is registered as CDS technical report n. CIT-CDS 95-006, February 1995.

In order to observe the motion of the object while holding our head still and one eye closed, we can choose either to track it (or a particular feature on its surface) by moving the eye, or to hold the eye still (by fixating some feature in the still background), and let the object cross our field of view. When it is us moving in the environment (or "object"), our eye constantly "holds" on some particular feature in the scene (smooth pursuit) or "jumps" between different features (saccadic motion).

From a geometric point of view there is no difference between the observer moving or the object moving, and the problem of estimating rigid motion from a sequence of projections is by now fairly well understood. In this paper we explore how the fixation constraint modifies the geometry of the problem, and whether it facilitates the task.

This problem has been in part addressed before in the literature of computational vision. In [6, 5], the fixation constraint is exploited for recovering the Focus of Expansion (FOE) and the time-to-collision using normal optical flow, and then computing the full ego-motion, including the portion due to the fixating motion. In [12], a pixel shift in the image is used in order to derive a constraint equation which is solved using static optimization in order to recover ego-motion parameters, similarly to what is done in [3, 10]. However, nowhere in the literature is the estimation of motion, performed by imposing the fixation constraint, directly compared with the estimation of a general rigid motion, due to the lack of a common framework. More seriously, most of the algorithms assume that perfect tracking of the fixation point has been performed, and it is not assessed how they degrade in the presence of inevitable tracking errors.

In this paper we study the motion estimation problem in the framework of dynamic epipolar geometry, and assess how such geometry is modified under the fixation assumption. Since dynamic motion estimation schemes have been proposed in the framework of epipolar geometry [11], we modify them in order to embed the fixation assumption. As a result, we can directly compare the estimates obtained by enforcing the fixation constraint with the estimates obtained by assuming general rigid motion. We also assess analytically how (small) perturbations of the fixation constraint affects the quality of the estimates, and we perform simulation experiments in order to probe the boundaries of validity of the fixation model.

1.1 Scenario

We will consider a system with a camera mounted on a two-degrees of freedom actuated joint (the eye) standing on a platform which is moving freely (with 6 degrees of freedom) in the environment (the head), as in figure 1. The architecture of the overall system is composed of two parts: an inner control loop that actuates the eye as to maintain a given feature in the center of the image-plane or to saccade to a different fixation point given from a higher-level decision system; an estimator then reconstructs the relative motion between the eye and the object which is due to the motion of the head within the environment. These estimates can then be used in order to elaborate control actions with different tasks, such as obstacle avoidance, "optimal" estimation of structure, target pursuing etc. .

The overall functioning of the scheme can be summarized as follows (see figure 1):

1. Select features.

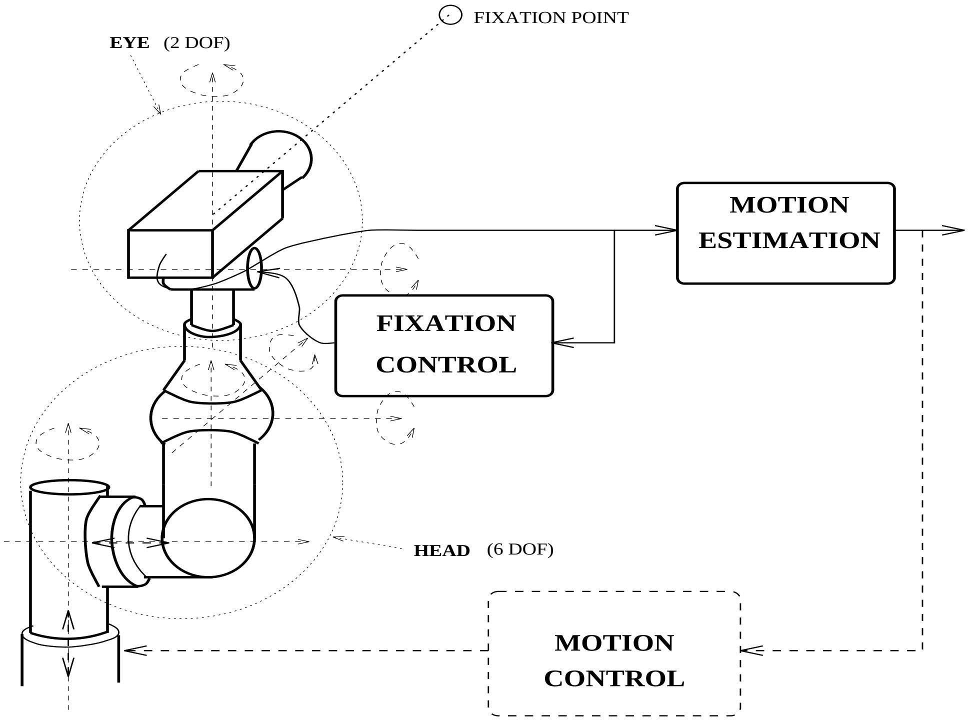

Figure 1: *Overall setup of motion from fixation: an inner tracking loop controls the two degrees of freedom of the eye as to maintain a given feature in the center of the image. The images are then fed into the motion estimation algorithm that recursively estimates the motion of the head within the environment. The estimates can possible be fed back to the head in order to accomplish different control tasks such as navigation, inspection, docking etc. (outer dashed loop).*

2. Select a target or fixation point. This could be the feature closest to the center of the image, or the best-conditioned feature, or the focus of expansion, or the singularity in the motion field or any other location assigned from a higher-level system.

3. Control the gaze of the eye to the fixation point. Simple control strategies can be implemented, such as a one-step deadbeat, or control on the sphere with exponential convergence. The kinematics and geometry of the eye mechanism must be included in the model (it will be a change of coordinates in the state-space sphere), the dynamics can be neglected in a first approximation.

4. Fine-tune fixation by shifting the origin of the image-plane.

5. Track features between successive time instants. This process (the correspondence problem) is greatly facilitated by two facts. First, since we fixate one point in the visible object, features only move little in the image, and always remain within the field of view. Second, knowledge of the motion of the camera from the actuators helps predicting the position of the features at successive frames.

6. Go to 3. (Inner, fast tracking loop).

7. Estimate relative motion between the object and the viewer. Both velocity or absolute orientation can be estimated. Check the quality of tracking.

8. Possibly take control action on the head in order to achieve specified tasks (outer loop).

We will only briefly describe the realization of the inner control loop (the "tracking" or "fixation" loop), which consists of a control system defined on a two-sphere, with measurements in the real projective plane (section 1.2). This problem is well-understood and extensive literature is available on the topic (see [4] and references therein). The rest of the paper assumes that tracking has been performed within some level of accuracy and analyizes the problem of estimating the remaining degrees of freedom. In section 2 we review the setup of epipolar geometry and show how it is modified by the fixation assumption. In section 3 we show how the epipolar representation can be used in order to formulate dynamic (recursive) estimators of motion. The fixation assumption modifies the parameter space, but not the structure of the estimator, which makes it possible to compare motion estimators embedding the fixation constraint, with estimators of general rigid motions. We present both a closed-form solution from two views and a recursive solution based upon the epipolar representation. In section 5 we describe a model for estimating absolute attitude under the fixation constraint.

While it is evident that fixation reduces the number of degrees of freedom, and therefore the estimator following the tracking loop will operate on a smaller-dimensional space and hence be more constrained, it is not trivial to assess how possible imprecisions in the tracking stage propagate onto the estimation stage. In section 4 we assess the sensitivity of the estimates with respect to the fixation constraint ,and define a measure of "goodness of tracking" that can be performed during the estimation phase.

In section 6 we substantiate our analysis with simulation experiments on noisy synthetic image sequences.

1.2 Fixation control

The task of the inner tracking loop is that of keeping a given point in the center of the image plane. Equivalently, we can enforce that a given direction (projection ray) in $\mathbb{R}^3$ coincides with the optical axis (see figure 2). In order to do so, we can act on two motors that drive the joint on top of which the camera is mounted. If we call $[\theta\ \phi]^T$ the angles at the joint which describe the local coordinates of the state s of the eye on the sphere, and u_1 and u_2 the torques applied to the motors, then the geometry, kinematics and dynamics of the eye can be described as a nonlinear dynamical system of the form:

$$\dot{s} = f(s, u) \qquad s \in \mathbf{S}^2. \tag{1}$$

If we call $\mathbf{x}_0$ the spherical coordinates of the target point in the reference centered in the optical center of the camera, with the Z-axis along the optical axis, then the motion of the camera $s(t)$ induces a vectorfield of the form

$$\dot{\mathbf{x}}_0 = g(\mathbf{x}_0, s) \qquad \mathbf{x}_0 \in \mathbf{S}^2. \tag{2}$$

However, we cannot measure directly the spherical coordinates of the target point, since it is projected on a flat image-plane, rather than on a spherical retina (figure 2). In fact, the actual measure is a local diffeomorphism

$$\begin{aligned} \pi : \mathbf{S}^2 &\rightarrow \mathbb{RP}^2 \\ \mathbf{x}_0 &\mapsto \mathbf{y}_0. \end{aligned} \tag{3}$$

Our overall dynamic model can be therefore summarized as

$$\begin{cases} \dot{s} = f(s, u) & s \in \mathbf{S}^2 \\ \dot{\mathbf{x}}_0 = g(\mathbf{x}_0, s) & \mathbf{x}_0 \in \mathbf{S}^2 \\ \mathbf{y}_0 = \pi(\mathbf{x}_0) + n_0 & \mathbf{y}_0 \in \mathbb{RP}^2 \end{cases} \tag{4}$$

where n_0 is a noise term due to the uncertainty in the tracking procedure. The goal of the inner tracking module can then be expressed as follows:

> *take the control action $u(t)$ such that $\mathbf{y}_0(t) \longrightarrow [0\ 0\ 1] \in \mathbb{RP}^2$ exponentially as $t \rightarrow \infty$.*

When we neglect the dynamics of the eye, and we assume that we are able to act on the velocity of the joints through our actuators, we can simplify our model into one of the form

$$\begin{cases} \dot{\mathbf{x}}_0 = u & \mathbf{x}_0 \in \mathbf{S}^2 \\ \mathbf{y}_0 = h(\mathbf{x}_0) + n_0 & \mathbf{y}_0 \in \mathbb{RP}^2 \end{cases} \tag{5}$$

which we can write in local coordinates, provided that $\mathbf{y}_0$ is close enough to $h(\mathbf{x}_0)$, as

$$\begin{cases} \dot{\mathbf{x}}_0 = u & \mathbf{x}_0 \in \mathbb{R}^2 \\ \mathbf{y}_0 = h(\mathbf{x}_0) + n_0 & \mathbf{y}_0 \in \mathbb{R}^2 \end{cases} \tag{6}$$

where h comprises a change of coordinates in the sphere and the perspective projection.

From the above expression it is immediate to formulate a proportional control law with exponential convergence to the target fixation point $\mathbf{y}_0$ either in the workspace,

$$u_w(\mathbf{x}, \mathbf{y}_0) = k_p \left(h^{-1}(\mathbf{y}_0) - \mathbf{x} \right), \tag{7}$$

or in the output space, represented for simplicity as the two-sphere

$$u_o(\mathbf{x}, \mathbf{y}_0) = J_h(\mathbf{x}) k_p v_G(\mathbf{x}, \mathbf{y}_0) \tag{8}$$

where k_p is the proportional constant, J_h is the jacobian of h:

$$J_h(\mathbf{x}) \doteq \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}) \tag{9}$$

and v_G is the geodesic versor

$$v_G(\mathbf{x}, \mathbf{y}_0) = \frac{(h(\mathbf{x}) \wedge \mathbf{y}_0) \wedge h(\mathbf{x})}{d} \tag{10}$$

with $d = \arccos(< h(\mathbf{x}), \mathbf{y}_0 >)$ the distance between the output and the target along the geodesic [4].

Exponential convergence is required as a mean of contrasting noise. In fact, if the control is fast, it can dump disturbances at a rate faster than they arrive, which helps the system not to diverge in the presence of noise and disturbances. The above controls can be easily shown to generate exponential convergence to the desired goal [4].

1.3 Tracking and shift registration

The purpose of the eye motion control is to keep a prescribed feature at the origin of the image plane using two degrees of freedom of the spherical joint of the eye. In principle, tracking of the target feature could be accomplished *locally* by shifting the origin of the image-plane at each step, provided that the feature remains within the field of view (see figure 2). In general, a combination of the two techniques is to be employed. The eye is rotated in order to maintain the target feature as close as possible to the center of the image, then the image plane is shifted, with a purely "software" operation, in order to translate the origin of the image-plane on the target feature. Provided that the feature tracking scheme achieves sub-pixel accuracy [2], the shift-registration allows us to perform the tracking within one pixel accuracy on the image-plane.

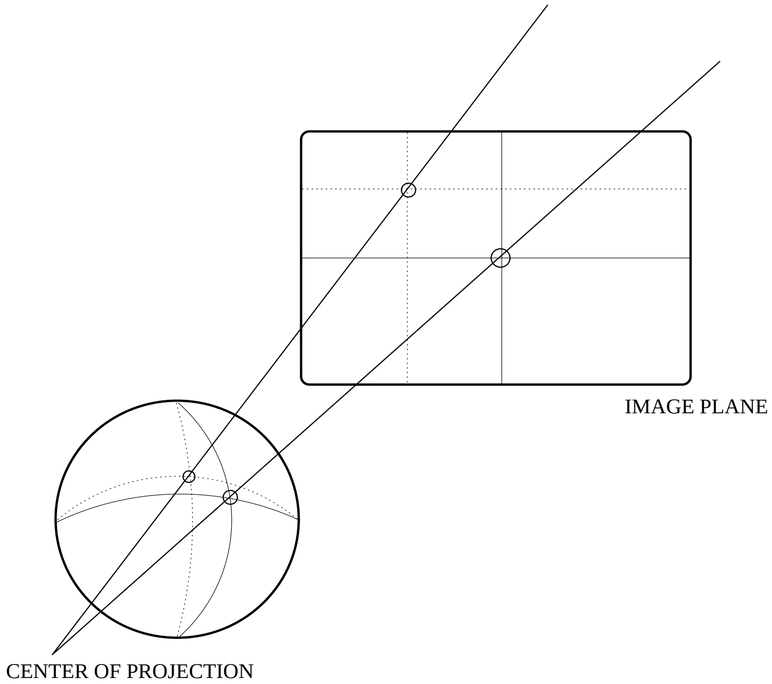

Figure 2: *Tracking amounts to controlling the camera as to bring one specified feature-point in the origin of the image plane. The same task can be accomplished locally by shifting the image-plane, a purely software operation. The two operations are equivalent locally to the extent in which the target feature does not exit the field of view.*

2 Epipolar geometry under fixation

In the present section we analyze the functioning of the second stage of the scheme depicted in figure 1, which consists of estimating the relative motion between the viewer and the object being fixated. Since one point of the object is still in the image plane, the object is free only to rotate about this point, and to translate along the fixation line. Therefore there are overall 4 degrees of freedom left from the fixation loop.

We start off with studying how the well-known setup of the epipolar geometry is transformed under the fixation conditions.

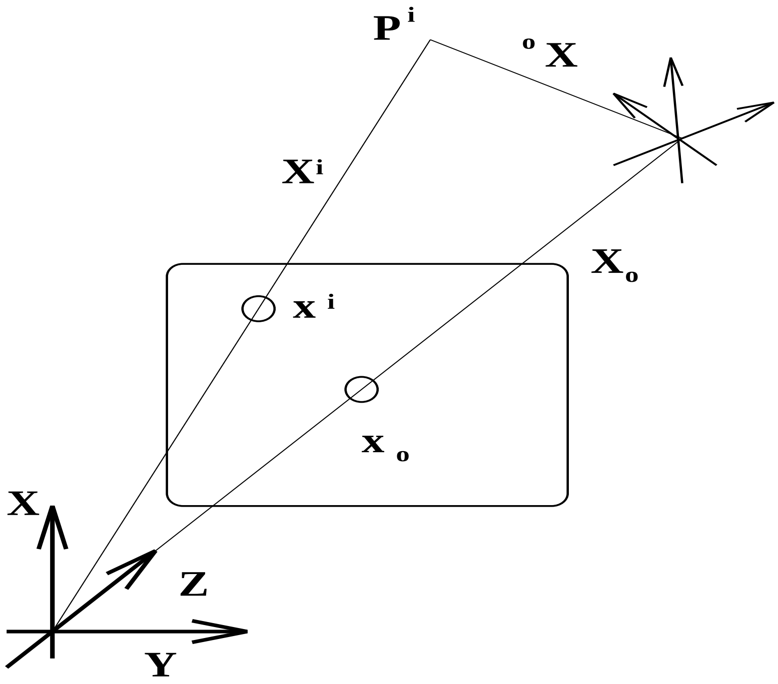

Figure 3: *Imaging geometry. The viewer-reference is centered in the center of projection, with the Z-axis pointing along the ptical axis. The object reference frame is centered in the fixation point. Under the fixation conditions the object can only rotate about the fixation point and translate along the fixation axis.*

2.1 Notation

We call $\mathbf{X} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T \in \mathbb{R}^3$ the coordinates of a generic point $\mathbf{P}$ with respect to an orthonormal reference frame centered in the center of projection, with Z along the optical axis and X, Y parallel to the image plane and arranged as to form a right-handed frame (see figure 3). The relative attitude between the camera and the object (or scene) is described by a **rigid motion** $g \in SE(3)$.

$$\begin{cases} \mathbf{P}^i(t) = {}^t g_o {}^o\mathbf{P}^i \\ \mathbf{P}^i(t+1) = {}^{t+1} g_o {}^o\mathbf{P}^i \end{cases} \Rightarrow \ \mathbf{P}(t+1) = {}^{t+1} g_o {}^t g_o^{-1} \mathbf{P}(t) \tag{11}$$

where ${}^\tau g_o \in SE(3)$ is the change of coordinates between the viewer reference frame at time τ and the object coordinate frame centered in the fixation point $\mathbf{P}_0(t) = [0 \ 0 \ d(t)]^T$. Since we are interested in the displacement relative to the moving frame (ego-motion), we can assume that the object reference is aligned with the viewer reference at time t, so that we can write the relative orientation between time t and $t+1$ in coordinates as

$$\mathbf{X}^i(t+1) = R(t)\left(\mathbf{X}^i(t) - \begin{bmatrix} 0 \\ 0 \\ d(t) \end{bmatrix}\right) + \begin{bmatrix} 0 \\ 0 \\ d(t+1) \end{bmatrix} \tag{12}$$

which we will write as

$$\mathbf{X}^i(t+1) = R(t)\mathbf{X}^i(t) + d(t)T(R,v) \tag{13}$$

where

$$T(R,v) \doteq \begin{bmatrix} -R_{13} \\ -R_{23} \\ -R_{33}+v \end{bmatrix} \tag{14}$$

and

$$v \doteq \frac{d(t+1)}{d(t)} \neq 0 \tag{15}$$

is the relative velocity along the fixation axis. The matrix $R \in SO(3)$ is an orthonormal rotation matrix that describes the change of coordinates between the viewer's reference at time t and that at time $t+1$ relative to the object. $T \in \mathbb{R}^3$ describes the translation of the origin of the viewer's reference frame.

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

$$\begin{aligned} \pi : \mathbb{R}^3 - \{0\} \quad &\rightarrow \quad \mathbb{RP}^2 \\ \mathbf{X} \quad &\mapsto \quad \mathbf{x} \doteq \begin{bmatrix} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{bmatrix}^T. \end{aligned} \tag{16}$$

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

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

Due to the fixation constraint, the camera is only allowed to translate along the fixation axis, rotate about the fixation axis (cyclorotation) and move on a sphere centered in the fixation point with radius equal to the distance from the fixation point to the optical center. Therefore there are 4 degrees of freedom in the velocity. These can also be easily seen from the object reference frame: the object reference is free to rotate about the fixation point (3 degrees of freedom) but can only translate along the fixation axis (1 degree of freedom).

In eq. (13), these 4 degrees of freedom are encoded into $R(t)$ (3 DOF) and $v(t)$ (1 DOF). However, note that also the distance from the fixation point $d(t)$ enters the model. The epipolar constraint, which will be derived in the next subsection, involves only relative orientation and measured projections, while it gets rid of the 3-D structure and of the absolute distance d.

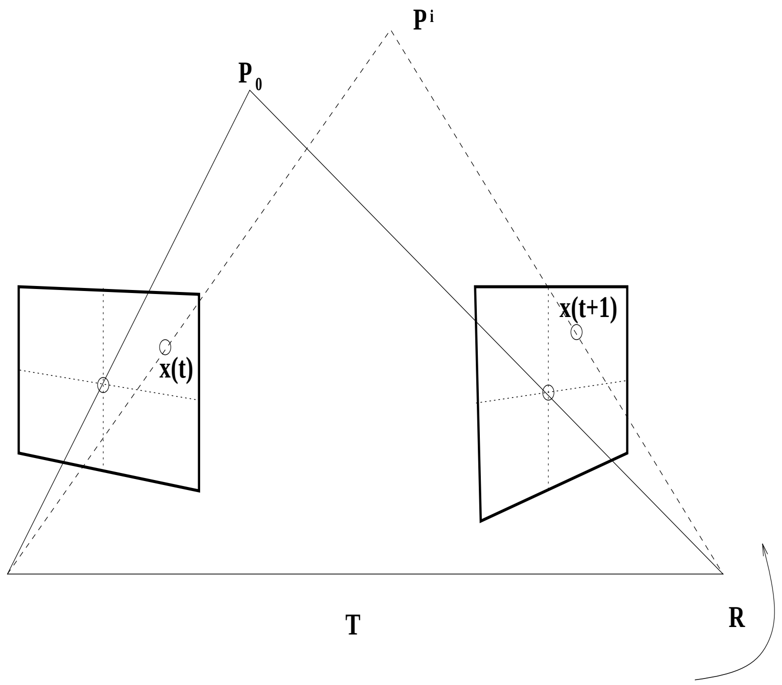

Figure 4: *Coplanarity constraint: the coordinates of each point in the reference of the viewer at time t, the coordinates of the same point at time t+1 and the translation vector are coplanar.*

2.2 Coplanarity constraint

The well-known coplanarity constraint (or "epipolar constraint", or "essential constraint") of Longuet-Higgins [8] imposes that the vectors $T(R(t), v(t))$, $\mathbf{X}^i(t+1)$ and $\mathbf{X}^i(t)$ be coplanar for all t and for all points $\mathbf{P}^i$ (figure 4). The triple product of the above vectors is therefore zero; if we multiply both sides of (13) by $\alpha \mathbf{X}^i(t+1)^T(T\wedge)$, where $\alpha \in \mathbb{R} - \{0\}$, we get

$$0 = \mathbf{X}^i(t+1)(T\wedge)R(t)\mathbf{X}^i(t) \tag{17}$$

which we will write as

$$\mathbf{X}^i(t+1)\mathbf{Q}(t)\mathbf{X}^i(t) = 0 \tag{18}$$

with

$$\mathbf{Q}(t) \doteq \mathbf{Q}(R(t), v(t)) = (T\left(R(t), v(t)\right)) \wedge R(t). \tag{19}$$

We will use the notation $\mathbf{Q}(t)$ when emphasizing the time-dependence, while we will use $\mathbf{Q}(R, v)$ when stressing the dependence of $\mathbf{Q}$ from the 3 rotation parameters contained in R and from the relative velocity along the fixation axix v. Note that $\mathbf{Q}$ is an element of a 4-dimensional differentiable manifold which is embedded in $\mathbb{R}^9$, since $\mathbf{Q}$ is realized as a 3×3 matrix.

Since the coordinates of each point $\mathbf{X}^i(t)$ and their projective coordinates $\mathbf{x}^i(t)$ span the same direction in $\mathbb{R}^3$, the constraint (18) holds for $\mathbf{x}^i$ in place of $\mathbf{X}^i$ (just divide eq (18) by $\mathbf{X}^i_3(t+1)\mathbf{X}^i_3(t)$):

$$\mathbf{x}^i(t+1)\mathbf{Q}(t)\mathbf{x}^i(t) = 0 \qquad \forall t\ ,\ \forall i. \tag{20}$$

2.3 Structure of the essential manifold

For a generic $T \in \mathbb{R}^3$ and a rotation matrix R, the matrix $\mathbf{Q} = (T\wedge)R$ belongs to the so-called "essential manifold"

$$\mathbf{E} \doteq \{SR \mid S \in so(3),\ R \in SO(3)\}, \tag{21}$$

which can be characterized as the tangent bundle to the rotation group $TSO(3)$ [11]. Under the fixation constraint, T has a special structure which restricts $\mathbf{Q}$ to a submanifold of the essential manifold. In this section we study the geometry of such a submanifold induced by the fixation constraint. We have already seen that the dimension of the space reduces from 6 down to 4, since two degrees of freedom are used in order to keep the projection of the fixation point still in the image plane.

After some simple algebra, it is easy to see that

$$\mathbf{Q}(R, v) = RS^T + vSR \tag{22}$$

where

$$S \doteq \begin{bmatrix} 0 & -\alpha & 0 \\ \alpha & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{23}$$

and α is an unknown scaling factor due to the homogeneous nature of the coplanarity constraint. If we restrict the essential matrices $\mathbf{Q} \in \mathbf{E}$ to have unit norm (as in the definition of the "normalized essential manifold" [11]), then α is fixed to be $\alpha = \frac{1}{\|\mathbf{Q}\|}$. Note that this arbitrary scaling does not affect neither the relative velocity v (which is already a scaled parameter) nor the rotation matrix R. We will see in section 2.4 that $\alpha = \frac{1}{\|\mathbf{Q}\|}$ is a necessary choice in order to avoid singularities in the representation. Under the fixation constraint, both the essential manifold $\mathbf{Q}$ and its normalized version $\frac{\mathbf{Q}}{\|\mathbf{Q}\|}$ belong to a four-dimensional submanifold of the essential manifold $\mathbf{E}$. The essential matrix is therefore defined, under the fixation constraints, by the Sylvester's equation (22), with strongly structured unknowns $R \in SO(3)$ and $v \in \mathbb{R}$. Other equivalent expressions can be derived as follows, assuming $\alpha = 1$:

$$\mathbf{Q} = \left(RS^T R^T + vS\right) R \tag{24}$$

$$\mathbf{Q} = \left(R \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + v \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \wedge R \tag{25}$$

$$\mathbf{Q} = \begin{bmatrix} -R_{.2} & | & R_{.1} & | & 0 \end{bmatrix} + v \begin{bmatrix} - & -R_{2.} & - \\ - & R_{1.} & - \\ - & 0 & - \end{bmatrix} \tag{26}$$

$$\mathbf{Q} = \begin{bmatrix} -R_{12} - vR_{21} & R_{11} - vR_{22} & -vR_{23} \\ -R_{22} + vR_{11} & R_{21} + vR_{12} & vR_{13} \\ -R_{32} & R_{31} & 0 \end{bmatrix}. \tag{27}$$

Another useful way of writing the epipolar constraint can be derived as follows. Since the constraints (20) are *linear* in the components of the essential matrix $\mathbf{Q}$, we can reorder them as

$$\chi(t)\mathbf{Q} = 0 \tag{28}$$

where $\chi(t)$ is a $N \times 9$ matrix which depends on the measurements $\mathbf{x}_i(t), \mathbf{x}_i(t+1)$ whose generic row can be written as

$$\chi_{i.} = \left[\begin{array}{ccccccccc} \mathbf{X}_1^i(t+1)\mathbf{X}_1^i(t) & \mathbf{X}_1^i(t+1)\mathbf{X}_2^i(t) & \mathbf{X}_1^i(t+1) & \mathbf{X}_2^i(t+1)\mathbf{X}_1^i(t) & \mathbf{X}_2^i(t+1)\mathbf{X}_2^i(t) & \mathbf{X}_2^i(t+1) & \mathbf{X}_1^i(t) & \mathbf{X}_2^i(t) & 1 \end{array} \right] \tag{29}$$

$\mathbf{Q}$ is now interpreted as a 9-dimensional column vector obtained by stacking the rows of $\mathbf{Q}$ one on top of each other. It is easy to verify that the above can be written as follows:

$$\chi(t)\mathcal{S}(v)R = 0 \tag{30}$$

where

$$\mathcal{S}(v) \doteq \left[\begin{array}{ccc} S & -vI & 0 \\ vI & S & 0 \\ 0 & 0 & S \end{array} \right] \tag{31}$$

is a skew-symmetric, 9×9 matrix with rank 8 which depends only upon the translational velocity v. I is the 3-dimensional identity matrix and R is the usual rotation matrix now interpreted as a nine-dimensional column vector obtained by stacking the rows of R on top of each other. We will not make a distinction between 3×3 matrices and $9-$dimensional column vectors, whenever it is clear from the context which representation is employed. Since both the last row and the last column of $\mathcal{S}$ are identically zero, we can delete them along with the last column of χ and the last element of R, which is now interpreted as a 8-dimensional column-vector.

From the above characterizations of the essential matrix constrained by the fixation hypothesis it is possible to draw some interesting conclusions. In particular, by left-multiplying the above equation by [0 0 1], we anihilate the second (rightmost) term of the right hand-side of (22), while the column vector $[0\ 0\ 1]^T$ anihilates the leftmost term, if right-multiplied. From this simple observation we can derive a necessary condition which acts as a consistency check for the quality of fixation:

$$\mathbf{Q}_{33} = 0. \tag{32}$$

In general, from a number of point matches, we can derive an approximate estimate of the matrix $\frac{\mathbf{Q}}{\|\mathbf{Q}\|}$ which, due to noise, will be such that $\mathbf{Q}_{33} \neq 0$; later in section 4 we will see how $|\mathbf{Q}_{33}|$ gives a measure of how accurate the inner tracking loop is.

2.4 Singularities and normalization of the epipolar representation

In the characterizations of the essential matrices described in the previous section, the unknown scaling factor has been taken into account by fixing the scalar $\alpha = 1$, and therefore the matrix $\mathbf{Q}$ is uniquely defined. However, there is a continuum of possible motions which correspond to the essential matrix

$$\mathbf{Q}(v, \Omega) = 0 \tag{33}$$

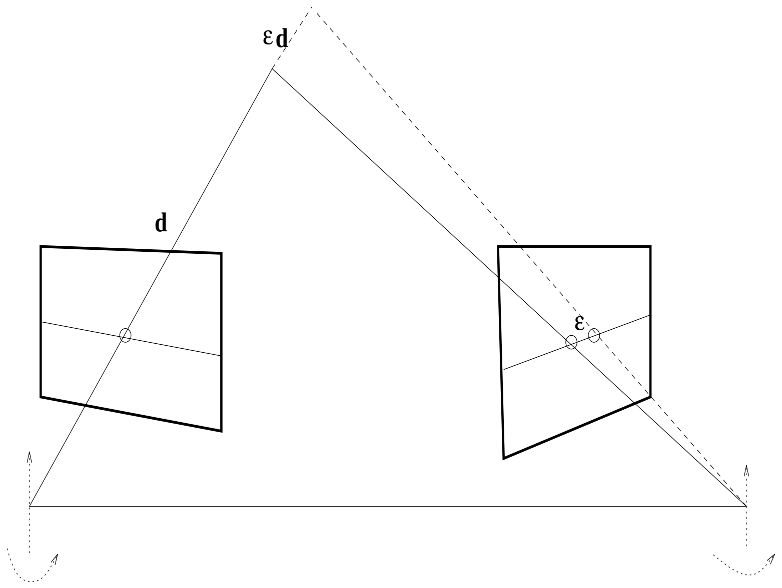

Figure 5: *Epipolar setup. Under the fixation constraint, both the centers of projection at time t and t+1, and the optical centers of the two cameras lie on the same plane, the epipolar plane. The intersection of the epipolar plane with the image planes is the epipolar line. The epipolar plane is invariant after fixation, for the camera can only translate along the plane, and rotate about a direction orthogonal to it.*

in particular

$$v = 1 \;\; \Omega = \begin{bmatrix} 0 \\ 0 \\ \theta \end{bmatrix} \;\; | \; \theta \in [0, \pi) \;\; \Rightarrow \;\; \mathbf{Q}(v, \Omega) = 0 \tag{34}$$

since $\mathbf{Q} = (T\wedge)e^{\Omega\wedge}$ with $T = 0$, and therefore all motions consisting of pure cyclorotation (rotation about the optical axis or fixation axis) generate a zero essential matrix or an undefined normalized essential matrix.

If we know that motion occurs only about the optical axis, we can easily estimate the amount of rotation θ by solving in a lest-squares sense the rigid motion equations (12), which reduce, in the case of pure cyclorotation, to

$$\mathbf{x}_i(t+1) = e^{\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\wedge\theta} \mathbf{x}_i(t). \tag{35}$$

In order to get rid of the singularity just mentioned, we need to normalize the essential matrices. Since the epipolar constraint is defined up to a scale, it can be arbitrarily multiplied by a constant. In particular, if we multiply it by $\frac{1}{\|\mathbf{Q}\|}$ we get rid of the singularity, since the translation vector T is constrained to be of unit norm. Note that we do not loose any degree of freedom in the representation, for the scaling does not affect the motion parameters v, Ω.

In section 3.3 we will see that this representation affects the convergence of the filter for estimating motion when away from the singular configuration. When the object purely rotates about the optical axis, the translation vector is undefined; we will see in section 3.3 how it is possible to sort out this situation.

3 Estimation from the epipolar constraint

The epipolar constraint, with the addition of the fixation assumption, can be used in order to estimate the 4 free parameters (three for rotation and one for relative translation along the fixation axis). The first solution we propose is a closed-form solution which is correct in the absence of noise, but is far from being efficient in the presence of uncertainty, since the structure of the epipolar constraint is not imposed in the estimation.

The second solution is a more correct one, for it enforces the structure of the epipolar constraint during the estimation. It consists of a dynamic estimator in the local coordinates of the essential manifold. The constraints are enforced by construction and the structure of the parameter manifold is exploited, while the computation is carried out by an Implicit Extended Kalman Filter (IEKF) in the lines of [11].

3.1 Closed-form, two-frames solutions

Consider N visible points $\mathbf{P}^i$, $\forall i = 1 \dots N$, and the N corresponding scalar constraints (20). The constraints are *linear* in the components of $\mathbf{Q}$, and can be used for estimating a *generic* 3×3 matrix $\hat{\mathbf{Q}}$ which is least-squares compatible with the measurements, in the same way as [8, 13, 11].

Once the matrix $\hat{\mathbf{Q}}$ has been estimated, we can derive a set of constraints for the components of the rotation matrix R. Just for the sake of simplicity, assume that we represent the rotation matrix *locally* using Euler angles $\alpha \neq 0, \beta \neq 0$ and $\gamma \neq 0$:

$$R = R_Z(\alpha)R_Y(\beta)R_Z(\gamma) = \begin{bmatrix} c_\alpha c_\beta c_\gamma - s_\alpha s_\gamma & -c_\alpha c_\beta s_\gamma - s_\alpha c_\gamma & c_\alpha s_\beta \\ s_\alpha c_\beta c_\gamma + c_\alpha s_\gamma & -s_\alpha c_\beta s_\gamma + c_\alpha c_\gamma & s_\alpha s_\beta \\ -s_\beta c_\gamma & s_\beta s_\gamma & c_\beta \end{bmatrix} \tag{36}$$

where $R_Z(\alpha)$ indicates a rotation about the $Z-$axis of α radiants

$$\begin{bmatrix} c_\alpha & -s_\alpha & 0 \\ s_\alpha & c_\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{37}$$

and similarly for $R_Y(\beta)$ and $R_Z(\gamma)$. From the above expression of R, and the expression for $\mathbf{Q}$ given in eq. (27), it is immediate to solve for the Euler angles:

$$\alpha = \arctan\left(-\frac{\mathbf{Q}_{13}}{\mathbf{Q}_{23}}\right) \tag{38}$$

$$\beta = \arcsin\sqrt{\mathbf{Q}_{31}^2 + \mathbf{Q}_{32}^2} \tag{39}$$

$$\gamma = \arctan\left(\frac{\mathbf{Q}_{31}}{\mathbf{Q}_{32}}\right) \tag{40}$$

$$\tag{41}$$

provided that $\mathbf{Q}_{23} \neq 0$ and $\mathbf{Q}_{32} \neq 0$. It is immediate to see that $\mathbf{Q}_{23} = \mathbf{Q}_{23} = 0$ only if rotation occurs only about the optical axis with an angle $\theta = \alpha + \gamma$. In such a case, equation (27) becomes

$$\mathbf{Q} = \begin{bmatrix} s_\theta(1-v) & c_\theta(1-v) & 0 \\ -c_\theta(1-v) & s_\theta(1-v) & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{42}$$

and we can solve for θ

$$\theta = \gamma + \alpha = \arctan\left(\frac{\mathbf{Q}_{22}}{\mathbf{Q}_{12}}\right) \tag{43}$$

provided that $\mathbf{Q}_{12} \neq 0$, in which case we have $\alpha = \beta = \gamma = 0$. Once the rotation parameters have been estimated, the translation parameter v can be recovered from the other elements of $\mathbf{Q}$. For instance, when $\beta = 0$,

$$v = 1 - \sqrt{\mathbf{Q}_{11}^2 + \mathbf{Q}_{21}^2}. \tag{44}$$

Alternatively, one may start with a different local coordinate parametrization of R, for example the exponential coordinatization

$$R = e^{\Omega\wedge} \tag{45}$$

and plug the result into equation (22), which can then be solved for the three unknowns $\Omega_1 \dots \Omega_3$ using an iterative optimization method such as a gradient descent.

It must be stressed that these methods do not enforce the structure of the parameter space during the estimation process. Rather, generic, non-structured parameters are estimated, and then their structure is imposed a-posteriori in order to recover an approximation of the desired estimates.

The epipolar constraints can also be used for formulating nonlinear filters that estimate the motion components over time, while taking into account the geometry of the parameter space. This is done in the next section.

3.2 Implicit dynamical filter for motion from fixation

Consider the local parametrization of the essential matrix $\mathbf{Q}(R, v)$, which is

$$\xi \doteq \begin{bmatrix} \Omega \\ v \end{bmatrix} \in \mathbb{R}^4 \tag{46}$$

where $\Omega \in \mathbb{R}^3$ is defined for $\|\Omega\| \in [0, \pi)$ by the equation [9]

$$e^{\Omega\wedge} \doteq R. \tag{47}$$

We can write a dynamic model in the local coordinates of the essential manifold, having as implicit measurement constraints the epipolar constraint (20) where the matrix $\mathbf{Q}$ is now expressed as a function of the local coordinates, $\mathbf{Q}(\xi)$:

$$\begin{cases} \mathbf{x}^i(t+1)^T\mathbf{Q}(\xi(t))\mathbf{x}^i(t) = 0 & \xi \in \mathbb{R}^4 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) & \forall i = 1 \dots N. \end{cases} \tag{48}$$

Estimating motion amounts to identifying the parameters ξ from the above model. This can be done using the local identification procedure presented in [11], which is the IEKF based upon the model

$$\begin{cases} \xi(t+1) = \xi(t) + n_\xi(t) \\ \mathbf{y}^i(t+1)^T\mathbf{Q}(\xi(t))\mathbf{y}^i(t) = \tilde{n}_i(t) & \forall i = 1 \dots N \end{cases} \tag{49}$$

where the second order statistic of the residual $\tilde{n}$ is computed according to [11]. An alternative way of writing the above model is

$$\begin{cases} \xi(t+1) = \xi(t) + n_\xi(t) \\ \chi(t)\mathcal{S}(\xi_1)R(\xi_2, \xi_3, \xi_4) = 0. \end{cases} \tag{50}$$

the equations of the estimator, as derived from [11], are:

prediction step:

$$\hat{\xi}(t+1|t) \;=\; \hat{\xi}(t|t) \qquad \hat{\xi}(0|0) = \xi_0 \tag{51}$$

$$P(t+1|t) \;=\; P(t|t) + Q_\xi \tag{52}$$

where Q_ξ is the variance of the noise n_ξ driving the random walk model and is intended as a tuning parameter, and P is the variance of the estimation error of the filter.

update step:

$$\hat{\xi}(t+1|t+1) \;=\; \hat{\xi}(t+1|t) + L(t+1)\begin{bmatrix} \vdots \\ \mathbf{y}^i(t+1)^T\mathbf{Q}(\hat{\xi}(t+1|t))\mathbf{y}^i(t) \\ \vdots \end{bmatrix} \tag{53}$$

$$P(t+1|t+1) \;=\; \Gamma(t+1)P(t+1|t)\Gamma(t+1)^T + L(t+1)R_nL^T(t+1) \tag{54}$$

where $L(t+1)$ is the Extended Kalman Gain [7], and $\Gamma = I - LC$, with $C \doteq \frac{\partial \mathbf{Q}}{\partial \xi}\Big|_{\hat{\xi}(t+1|t)}$.

3.3 Dealing with singularities in the representation

In section 2.4 we have pointed out a singularity in the non-normalized epipolar representation when the relative motion between the scene and the object consists of pure rotation about the optical axis. This phenomenon is to be expected, for pure rotation about the optical axis generates zero ego-motion translation

$$T = -R_{.3} + \begin{bmatrix} 0 \\ 0 \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{55}$$

which is a singular configuration for the motion estimation problem [11]. As long as there is a non-zero translation (that is, as long as there is some components of rotation about an axis non corresponding to the optical axis), the constraints are well-defined. However, serious problems may occur while estimating motion even when the motion parameters are far away from the singular point.

In order to visualize that, we can imagine the innovation of the filter as living on a residual surface that maps some particular motion v, Ω onto $\mathbb{R}^N$ when N feature points are visible. The filter will try to update the state $\hat{v}, \hat{\Omega}$ as to reach the minimum of the residual. Of course the motion that generated the data v, Ω corresponds to a minimum of the residual surface (it would be zero in absence of noise). However, also the location $v = 1,\ \Omega = [0\ 0\ \theta]^T$ corresponds to a zero of the residual, which is a hole in the residual surface. Therefore the filter must be able to reach the minimum without falling into the singularity (see figure 6).

This can be done provided that the initial conditions are close to the minimum of the residual surface corresponding to the true motion. However, in the presence of high measurement noise levels, the residual surface becomes increasingly more irregular, and eventually the filter falls into the singularity. This effect will be illustrated in the experimental section, where we will show that in the presence of high noise levels, the filter initialized fare enough from the true value of the state falls into the singularity, the innovation goes to zero and the variance of the state increases.

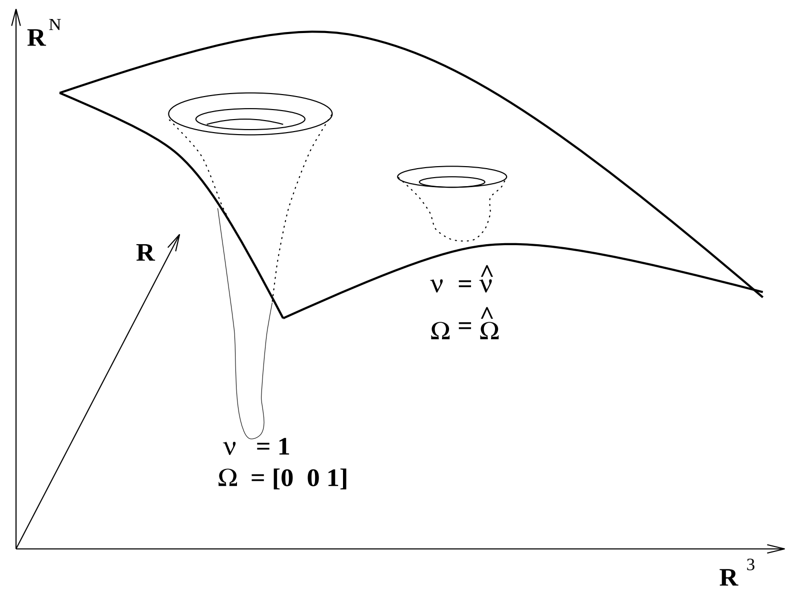

Figure 6: *Singularity in the non-normalized epipolar representation. The residual surface, where the innovation of the filter takes values, has a minimum corresponding to the true motion, but also a minimum corresponding to cyclorotation. The filter must be able to converge to the true minimum without falling into the singularity. The normalized epipolar representation is a way of getting rid of the singularity, for the translation vector is constrained to having unit norm.*

One way of getting rid of this singularity is to use the normalized essential matrix, which corresponds to dividing the epipolar constraint by the norm of translation. This eliminates the singularity, since T is constrained to having unit norm. However, the motion corresponding to pure cyclorotation gives an essential matrix which is undefined, and therefore the filter will give arbitrary estimates.

In order to sort out the case of pure rotation about the optical axis, we can first try to fit a θ into the purely cyclo-rotational model

$$\mathbf{x}(t+1) = R_Z(\theta)\mathbf{x}(t). \tag{56}$$

If the residual is big enough it means that rotation is not purely about the optical axis. Therefore the translation induced in the viewer's reference is non-zero, and the normalized epipolar constraint is well-defined. We will see in the experimental section how the filter based upon the normalized epipolar representation performs where the non-normalized filter would fall into the singularity.

4 Vergence control, quality of fixation and sensitivity of constraints

One may argue that, in the proposed architecture, the estimation scheme that follows the fixation loop is "blind", in the sense that it cannot reject disturbances due to imperfect tracking. In the present section we analyze how the estimation algorithm is modified in the presence of non-perfect tracking, and how it can assess the quality of the fixation.

We will consider two different kinds of non-perfect tracking. One in which the two optical axes (at time t and $t+1$) intersect at a point which is not the desired fixation point, and one in which the two optical axes do not intersect at all.

4.1 Vergence control

Let us assume that the optical axis of the camera at time t intersects the optical axis at time $t+1$ in a "vergence point" which is different from the desired fixation point (see figure 5). Consider the plane determined by the two centers of projection and the optical center (fixation point) in the camera at time t, which is called the epipolar plane at time t. If the optical axes intercept, there must exist one point on the projection of the optical axis of the camera at time t which passes through the optical center of the camera at time $t+1$. Equivalently, the optical center at time $t+1$ must belong to the epipolar plane. It is immediate to see that this can happen only if the direction of rotation is orthogonal to the direction of translation, which is constrained to belong to the epipolar plane (see fig. 5). In brief, the epipolar plane is invariant under the vergence conditions.

Therefore, under the vergence conditions, we can identify one point $\mathbf{P}_0$ at the intersection of the optical axes, for which the fixation constraint is satisfied, although it is not the desired fixation point. From Chasles' theorem [9] we can conclude that the algorithm proposed in the previous section estimates the motion of the object relative to the point $\mathbf{P}_0$, rather than relative to the desired fixation point. If the mismatch between the target point and the actual vergence point is ϵ along the epipolar line, then the mismatch along the optical axis is approximately $d\epsilon$, where d is the distance between the optical center and the target fixation point.

A natural question to ask at this point is how the algorithm following the fixation loop can verify whether the vergence conditions are satisfied and, if they are not, send a feedback signal to the fixation loop.

4.2 Vergence conditions, quality of fixation

When the optical axes do not intercept, the epipolar constraint is not satisfied for the optical center. The vergence constraint between two time instants can be expressed by saying that

> the two optical axes intersect $\Leftrightarrow \exists\ \mathbf{X}_0$ such that $\mathbf{x}_0(t) = [0\ 0\ 1]^T \Rightarrow \mathbf{x}_0(t+1) = [0\ 0\ 1]^T$.

It is immediate to verify that the above conditions hold if and only if the direction of translation or orthogonal to the direction of rotation. Indeed, a more synthetic condition that can be derived by observing that

> the optical axes intersect $\Leftrightarrow \mathbf{Q}_{33} = 0$.

In fact, clearly if the optical axes intersect, the optical center $\mathbf{x}_0$ must satisfy the epipolar constraint:

$$\mathbf{x}_0(t+1)^T \mathbf{Q} \mathbf{x}_0(t) = 0 \ \Rightarrow \mathbf{Q}_{33} = 0. \tag{57}$$

Vice-versa, assume that $\forall \mathbf{x}_0$, the condition $\mathbf{x}_0(t+1) \neq [0\ 0\ 1]^T$ implies $\mathbf{x}_0(t) \neq [0\ 0\ 1]^T$ while $\mathbf{Q}_{33} = 0$. Write $\mathbf{x}_0(t+1)$ as $[\alpha\ \beta\ 1]^T$ with $\alpha\beta \neq 0$. Then the epipolar constraint must be violated for all correspondence points of the form $[0\ 0\ 1]^T$:

$$[\alpha\ \beta\ 1]\mathbf{Q}[0\ 0\ 1]^T \neq 0 \ \Rightarrow \alpha\mathbf{Q}_{13} + \beta\mathbf{Q}_{23} + \mathbf{Q}_{33} \neq 0. \tag{58}$$

If $\mathbf{Q}_{13} = \mathbf{Q}_{23} = 0$, then we conclude that $\mathbf{Q}_{33} \neq 0$, from which the contradiction. If at least one of $\mathbf{Q}_{13}, \mathbf{Q}_{23} \neq 0$, by choosing $\alpha = -\mathbf{Q}_{23}, \beta = \mathbf{Q}_{13}$, we conclude again $\mathbf{Q}_{33} \neq 0$, which contradicts the hypotheses.

Therefore, when the vergence conditions are not satisfied and the optical axes do not intersect, the scalar $|\mathbf{Q}_{33}|$ is a measure of the quality of vergence. From a geometrical point of view, $\mathbf{Q}_{33}$ is the volume of the parallelepiped with sides equal to the translation vector, the optical axis of the camera at time t and the one at time $t+1$.

Since at each step we can estimate the matrix $\mathbf{Q}$ from all the visible points, we could use $\mathbf{Q}_{33}$ as a sensory signal to be fed-back to the fixation loop. This would allow us to design a vergence control that exploits all the visible features, rather than the projection of the fixation point alone. This issue is not explored in the present paper and is an object of future research.

4.3 Sensitivity and degradation of the constraint

In the previous sections we have treated the problem of motion estimation as an identification task where the class of models was determined by the epipolar constraint under the fixation assumption. We now want to ask ourselves: suppose the actual process generating the date does not exactly fall within the given class of models, how do small deviations from the class affect the quality of the estimates?

More specifically, suppose that our camera is not tracking exactly the fixation point. The measurements we get from the image plane do not satisfy the epipolar constraint of eq. (22) for any choice of the parameters. However, if the deviation from the constraints is small, we would like our estimates to deviate little from the true motion parameters.

Suppose that our measurements are generated for an object which rotates about the fixation point with Ω, translates along the fixation axis by v and also drifts away from the fixation point with some velocities ϵ_1 and ϵ_2 along X and Y respectively. Therefore the model generating the data looks like

$$\mathbf{X}^i(t+1,\epsilon) = R(\Omega)\left(\mathbf{X}^i(t) - \begin{bmatrix} 0 \\ 0 \\ d(t) \end{bmatrix}\right) + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ d(t+1) \end{bmatrix} \tag{59}$$

where we measure

$$\mathbf{x}^i(t,\epsilon) = \pi(\mathbf{X}^i(t,\epsilon)) \tag{60}$$

which we collect into the matrix

$$\chi(t,\epsilon) \tag{61}$$

as in equation (29). For $\epsilon = 0$ the epipolar constraint is satisfied by the actual motion parameters v, Ω:

$$\chi(t,0)S(v)R(\Omega) = 0 \tag{62}$$

where S and R are a 9×9 matrix and a $9-$vector defined as in (30). However, in the presence of disturbances ϵ, there is no element in the class of models that satisfies the constraints, i.e.

$$\forall \epsilon > 0, \ \Rightarrow \ \chi(t,\epsilon)S(\tilde{v})R(\tilde{\Omega}) \neq 0 \ \forall \ \tilde{v} \in \mathbb{R} \ \forall \ \tilde{\Omega} \in \mathbb{R}^3. \tag{63}$$

At this point, assuming ϵ small, we may seek for the perturbations $\tilde{v} = v - \delta v$ and $\tilde{\Omega} = \Omega - \delta\Omega$ that make the above residual zero up to second order terms:

$$\delta v, \ \delta\Omega = \arg\min \|\chi(t,\epsilon)S(v-\delta v)R(\Omega - \delta\Omega)\|. \tag{64}$$

This is essentially the task of the recursive filter described in the previous sections, where the process to be minimized is the innovation. Expanding around the zero-perturbation conditions, we have

$$\begin{aligned} \chi(t,\epsilon)S(v-\delta v)R(\Omega-\delta\Omega) \ &= \ \chi(t,0)S(v)R(\Omega) + \frac{\partial\chi}{\partial\epsilon_1}S(v)R(\Omega)\epsilon_1 + \frac{\partial\chi}{\partial\epsilon_2}S(v)R(\Omega)\epsilon_2 + \\ -\chi(t,0)\frac{\partial S}{\partial v}(v)R(\Omega)\delta v \ &- \ \chi(t,0)S(v)\frac{\partial R}{\partial\Omega}(\Omega)\delta\Omega + \mathcal{O}(\epsilon,\delta v,\delta\Omega). \end{aligned} \tag{65}$$

We can now find the perturbations $\delta v = \delta v(\epsilon, v, \Omega)$ and $\delta\Omega = \delta\Omega(\epsilon, v, \Omega)$ that make the residual zero up to higher order terms from

$$\left[\begin{array}{cc} \frac{\partial \chi}{\partial \epsilon_1} SR & \frac{\partial \chi}{\partial \epsilon_2} SR \end{array}\right] \epsilon = \chi(t,0) \left[\begin{array}{cc} \frac{\partial S}{\partial v} R & S \frac{\partial R}{\partial \Omega} \end{array}\right] \left[\begin{array}{c} \delta v \\ \delta\Omega \end{array}\right] \tag{66}$$

which we will write as

$$\mathcal{B}(v,\Omega)\epsilon = \mathcal{A}(v,\Omega) \left[\begin{array}{c} \delta v \\ \delta\Omega \end{array}\right]. \tag{67}$$

The $N \times 4$ matrix $\mathcal{A}$ loses normal column rank only at the singular configuration $v = 1$, $\Omega = [0\ 0\ \theta]^T$ for all $\theta \in [0, \pi)$. However, this configuration does not belong to the state-space of the filter, for it has been eliminated by the normalization constraint. Therefore we can conclude

$$\left[\begin{array}{c} \delta v \\ \delta\Omega \end{array}\right] = \left(\mathcal{A}^T \mathcal{A}\right)^{-1} \mathcal{A}^T \mathcal{B}\epsilon \doteq \mathcal{C}(v,\Omega)\epsilon \tag{68}$$

and the induced norm of the matrix $\mathcal{C}(v,\Omega)$ is a measure of the "gain" between (small) disturbances in the constraints (or drifts outside the model class) and the errors in the estimates. In the experimental section we will show the result of a simulation where the disturbance level was increased up to the point in which the filter based upon the fixation constraint did not converge.

5 Attitude estimation from fixation

In some cases it may be desirable to reconstruct not only the relative velocity between the object being fixated and the viewer, but also their relative configuration, in the lines of [1]. Of course the relative configuration, assuming the initial time as the base frame, can be obtained by integrating velocity information, and this is indeed the only feasible solution when the motion of the viewer induces drastic changes in the image, such as occlusion, appearance of new objects etc. .

While in most applications the scene changes significantly and we cannot assume that the same features are visible over extended periods of time, in the case of fixation we can assume that the object stays in the field of view and we can integrate structure information from the same features to the extent in which they are visible.

Notice that, while in all the previous cases involving estimation of *velocity* (or relative configuration in the moving frame), we could decouple the motion parameters from the structure and therefore formulate filters involving only motion parameters and measured projections, in the case of the absolute orientation, it is necessary to include structure in the state of the filter.

The fixation assumption gives the strong constraint that the motion of the object being fixated rotates about the fixation point and translate along the fixation axis. This results in the fact that the object remains in the field of view as long as we fixate it. Therefore we will adopt an object-centered model, where the coordinates of each point are constant over time:

$${}^{o}\mathbf{P}^{i} = \text{const.} \tag{69}$$

Since we measure the projection of the coordinates of the point in the reference frame of the camera, we can enforce that the coordinates relative to the camera reference at the first instant are constant:

$$ {}^{t_o}\mathbf{P}^i \doteq {}^{o}\mathbf{P}^i - \begin{bmatrix} 0 \\ 0 \\ d_0 \end{bmatrix} = \text{const} \tag{70}$$

which relates to the measured projection via

$$ \mathbf{y}^i(t) = \pi\left({}^tR_{t_o}\,{}^{t_o}\mathbf{P}^i + \begin{bmatrix} 0 \\ 0 \\ d(t) \end{bmatrix}\right). \tag{71}$$

where ${}^tR_{t_o}$ is the relative orientation between the viewer reference at time t and the same reference frame at the initial time t_0.

We may conceive at this point a dynamic model having the trivial constant dynamics of the points in the state, and the above projection as the measurement constraint. In order to do so, we have to insert ${}^tR_{t_0}$ and $d(t)$, along with their derivatives, into the state of the filter, which becomes therefore $3N+8$-dimensional:

$$ \begin{cases} {}^{t_0}\mathbf{P}^i(t+1) = {}^{t_0}\mathbf{P}^i(t) & {}^{t_0}\mathbf{P}^i(0) = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ 1 \end{bmatrix} \quad \forall i = 1 \dots N \\ {}^tR_{t_0}(t+1) = {}^tR_{t_0}(t)e^{\Omega\wedge} & {}^tR_{t_0}(0) = I \\ \Omega(t+1) = \Omega(t) + n_\Omega & \Omega(0) = 0 \\ d(t+1) = d(t) + v(t) & d(0) = d_0 \\ v(t+1) = v(t) + n_v(t) & v(0) = 0 \\ \mathbf{y}^i(t) = \pi\left({}^tR_{t_0}\,{}^{t_0}\mathbf{P}^i + \begin{bmatrix} 0 \\ 0 \\ d(t) \end{bmatrix}\right). \end{cases} \tag{72}$$

where π denotes an ideal perspective projection. In the case of weak-perspective, the last measurement equation transforms into

$$ \mathbf{y}^i(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} {}^tR_{t_0}\frac{{}^{t_0}\mathbf{P}^i}{d}. \tag{73}$$

There is an additional constraint that can be imposed in order to set the overall scaling, which is

$$ {}^{t_0}\mathbf{P}^0(t) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \forall t. \tag{74}$$

The above can be imposed either as a measurement constraint, or as a model constraint by setting the variance of the corresponding state to zero, as in [1].

The above model may be reduced into a minimal one by removing the dynamics of the absolute orientation $d(t), R(t)$, and by exploiting the fact that

$$ {}^{t_0}\mathbf{P}^i = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}^i = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}^i (0)Z_0^i. \tag{75}$$

Since we measure the initial projection of each feature point, we can leave only the scaling (initial depth) Z_0^i in the state. It must be noticed, however, that the error in the location of the initial features is propagated through time, since we do not update the states corresponding to the measured projections. If one is willing to trade the drift due to the initial measurement error with eliminating $2N$ states from the model, he ends up with the following system

$$\begin{cases} Z_0^i(t+1) = Z_0^i(t) \qquad Z_0^i(0) = 1 \\ \Omega(t+1) = \Omega(t) + n_\Omega(t) \qquad \Omega(0) = 0 \\ v(t+1) = v(t) + n_v(t) \qquad v(0) = 0 \\ \mathbf{y}^i(t) = \pi\left(R(t) \begin{bmatrix} \mathbf{y}_1(0) \\ \mathbf{y}_2(0) \\ 1 \end{bmatrix} Z_0^i(t) + \begin{bmatrix} 0 \\ 0 \\ d(t) \end{bmatrix}\right). \\ Z_0^0(t) = 1 \end{cases} \tag{76}$$

where $R(t)$ and $d(t)$ are computed from the states $\Omega(t)$ and $v(t)$ at each time by integrating

$$\begin{cases} R(t+1)\ R(t) = e^{\Omega(t)\wedge} \qquad R(0) = I \\ d(t+1) = d(t) + v(t) \qquad d(0) = 1. \end{cases} \tag{77}$$

A simple EKF based upon the model above recovers the structure modulo the initial distance from the fixation point d_0. If such a distance is known, it is possible to recover the full structure, as well as the motion parameters $\Omega(t)$ and $v(T)$.

6 Experiments

6.1 Experimental conditions

In order to test the effectiveness of the schemes proposed, and compare it against equivalent motion estimation techniques that do not take into account the fixation constraint, we have generated a cloud of dots within a cubic volume at $d = 2m$ in front of the viewer. These dots are projected onto and ideal image plane with unit focal length and 500×500 pixels, corresponding to a visual angle of approximately 30^o. Noise has been added to the projections with 1 pixel std, corresponding to the average performance of current feature tracking techniques [2]. One random point in the cloud is chosen as the fixation point, and the cloud is then made rotate about this point and translate along the fixation axis with smooth but non-constant velocity.

6.2 Recursive filters

In figure 7 (top-left), the 4 components of the state of the filter described in section 3.2 are plotted, along with the ground truth in dotted lines. The plot on the right shows the absolute estimation error.

The same data have been fed to the essential filter [11], which estimates 5 states corresponding to the direction of translation and the rotational velocity without enforcing the fixation constraint. The states corresponding to the same motion described above, as long as ground truth, are plotted in the left-plot of figure 7 (bottom). The estimation error is marginally higher than the one of the filter with the fixation constraint.

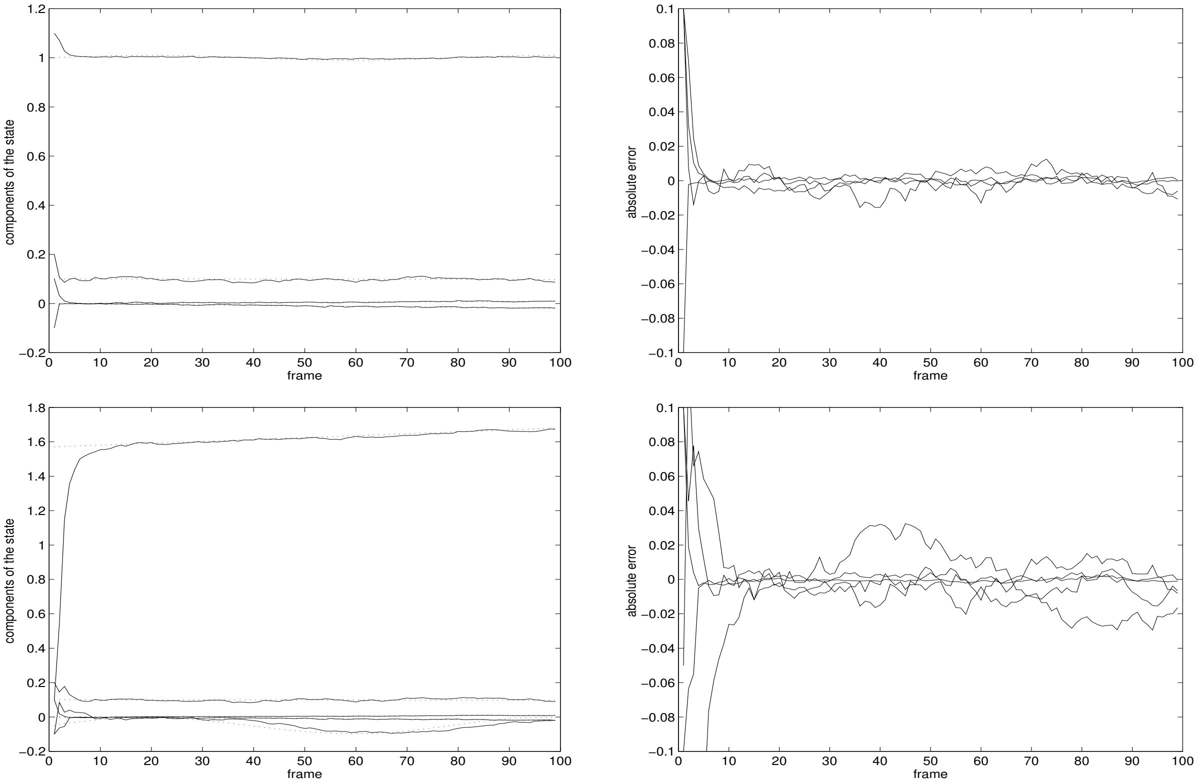

Figure 7: *(top-left) Estimates of the 4-dimensional state of the filter for estimating relative orientation under the fixation constraint. Filter estimates are in solid lines, while ground truth is in dotted lines. The estimation error (top-right) is smooth and strongly correlated, which is a symptom of poor tuning of the filter. If we do not enforce the fixation constraint, we need to estimate 5 motion parameters. The filter which does not enforce the fixation constraint converges faster (bottom-left) and the estimation error is larger but far less correlated (bottom-right), which indicates that the potential limits of the scheme have been achieved.*

In our preliminary set of experiments we have observed a higher robustness level in the filter enforcing the fixation constraint. For example, the maximum noise level tollerable by the filter not enforcing the fixation constraints in this particular experimental setup is 1.5 pixels, while the filter enforcing fixation performs up to 2.5 pixels, as reported in figure 8.

6.3 Attitude estimation

In figure 9 we report the estimates of the absolute orientation and structure as estimated by the filter described in section 5. The structure parameters (initial depth of all points) has been plotted against the true parameters, assuming that the initial distance of the fixation

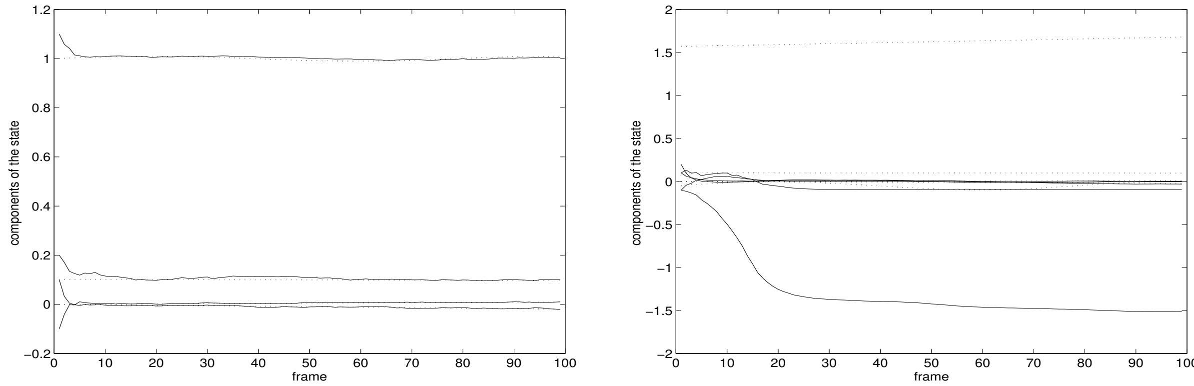

Figure 8: *(left) Convergence of the states of the filter enforcing the fixation constraint for a noise level in the feature tracking of 3 pixels. The filter that does not enforce the fixation constraint does not converge in the same experimental situation. Initial conditions, tuning of the filters and noise levels are the same for both filters.*

point is known. In general, structure can be recovered only up to a scale factor. The four motion components are also plotted, along with the estimation error, in the right plot.

It must be noticed that this filter has a $N+4$-dimensional state, unlike the one described above which has dimension 4. Furthermore, the filter has proven very sensitive to the initial conditions in the motion parameters, while the structure parameters can be safely initialized to 1, which corresponds to having the visible objects flat on the image plane. The error is significantly correlated and convergence is slow for the motion parameters, which are observable only through 2 levels of bracketing with the state equation.

In case occlusions occur in the image plane or some features disappear or exit the field of view, it is necessary to resort to the schemes described in section 3.2, unless we are willing to deal with a filter with a variable number of states.

6.4 Singularities and normalization

As we have mentioned in section 2.4, the non-normalized epipolar representation contains a singularity in $v = 1, \Omega = [0\ 0\ \theta]^T$, where the innovation of the filter becomes zero. Therefore, even when motion does not correspond to pure translation about the optical axis (the singular configuration), the filter may converge to the singular configuration whenever initialized far enough from the true state. In particular, when the noise level increases, the residual surface becomes more and more irregular, and it becomes easier for the filter to fall into the singular configuration.

In figure 10 (left) we show the state of the filter that is initialized far from the true inital conditions for a measurement noise level of 1 pixel. The filter converges to a state corresponding to $v = 1$ and $\Omega = [0\ 0\ \theta]^T$ with some θ. Correspondingly, the innovation goes to zero (fig. 10 right) and the filter saturates. The variance of the estimation error keeps increasing after the filter has saturated. In figure 10 (bottom) we plot the state with errorbars

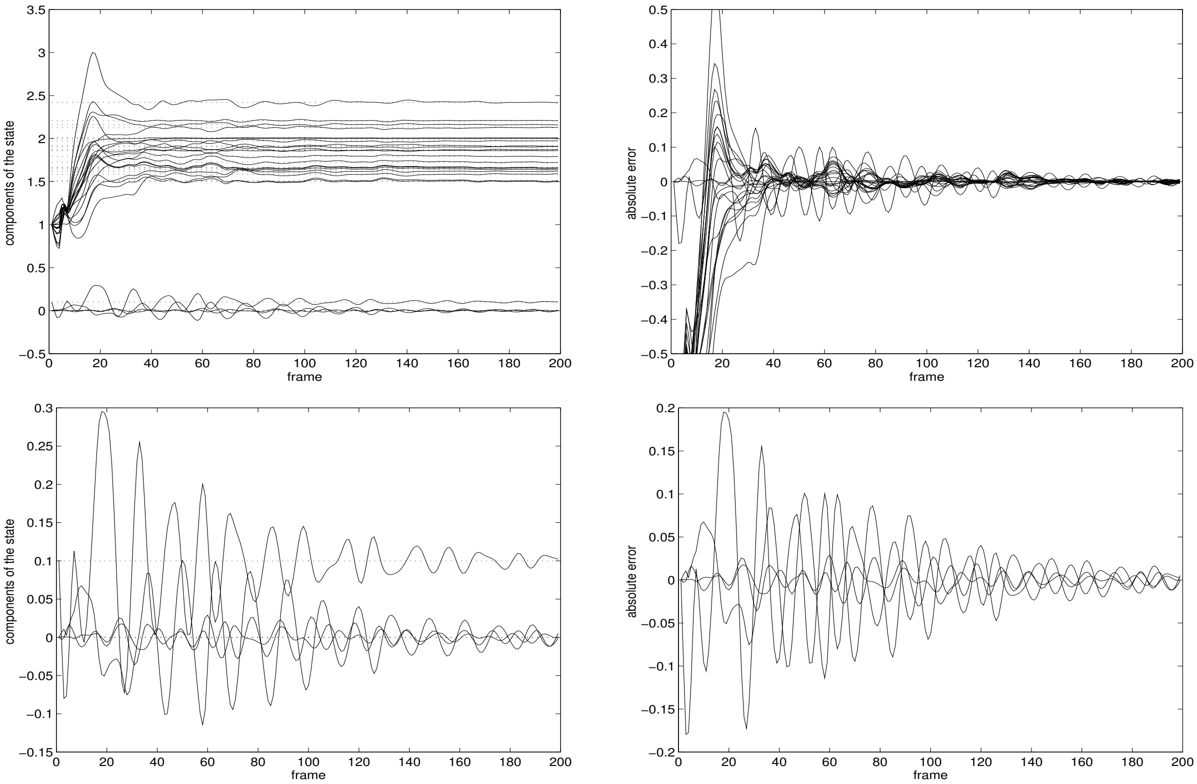

Figure 9: *(top-left) Estimates of the $N+4$-dimensional state of the filter for estimating absolute orientation and structure. Success in the estimation process depends crucially on the initial conditions of the motion parameters (bottom-left), while the structure-parameters can be safely initialized to 1, which corresponds to having the visible objects flat on the image-plane. The estimation error (top-right) is strongly correlated and decays slowly. The estimation error for the motion parameters, initialized within 1% off the true values, is plotted in (bottom-right) for comparison with the relative motion estimation scheme.*

corresponding to the diagonal elements of the variance/covariance matrix of the estimation error. It can be seen that, after the variance decreases due to the initial convergence towards the minimum, it keeps increasing steadily once the filter has saturated.

When the same initial conditions and noise levels are applied to the filter based upon the normalized essential matrices, convergence is achieved without any problems of saturation (figure 11).

6.5 Sensitivity to the fixation constraint

In order to experiment with the degradation of the filter enforcing the fixation constraint in presence of motions that violate the fixation assumptions, we have perturbed the experiments described above by translating the cloud on a plane orthogonal to the fixation axis at random within a standard deviation ranging from 1% to 6% of the norm of the essential matrix. We have started from the true initial conditions and added no noise to the measurements. For each level of disturbance, we have run 100 experiments, and computed the estimation error for the translation along the fixation axis and for the rotation components. The results are plotted in figure 12, where we show the average error across different trials, with the standard deviation showed as an errorbar. The results seem to confirm that the degradation of the estimates is graceful for small disturbances. However, when the disturbance exceeds 6% of the overall norm of the current relative motion, the filter does not reach convergence.

7 Conclusions

We have studied the problem of estimating the motion of a rigid object viewed from a monocular perspective camera which is actuated as to track one particular feature-point in the scene. We have cast the problem in the framework of epipolar geometry, and formulated both closed-form and recursive schemes for recursively estimating motion and attitude using the fixation constraint. The framework of dynamic epipolar geometry allows us to compare the proposed scheme directly against the equivalent scheme that does not enforce the epipolar constraint. Also, the degradation of the performance in the presence of disturbance in the fixation hypothesis is assessed.

The performance of the estimators have been compared via simulations to the equivalent estimation schemes that does not enforce the fixation constraint. The results seem to indicate that using the fixation constraint helps achieving better accuracy, in the presence of perfect tracking. Degradation of the performance in the presence of disturbance in the fixation constraint is graceful for small disturbances. It will be subject to future research to study how to compensate for non-perfect tracking by feeding back a measure of "goodness of fixation" and performing a shift-registration of the origin of the image plane.

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

[3] M. J. Barth and S. Tsuji. Egomotion determination through an intellingent gaze control strategy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1993.

[4] F. Bullo, R. M. Murray, and A. Sarti. Control on the sphere and reduced attitude stabilization. In *Proceedings of the IFAC Symposium on Nonlinear Control Systems NOLCOS*, Tahoe City, June 1995.

[5] C. Fermüller and Y. Aloimonos. Tracking facilitates 3-d motion estimation. *Biological Cybernetics (67), 259-268*, 1992.

[6] C. Fermüller and Y. Aloimonos. The role of fixation in visual motion analysis. *Int. Journal of Computer Vision (11:2), 165-186*, 1993.

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

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

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

[10] D. Raviv and M. Herman. A unified approach to camera fixation and vision-based road following. *IEEE Trans. on Systems, Man and Cybernetics vol. 24, n. 8*, 1994.

[11] 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.

[12] M. A. Taalebinezhaad. Direct recovery of motion and shape in the general case by fixation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1992.

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

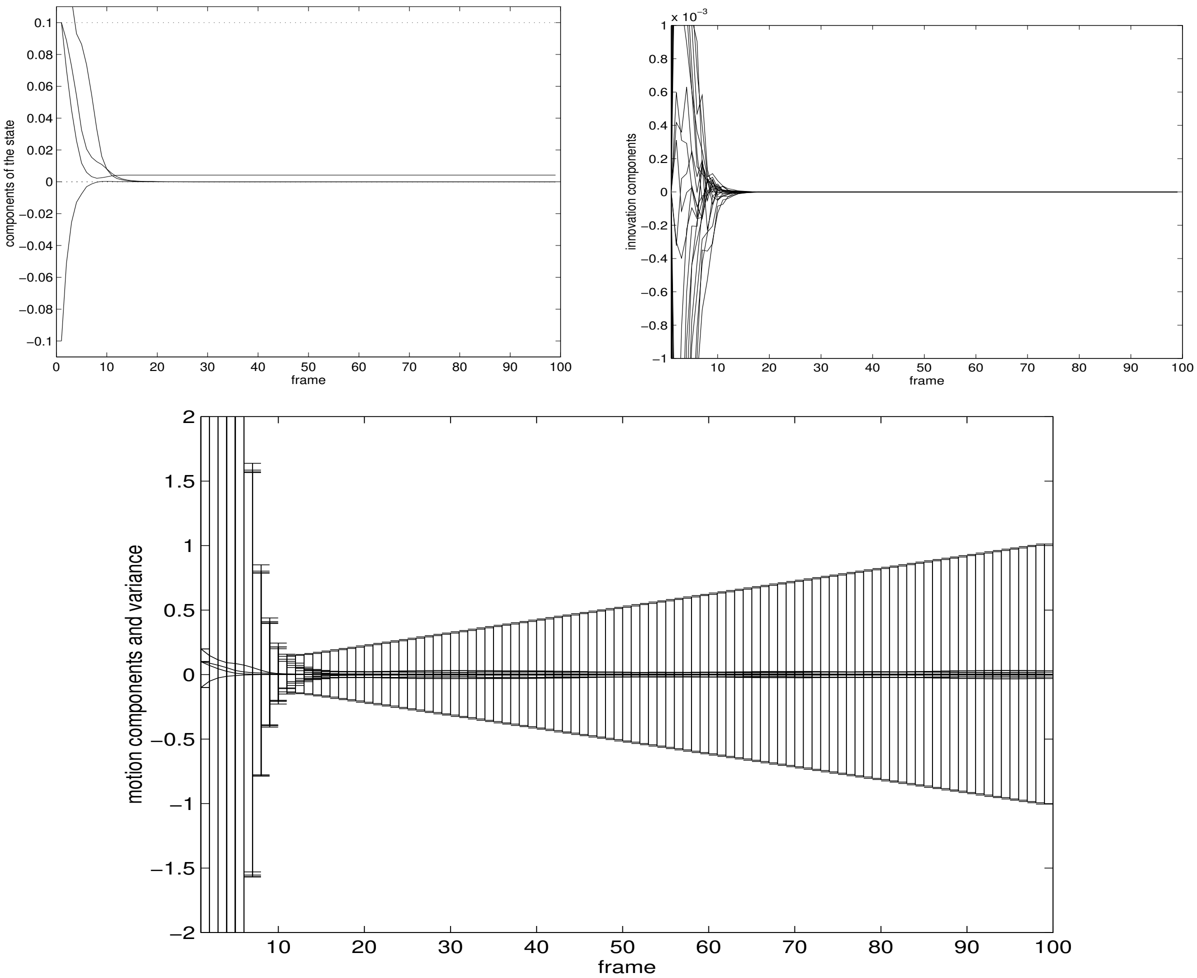

Figure 10: *(top-left) Convergence of the filter to the singular configuration. For a noise level of 1 pixel and the initial conditions far enough from the true values, the state of the filter ends up in the minimum of the residual surface corresponding to cyclorotation (all states are zero but* Ω_3 *which is arbitrary). Correspondingly the innovation becomes zero (top-right) and the variance increases (bottom plot). The variance is represented via the errorbars in the motion estimates, which are the diagonal elements of the variance/covariance matrix of the estimation error.*

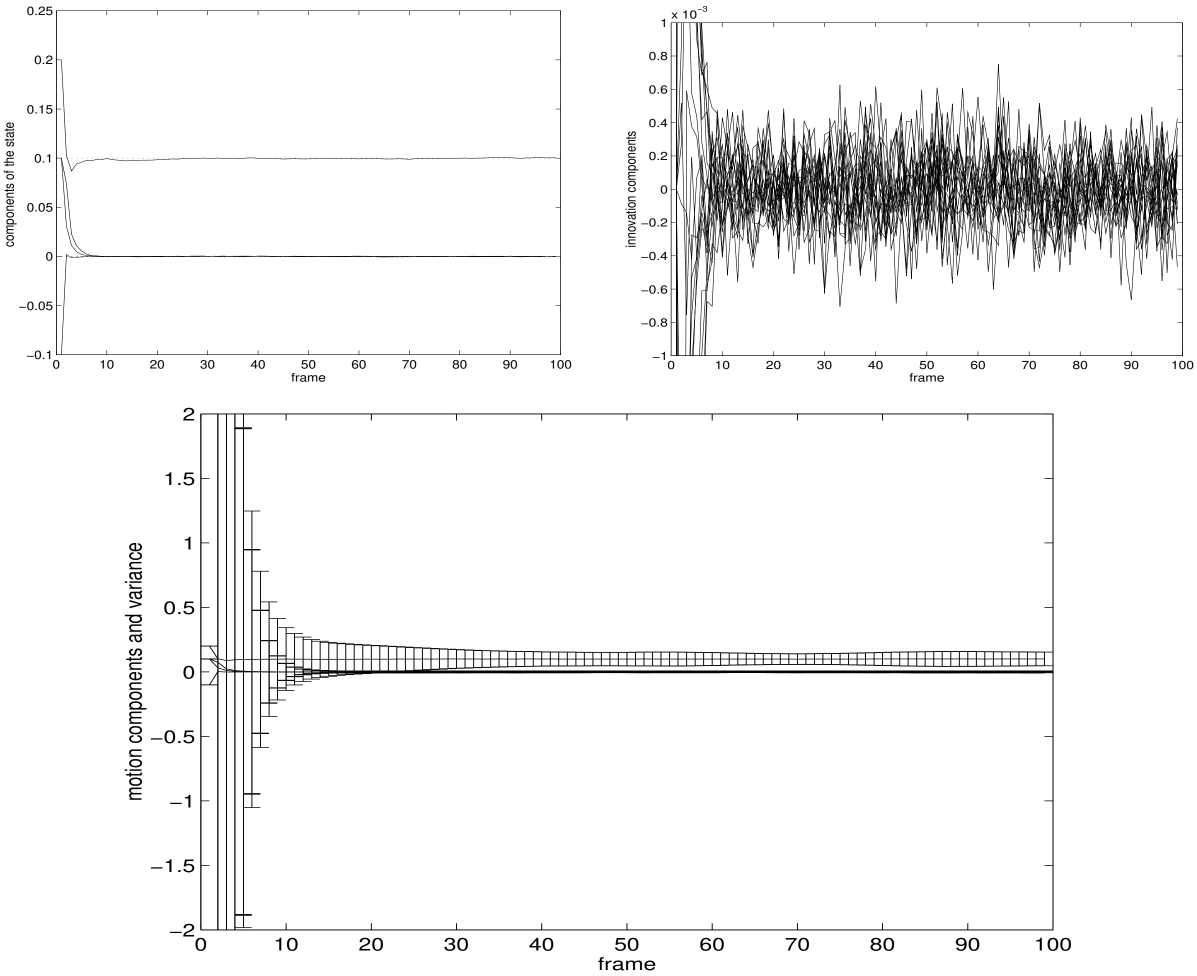

Figure 11: *(top-left) Convergence of the filter enforcing the normalization constraint. There are no singular configurations in the state manifold, and the filter converges fast to the correct estimate. The innovation is small but non-zero (top-right), and the variance of the state decreases as time grows (bottom).*

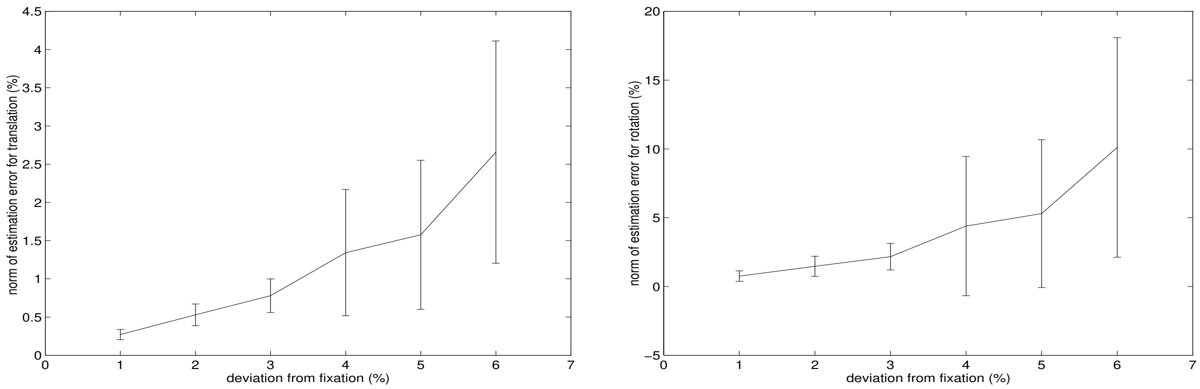

Figure 12: *Estimation error versus disturbances in the fixation constraint. The plots show the average over 100 trials, with the standard deviation across trials shown as an errorbar. When the fixation constraint is violated by adding spurious translation components ranging from 1 to 6 percent of the norm of the fixating motion, the estimation error increases gracefully. In the left plot the estimation error for the translation along the optical axis, on the right the norm of the estimation error for the rotational velocity.*