Motion from "X" by Compensating "Y"*

A unified framework for motion estimation from image sequences by compensating for the image-motion of a point, a line or a plane

Stefano Soatto and *Pietro Perona*

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

March 7, 1995

keywords: Motion and structure estimation, essential manifold, motion decoupling, motion compensation, fixation, plane plus parallax, plane fitting, direct methods, real-time motion estimation.

Abstract

This paper analyzes the geometry of the visual motion estimation problem in relation to transformations of the input (images) that stabilize particular output functions such as the motion of a point, a line and a plane in the image. By casting the problem within the popular "epipolar geometry", we provide a common framework for including constraints such as point, line of plane fixation by just considering "slices" of the parameter manifold. The models we provide can be used for estimating motion from a batch using the preferred optimization techniques, or for defining dynamic filters that estimate motion from a causal sequence. We discuss methods for performing the necessary compensation by either controlling the support of the camera or by pre-processing the images. The compensation algorithms may be used also for recursively fitting a plane in 3-D both from point-features or directly from brightness. Conversely, they may be used for estimating motion relative to the plane independent of its parameters.

*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 95-009, March 1995.

1 Introduction

Suppose you are looking at a scene through a moving camera. The problem of visual motion and structure estimation deals with reconstructing both the relative motion between the scene and the camera, and the "structure" of the scene. The strategies for solving the problem depend on how we represent the "structure" of the scene and its motion relative to the viewer.

Suppose that our scene is described by a number N of *point-features* in 3-D space, with coordinates $\mathbf{X}^i \; \forall i = 1 \dots N$ relative to some reference frame centered in the optical center of the camera, which move *rigidly* between one time-instant and another, with some relative translation T and relative orientation R. Suppose we are able to measure the *perspective projection* of each point-feature onto the 2-D image plane, through the projective coordinates $\mathbf{x}^i$. We also assume we are able to assess which feature corresponds to which across different views (the correspondence problem; see [1] for a number of techniques for addressing this problem).

1.1 Motion and structure estimation as an optimization problem

Once the geometric constraints involved in the problem (namely the rigidity constraint and the point-wise representation of structure) and the measurement model (perspective projection) have been formalized, one can set up an optimization problem in order to estimate the $3N + 6M$ unknown parameters (3 space coordinates for each feature-point and 6 components of motion across M time instants), from the $2NM$ image projections of the N points at each of the M images.

There are two aspects which are tightly related in formulating the optimization task: the *model* being used, and the *estimation* techniques employed. A variety of models have been proposed for estimating structure and motion from images, which were then employed in batch optimization techniques (closed-form from two or more views or iterative) or in recursive estimation methods.

A simple counting of the dimensions involved will soon convince the reader that, regardless the estimation method employed, the huge dimensionality of the problem and the highly nonlinear nature of the parameter space make the optimization so complicate that the issue of an appropriate *modeling* becomes crucial.

A typical number of feature-points visible on each frame of a realistic scene is, say, $N = 100$. If we consider a sequence of $M = 30$ images, corresponding to one second of video, we have 480 unknown parameters, with 6000 available measurements. The unknowns live on a parameter space that is diffeomorphic to

$$\mathbb{R}^{3^N} \times SE(3)^M \tag{1}$$

where $SE(3)$ is the Lie-group of Euclidean motions in $\mathbb{R}^3$ [9]. We are going to be able to recover only 479 parameters, since there is an overall scaling ambiguity that affects the depth of each point and the norm of the direction of translation [8]. Even if we consider the camera as moving with *constant velocity* during the 1 second video sequence, we still have 305 parameters to estimate.

1.2 Decoupling as a modeling strategy

When facing a high-dimensional optimization problem it is important to understand the geometry of the parameter space in order to see whether there are "slices" of it where the parameters evolve independently in the cost objective. Suppose for instance that our optimization task can be written in the form

$$\hat{x}, \hat{y} = \arg \min_{x \in X, y \in Y} f(x, y) \tag{2}$$

and suppose that we can identify a subspace of the space X, of the form

$$\{x = g(y) \mid y \in Y\} \subset X \tag{3}$$

such that, when $\hat{y}$ solves the above optimization problem, the corresponding $\hat{x}$ is given by $\hat{x} = g(\hat{y})$. Then we can decompose the original optimization problem (locally) into a smaller-dimensional one of the form

$$\hat{y} = \arg \min_{y \in Y} f(g(y), y) \tag{4}$$

whose solution can be used for computing

$$\hat{x} = g(\hat{y}). \tag{5}$$

This procedure responds to the need of decomposing a high-dimensional optimization task into the solution of a number of smaller, simpler and better constrained problems by exploiting the geometric structure of the parameter space.

In the case of structure and motion estimation, the work of Longuet-Higgins [8] follows this direction, by decoupling the structure parameters $\mathbf{X}^i$ from the motion parameters T, R, which are encoded as elements of an $8-$dimensional space, called the *essential manifold* [13]. Heeger and Jepson [5] further decouple the translational velocity from the rotational velocity in the continuous-time approximation. Therefore, the algorithms of Longuet-Higgins and Heeger and Jepson, applied to the original task of estimating structure and motion, formulate a constraint involving only $8M$ and $2M$ unknown parameters respectively, from which all the other unknowns can be recovered a-posteriori.

The models described by Longuet-Higgins and Heeger-Jepson are essentially *static*, in the sense that the estimates of motion at the frame m depend only upon measurements of the neighboring frames m and $m-1$. The coherency of the structure and motion across multiple frames may be exploited; in [13], the constraints formulated by Longuet-Higgins and Heeger and Jepson are viewer as implicit dynamical systems of some particular class (Exterior Differential Systems), and a recursive estimation scheme is proposed for integrating information over time in a *causal* fashion (the estimates at the frame m depend upon measurements from the images $1 \ldots m$).

1.3 Compensation of image-motion

Motivated by the mechanics of the oculomotory system in most mammals, a number of studies have suggested that the task of estimating motion is made easier if some particular point on the image-plane is being "fixated" [4, 11, 15].

The claim is that fixation, intended as a "pre-processing" stage, facilitates motion analysis by reducing the number of residual degrees of freedom. The pre-processing can be accomplished both "mechanically" by rotating the eye, or "algorithmically" by shifting the coordinate system of the image-plane.

In a completely different context, alternative representation of the scene have been proposed, which refer the structure to some plane in the scene. After "warping" the image so as to stabilize the image of the plane, the residual image-motion is simpler to analyze and is related only to a small number of free parameters, while the others have been "factored out" by the warping procedure [12, 10].

Both operations, fixation and warping, can be viewer as a pre-processing stage in which we try to compensate for the image motion of a point or a plane. We can imagine another situation in between these two extrema, which consists in compensating for the motion of a point and the orientation of a line in the image plane.

Alternatively we could view these pre-processing operations as a closed control loop that stabilizes the image motion of a point, a point and a line, or a plane.

1.4 Compensation for decoupling: geometric stratification

In this paper we show that the concepts of image compensation (or stabilization) and decoupling of motion and structure parameters are closely related.

We start off by recalling the setup of epipolar geometry [8] in order to decouple structure from motion, without any compensation. Motion estimation is qualified as an optimization task with the parameters on the essential manifold, which can be solved in closed-form from two views [8, 17, 3], iteratively from two views [7] or recursively from an image sequence [13].

Then we explore how the setup of epipolar geometry is modified under the assumption that the motion of a point, a line or a plane has been compensated. We will see that such compensations allow us to identify "slices" of the essential manifold and therefore define smaller, simpler and better-constrained models for estimating motion.

In the general case, the parameters evolve on the $5-$dimensional essential manifold; once we compensate for the motion of a point, a line or a plane, we reduce the problem to a $4, 3$ and $2-$dimensional submanifold respectively. The table below summarizes this geometric stratification. Note that, while fixation of a point, or a point and a line, can be achieved both mechanically and algorithmically, there is no physical 3-D relative motion between the camera and the scene that stabilizes the image-motion of a plane. Therefore, this may only be accomplished in software.

Geometric stratification of the problem of estimating motion under the compensation of the image-motion of a point, a point and a line, and a plane.

Stabilized feature	**Compensating 3-D motion**	**Corresponding image deformation**	**Residual DOFs**	**State-space manifold**
none	none	none	5	**E** Essential mfd
point	2-D camera rotation	image center displacement	4	$\mathcal{S}^4$ Sylvester mfd
point+line	rotation about optical center	image center shift + rotation	3	$\mathcal{S}^3$ 3-dimensional Sylvester mfd
plane	no feasible 3-D rigid motion	quadratic warping	2	$so(3)$ skew-symmetric unit-norm 3-matrices

1.5 Relation to previous work

This paper analyzes the geometry of the motion estimation problem in relation to transformations of the input images that stabilize particular output functions such as the motion of a point, a line and a plane in the image. As a side-effect, it outlines a unified modeling framework for estimating rigid 3-D motion under compensation of image-motion. The geometric framework is the popular "epipolar geometry", which has been object of extensive study over the past decade (see [3] for a review). Diverse studies on motion fixation [4, 11, 15] and structure representation [12, 10] are cast in the same framework, which allows us to compare the estimates of motion under the different fixation assumptions. Another side-effect is the derivation of a discrete-time equivalent of the model proposed by Heeger and Jepson [5] under the instantaneous approximation.

Most of the paper is concerned with *modeling*. However, for each model proposed, we suggest a formulation of a dynamic filter that recursively estimates the parameters of the model. These filters are based upon the general techniques presented in [13].

The paper also describes how to actually design the image compensations which the models are based upon. These can be derived both from point-features, or directly from brightness, and therefore fall in the category of the so-called "direct methods" [6]. The models for image warping from brightness can be easily extended for estimating the motion of a plane or the direction of translation from point-features or directly from the image brightness.

1.6 Organization of the paper

Section 2 serves to establish the notation and introduce the well-known setup of epipolar geometry. The coplanarity constraint introduced by Longuet-Higgins [8] is derived, and possible estimation techniques that exploit it are described, which include closed-form and iterative solutions from two views, or recursive multi-frame estimation. The parameters of any estimation scheme based upon the epipolar constraint evolve in the so-called "essential manifold", which is a differentiable (smooth) manifold whose structure is briefly described in section 2.2.

Section 3 studies how the setup of epipolar geometry is modified when one point is being fixated on the image plane. We show that the fixation constraint defines a simple submanifold of the essential manifold, and therefore all the techniques used for estimating a general motion can be particularized to this case by just restricting the parameter to the corresponding "slice" of the essential manifold. As far as actually stabilizing the motion of a point on the image-plane, we refer the reader to the appropriate literature.

In section 4 we further constrain the motion by assuming that the position of a point and the orientation of a line are fixed in the image plane.

In section 5 we study the case when the image has been warped such as the motion of a plane in the scene has been compensated. We describe the so-called "plane-plus-parallax" representation [12, 10], and unreveal the geometric structure that induces on the essential manifold. In section 5.4 we discuss methods for actually performing the warping, both from point-features and directly from image brightness.

As a side-effect, we introduce a model for recursively fitting a plane in the scene both from feature-point correspondence and from brightness (section 5.5), as well as a model for estimating motion relative to the plane.

2 Epipolar geometry

We call $\mathbf{X} = \left[\begin{array}{ccc} X & Y & Z \end{array} \right]^T \in \mathbb{R}^3$ 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. Since we are interested in the displacement relative to the moving frame (ego-motion), we can write the rigid motion of the point of coordinates $\mathbf{X}^i$ between time t and $t+1$ as

$$\mathbf{X}^i(t+1) = R(t)\mathbf{X}^i(t) + T(t) \tag{6}$$

The matrix $R \in SO(3)$ is an orthonormal rotation matrix that describes the change of orientation 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. The 3×3 rotation matrix R comprises 3 degrees of freedom, which we represent as the three-dimensional vector of exponential coordinates Ω, defined such that $R = e^{\Omega\wedge}$ [9].

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 $\mathbb{RP}^2$:

$$\begin{array}{rcl} \pi : \mathbb{R}^3 - \{0\} & \rightarrow & \mathbb{RP}^2 \\ \mathbf{X} & \mapsto & \mathbf{x} \doteq \left[\begin{array}{ccc} \frac{X}{Z} & \frac{Y}{Z} & 1 \end{array} \right]^T. \end{array} \tag{7}$$

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:

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

2.1 Coplanarity constraint

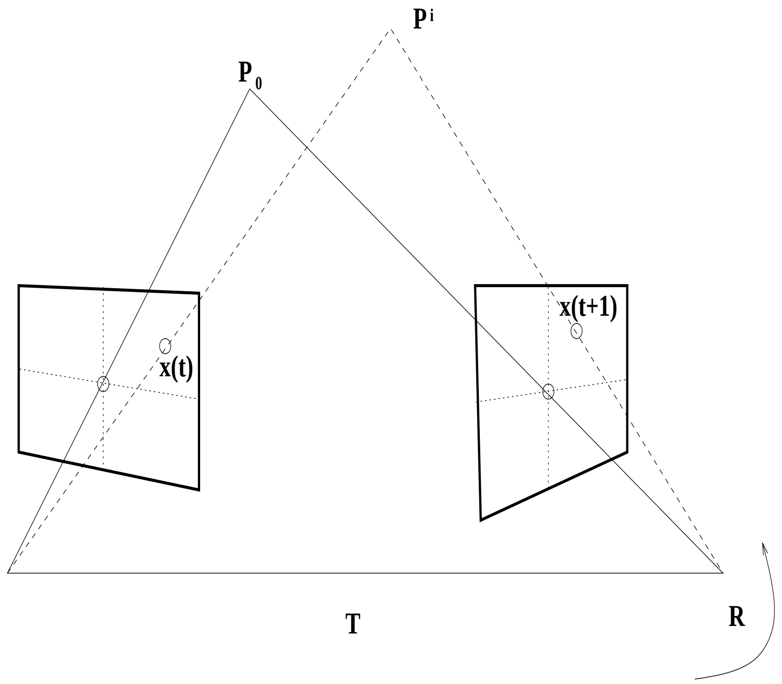

Figure 1: *Coplanarity constraint: the coordinates of each point in the reference of the viewer at time t, the coordinates of the same point at time t+1 and the translation vector are coplanar.*

The well-known coplanarity constraint (or "epipolar constraint", or "essential constraint") of Longuet-Higgins [8] imposes that the vectors $T(t)$, $\mathbf{X}^i(t+1)$ and $\mathbf{X}^i(t)$ be coplanar for all t and for all points $\mathbf{P}^i$ (figure 1). The triple product of the above vectors is therefore zero. In order to write the triple product in a common coordinate system, we multiply both sides of (6) by $\alpha\mathbf{X}^i(t+1)^T(T\wedge)$, where $\alpha \in \mathbb{R} - \{0\}$, ending up with

$$0 = \mathbf{X}^i(t+1)(T\wedge)R(t)\mathbf{X}^i(t) \tag{8}$$

which we will write as

$$\mathbf{X}^i(t+1)\mathbf{Q}(t)\mathbf{X}^i(t) = 0 \tag{9}$$

with

$$\mathbf{Q}(t) \doteq \mathbf{Q}(R(t), T(t)) = (T(t)) \wedge R(t). \tag{10}$$

We will use the notation $\mathbf{Q}(t)$ when emphasizing the time-dependence, while we will use $\mathbf{Q}(R,T)$ when stressing the dependence of $\mathbf{Q}$ from the 3 rotation parameters contained in R and from the normalized translation T.

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 (9) holds for $\mathbf{x}^i$ in place of $\mathbf{X}^i$ (just divide eq. (9) by $\mathbf{X}^i_3(t+1)\mathbf{X}^i_3(t)$):

$$\mathbf{x}^i(t+1)\mathbf{Q}(t)\mathbf{x}^i(t) = 0 \qquad \forall t\ ,\ \forall i. \tag{11}$$

2.2 The essential manifold

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

$$\mathbf{E} \doteq \{SR \mid S \in so(3),\ R \in SO(3)\}, \tag{12}$$

whose structure of an algebraic variety has been object of massive study over the past decade (see [3] for a review). Only very recently, however, it has been realized that the essential manifold is indeed a differentiable (smooth) manifold, since it can be characterized as the tangent bundle to the rotation group $TSO(3)$ [13], which is a six-dimensional smooth manifold. It is possible to characterize the topological properties of the essential manifold by defining a local coordinate chart, in the lines of [13].

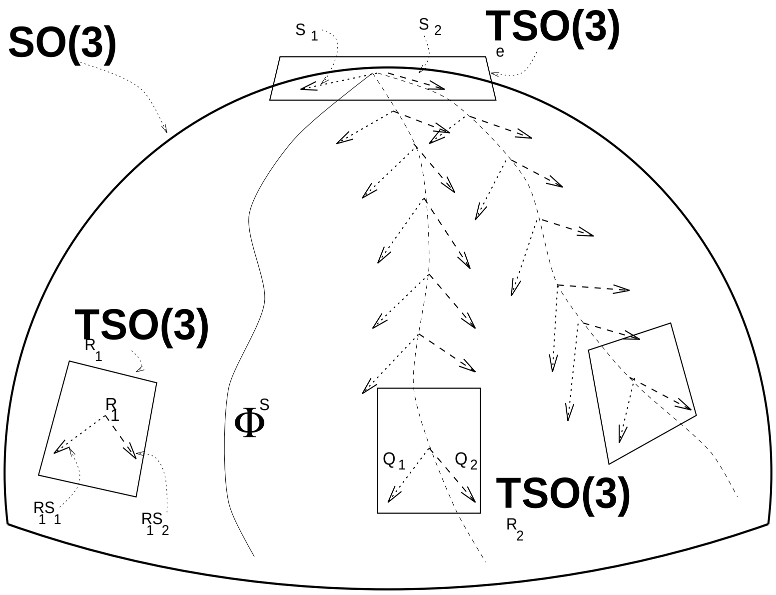

Figure 2: *The essential manifold as the tangent bundle of the rotation group*

2.3 Motion estimation from the epipolar constraint

The coplanarity constraint has been used for over a decade in order to estimate rigid motion from images. The schemes available can be roughly classified as two-frames, closed-form solutions, two-frames iterative solutions or recursive, multi-frame algorithms.

Closed-form solutions consist of first estimating the parameters of a *generic* matrix $\mathbf{Q}$ from a number $N \geq 8$ of epipolar constraints (11), and then *unfolding* the parameters T and R from the estimated $\mathbf{Q}$, in the lines of [8, 16, 3] and many other modifications of the basic scheme of Longuet-Higgins [8].

These schemes are quasi-linear, in the sense that both estimating $\mathbf{Q}$ from the epipolar constraints and unfolding the motion parameters from it can be accomplished using essentially linear techniques. However, the procedure is not optimal, because the structure of the matrix $\mathbf{Q}$ is not enforced in the estimation stage, but rather "a posteriori", so that the estimate of $\mathbf{Q}$ is not guaranteed to belong to the essential matrix. In order to overcome this problem, one could substitute the parameters T and Ω, where Ω are the exponential coordinates of R, into the epipolar constraint, and then solve iteratively for this parameter for a number of constraints, in the lines of [7]. This procedure is more robust than the closed-form, but unpredictable due to the sensitivity of the iterative descent procedure in the presence of foldings of the error surface or local minima.

Another possibility consists in viewing the epipolar constraint (11) as an *implicit dynamical system with parameters on the essential manifold.* The so-called "Essential filter" described in [13] provides a principled way of identifying the motion parameters recursively from the dynamical model

$$\begin{cases} \mathbf{x}^i(t+1)\mathbf{Q}(t)\mathbf{x}^i(t) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} \in \mathbf{E}. \tag{13}$$

3 Compensating for a point: motion from fixation

Suppose now that some device provides us with a sequence of images where the projection of a given point on the image-plane remains fixed. This is the case of a viewer moving while fixating some object in the scene. In section 3.1 we show how the setup of epipolar geometry is modified under the fixation assumption. In the following section 3.3 we describe how it is possible to design both an "hardware" device of a simple "software" device that controls fixation of a point.

3.1 Motion from fixation

Since the projection of the fixation point is still in the image plane, the object (scene) 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. These four degrees of freedom are encoded into the rotation matrix $R = e^{\Omega\wedge}$, and in the relative translation along the fixation axis $v \in \mathbb{R}$. It is easy to see that the representation presented in the previous section generalizes easily once we represent the translation T as

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

and

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

is the ratio between the distance of the fixation point at time $t+1$ and the same distance at time t.

3.2 Modification induced on the essential manifold

The coplanarity constraint (11) also holds in the case of fixation, once we have substituted the appropriate expression for T. Since there are now fewer degrees of freedom (4, out of 5 that were present in the general case), the parameters Ω and v will now lie on a four-dimensional subspace of the essential manifold. Indeed, it can be shown [14] that the essential matrices under the fixation constraint are all and only the 3×3 essential matrices that satisfy the following Sylvester's equation

$$\mathbf{Q}(R, v) = RS^T + vSR \tag{16}$$

where

$$S \doteq \begin{bmatrix} 0 & -\alpha & 0 \\ \alpha & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{17}$$

and α is the arbitrary scaling factor due to the homogeneous nature of the coplanarity constraint. We will call $\mathcal{S}^4$ the four-dimensional submanifold of the essential manifold which is defined by the above equation. The $\mathcal{S}^4$ manifold is locally diffeomorphic to $\mathbb{R} \times SO(3)$ and hence to $\mathbb{R}^4$.

Therefore, in order to estimate motion under the fixation constraint, it is sufficient to consider the epipolar constraint where now the parameters are constrained not on the essential manifold, but on the $\mathcal{S}^4$-manifold.

$$\begin{cases} \mathbf{x}^i(t+1)\mathbf{Q}(t)\mathbf{x}^i(t) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} \in \mathcal{S}^4 \tag{18}$$

where

$$\mathcal{S}^4 = \{\mathbf{Q} \in \mathbf{E} \mid \mathbf{Q} = RS^T + vSR, R \in SO(3), v \in \mathbb{R}, S = [0\ 0\ 1]^T \wedge\}. \tag{19}$$

In [14] we have presented both recursive multi-frame and batch motion estimation techniques based upon the fixation constraint.

3.3 Fixation control

Keeping a single feature point fixed on the image plane can be accomplished both by rotating the camera about the center of projection (or about any other point in space), or by shifting the center of the image-coordinates by a purely software operation. As far as the effects are concerned for motion estimation, the two methods are equivalent. A gaze-control technique based upon geodesic control on a sphere is described in [14] and based upon [2], while image-shift registration techniques are described, for instance, in [15].

4 Compensating for a point + a line: motion from planar fixation

Suppose now that some external device is capable of not only keeping the fixation point still on the image plane, but also of maintaining one additional feature on a line passing through the fixation feature. In this section we explore how this constraint affects the epipolar framework (section 4.1) and how it is possible to achieve such a fixation (section 4.3).

4.1 Motion from planar fixation

Suppose that we maintain a point and a line passing through it fixed in the image plane. We are essentially in the same situation described in the previous section once we have "frozen" the degree of freedom corresponding to cyclorotation (rotation about the optical axis). Therefore there are overall 3 degrees of freedom.

4.2 Modification induced on the essential manifold

The essential matrices corresponding to motions that obey the point plus line fixation constraint must lie on a three-dimensional submanifold of the submanifold $\mathcal{S}^4$ of the essential manifold $\mathbf{E}$, since the point-fixation constraint described in the previous section is satisfied. The only modification that occurs is that now there is no translation about the $Z-$ axis (cyclorotation). Therefore the parameter space becomes

$$\mathcal{S}^3 = \{\mathbf{Q} \in \mathbf{E} \mid \mathbf{Q} = RS^T + vSR, R = e^{\left[\begin{array}{ccc} \omega_1 & \omega_2 & 0 \end{array}\right]^T \wedge}, v, \omega_1, \omega_2 \in \mathbb{R}, S = [0\ 0\ 1]^T \wedge\} \quad (20)$$

Therefore, under the point plus line fixation assumption, we can still use the standard estimation techniques based upon the epipolar constraint (closed-form, iterative or recursive) provided that we restrict the parameter manifold to the 3-dimensional submanifold of the essential manifold described by the above equations

$$\begin{cases} \mathbf{x}^i(t+1)\mathbf{Q}(t)\mathbf{x}^i(t) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} \in \mathcal{S}^3. \quad (21)$$

4.3 Line fixation control

Fixating a line on the image plane can be easily achieved by fixating a point and then rotating the image until the other point comes to the desired line. This can be accomplished both by rotating the camera about the fixation axis, or by rotating the image about the optical center with a purely software operation.

5 Compensating for a plane: plane plus parallax

We now proceed in our stratification by assuming that we are able to "compensate" the image sequence in such a way that the points that lie on an "average plane" of the scene (or on any other arbitrary plane) remain fixed in the image plane. In this case there is no physical motion of the camera that achieves this compensation (besides the trivial still configuration). Therefore we need to "deform" the images of the sequence in order to account for the motion of the plane. In section 5.4 we will show how it is possible to achieve such a compensation purely from image brightness or from point-correspondences, without direct knowledge of the motion of the camera or of the parametrization of the plane. In the next section 5.1, instead, we will see how the epipolar geometry is modified by this constraint. We will show, as it has already been noticed [12, 10], that, after the motion of the plane has been compensated, the residual motion depends only upon translation, while rotation has been "factored out". Therefore, only the two parameters of the direction of translation are left in the epipolar constraint. The subspace of the essential matrix that corresponds to the plane-fixation has an appealing geometric description and the factorization of the rotational component of motion from the translational part is complete.

5.1 Plane-plus-parallax representation

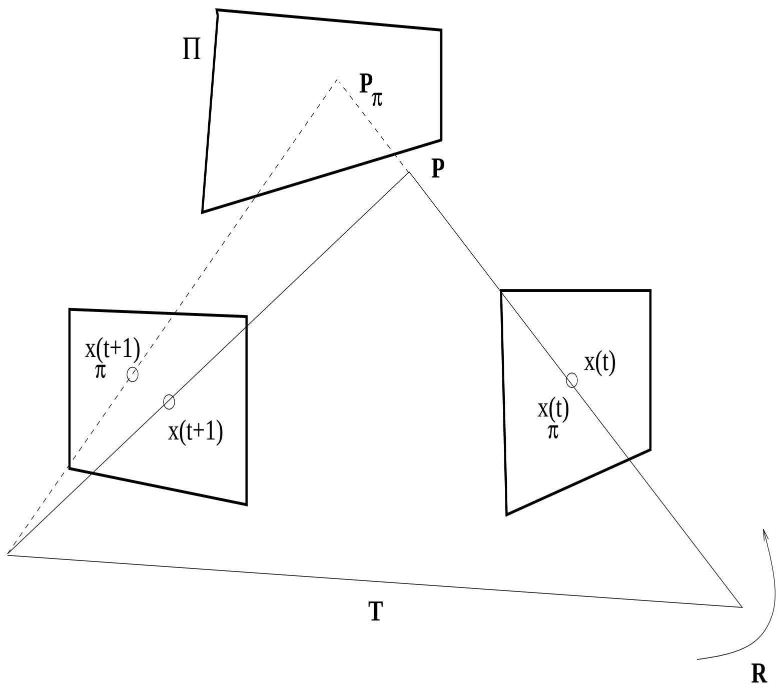

Figure 3: *Plane plus parallax representation*

Suppose that we are given a plane in the image which does not pass through the center of projection, described by

$$\Pi = \{X \in \mathbb{R}^3 \mid \mathbf{a}^T X = 1\} \tag{22}$$

where $\mathbf{a} = [a_1 \; a_1 \; a_3]^T$ are the parameters describing the planar surface. This plane could be the least-square fit of the scene, or it could be any planar surface not intersecting the center of projection. Suppose at time t we observe some point $\mathbf{P} \notin \Pi$, through its coordinates $\mathbf{x}(t)$. Now call $\mathbf{P}_\Pi$ the point obtained by intersecting the plane with the vector $\mathbf{x}(t)$ (see figure 3). Its projection clearly coincides with the one of $\mathbf{P}$:

$$\mathbf{x}_\Pi(t) = \mathbf{x}(t). \tag{23}$$

Now suppose that the camera moves between time t and $t+1$, and that the coordinates of each point $\mathbf{x}^i(t+1)$ is warped in such a way that the coordinates of the points lying on the plane Π remain unchanged (we will see later on how to accomplish such a warping):

$$\mathbf{y}_\Pi^w(t+1) = \mathbf{y}_\Pi(t) \qquad \forall \; \mathbf{y}_\Pi \in \Pi. \tag{24}$$

Therefore

$$\mathbf{x}_\Pi^w(t+1) = \mathbf{x}_\Pi(t) = \mathbf{x}(t) \tag{25}$$

in the coordinate frame of the viewer at time $t+1$. The epipolar constraint imposes that $\mathbf{x}_\Pi^w(t+1), \mathbf{x}^w(t+1)$ and $T(t)$ be coplanar (see figure 3). Note that these three vectors are all defined in the same reference frame, the one of the viewer at time $t+1$. By writing the triple product as

$$\mathbf{x}^w(t+1)^T(T(t) \wedge \mathbf{x}_\Pi^w(t+1)) = 0 \tag{26}$$

and remembering that $\mathbf{x}_\Pi^w(t+1) = \mathbf{x}_\Pi(t) = \mathbf{x}(t)$, we end up with the usual epipolar constraint (11), where now the matrix $\mathbf{Q} = T\wedge$ is now just a skew-symmetric matrix depending upon translation

$$\begin{cases} \mathbf{x}^{i^w}(t+1)\mathbf{Q}(t)\mathbf{x}^i(t) = 0 \\ \mathbf{y}^i(t) = \mathbf{x}^i(t) + n_i(t) \end{cases} \qquad \mathbf{Q} = T\wedge \in so(3). \tag{27}$$

The effect of rotation has been canceled out by the image warping.

5.2 Modification induced on the essential manifold

We have seen that the plane-fixation constraint corresponds to essential matrices which are of the form $\mathbf{Q} = T\wedge$. Due to the normalization constraint on T, we have only two degrees of freedom left, and rotation has been fully decoupled from translation.

If we follow the interpretation of the essential manifold as the tangent bundle of the rotation group, presented in [13], we can give a simple geometric plot of the effect of the plane-fixation constraint on the essential manifold. In particular, each essential matrix $\mathbf{Q} = T \wedge R$ is a tangent vector in the direction $T\wedge$ to the point R of the set of rotation matrices $SO(3)$. The tangent plane to the origin (identity matrix) of the rotation group is just the set of skew-symmetric matrices $so(3)$, which is the lie algebra corresponding to the lie group $SO(3)$. Now the effect of the plane-fixation constraint is that of mapping an arbitrary tangent vector to $SO(3)$ at an arbitrary point, onto a tangent vector to the origin by right-operation (see figure 2).

Therefore, among all possible tangent vectors at all possible rotations (i.e. among all possible essential matrices), the ones that correspond to a plane-fixation situation are all and only the ones that are tangent to the origin (identity).

5.3 Motion estimation under plane-compensation

The plane-compensation has the effect of decoupling rotation from translation. Any motion estimation scheme based upon the epipolar constraint, with the parameters on $so(3)$ – the space of 3×3 skew-symmetric matrices, estimates the two parameters corresponding to the direction of translation. Note that such schemes would be *linear*, for $so(3)$ is isomorphic to $\mathbb{R}^3$ (i.e. there is a linear and bijective transformation between matrices $S \in so(3)$ and vectors $T \in \mathbb{R}^3$, which is indeed $S = T\wedge$). Rotation can be estimated separately from the parameters of the plane-compensation, as we will see in the next sections.

5.4 Plane-compensation: quadratic warping

In this section we formulate a differential constraint on the projection of points on the plane Π. This constraint can be used for finding the transformation of the projective coordinates of points on the plane along time. The transformation can be inverted in order to compensate for the motion and maintain the points on the plane fixed in image coordinates.

Consider the generic point $\mathbf{X}_\Pi \in \Pi$. At a generic time instant, due to the motion of the camera with translational velocity V and rotational velocity Ω, its coordinates change in the viewer's reference according to

$$\dot{\mathbf{X}}_\Pi(t) = \Omega(t) \wedge \mathbf{X}_\Pi(t) + V(t) \tag{28}$$

where V, Ω are related to T and R via exponential coordinates [9]. Since $\mathbf{X}_\Pi \in \Pi$, it must be $\mathbf{a}^T \mathbf{X}_\Pi = 1$ and therefore

$$\frac{1}{Z_\Pi} = \mathbf{a}^T \mathbf{x}_\Pi \tag{29}$$

so that the motion field for points X_Π^i on the plane can be written as

$$\dot{\mathbf{x}}_\Pi^i(t) = \left[\mathbf{a}^T \mathbf{x}_\Pi^i \mathcal{A}_i \mid \mathcal{B}_i\right] \begin{bmatrix} V(t) \\ \Omega(t) \end{bmatrix} \tag{30}$$

where

$$\mathcal{A}_i \doteq \begin{bmatrix} 1 & 0 & -x_i \\ 0 & 1 & -y_i \end{bmatrix} \qquad \mathcal{B}_i \doteq \begin{bmatrix} -x_i y_i & 1 + x_i^2 & -y_i \\ -1 - y_i^2 & x_i y_i & x_i \end{bmatrix}. \tag{31}$$

We can rewrite an alternative expression for the optical flow as

$$\dot{\mathbf{x}}_\Pi = \mathbf{A}(\mathbf{a}, V, \Omega)[1 \ x \ y \ xy \ x^2 \ y^2]^T \doteq \mathbf{A}(\mathbf{a}, V, \Omega) u(\mathbf{x}_\Pi) \tag{32}$$

where $\mathbf{A}$ is a 2×6 matrix that depends upon the choice of the plane Π and the motion of the viewer V, Ω:

$$\mathbf{A} \doteq \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & 0 \\ a_6 & a_7 & a_8 & a_5 & 0 & a_4 \end{bmatrix}. \tag{33}$$

Now, given a number of flow vectors $\dot{\mathbf{x}}_i$ at a number of locations $\mathbf{x}_i$, one may solve via linear least-squares for the 8 parameters of $\mathbf{A}$ *without imposing any structure* on them.

Alternatively, one may use the above constraint for two other purposes: one for estimating a best plane-fit from correspondences, by decoupling the plane parameters $\mathbf{a}$ from the motion parameters, and another for estimating ego-motion when the visible structure lies on a plane, by decoupling motion from the plane parameters. This will be done in the next two sections.

We end this section by defining the "warp operation" on a generic image point $\mathbf{x}$ (not image of a point on the reference plane) as

$$\mathbf{x}^w(t+1) \doteq \mathbf{x}(t+1) - \mathbf{A} u(\mathbf{x}(t)) \tag{34}$$

Note that, if the point $\mathbf{x}_\Pi \in \Pi$, then we have

$$\mathbf{x}_\Pi^w(t+1) \doteq \mathbf{x}_\Pi(t+1) - \mathbf{A} u(\mathbf{x}_\Pi(t)) = \mathbf{x}_\Pi(t) \tag{35}$$

provided that we approximate the derivative with the first difference. In the presence of strong temporal aliasing, we can refine the warping iteratively, by applying it over and over on the residual image motions.

5.4.1 Direct methods for quadratic warping from image brightness

Note that the warping can also be performed directly from image intensities. In fact, from the image brightness constraint equation

$$\frac{d}{dt}I(\mathbf{x},t) = 0 \tag{36}$$

we get

$$\nabla_{\mathbf{x}} I(\mathbf{x},t)\dot{\mathbf{x}} + I_t = \nabla_{\mathbf{x}} I(\mathbf{x},t)\mathbf{A}u(\mathbf{x}) + I_t = 0 \tag{37}$$

which is a constraint that can be solved in a least-squares sense for the parameters of the matrix $\mathbf{A}$.

5.5 Motion-independent plane fitting

Consider the expression of the motion field (30), which we rewrite as

$$\dot{\mathbf{x}} = \mathcal{C}(\mathbf{x},\mathbf{a}) \begin{bmatrix} V \\ \Omega \end{bmatrix}. \tag{38}$$

Given the above constraint at a sufficient number of locations $\mathbf{x}$, we can solve for motion as a function of the plane parameters $\mathbf{a}$, and substitute back the result, ending up with a subspace constraint involving only the plane parameters $\mathbf{a}$ and measured image coordinates:

$$\begin{bmatrix} \hat{V} \\ \hat{\Omega} \end{bmatrix} = \mathcal{C}^{\dagger}(\mathbf{x},\mathbf{a})\dot{\mathbf{x}} \tag{39}$$

$$\mathcal{C}^{\perp}(\mathbf{x},\mathbf{a})\dot{\mathbf{x}} = 0 \qquad\qquad \mathbf{a} \in \mathbb{R}^3 \tag{40}$$

where $\mathcal{C}^{\perp} \doteq I - \mathcal{C}\mathcal{C}^{\dagger}$. The above is an implicit dynamical system with parameters $\mathbf{a}$, and the Essential filter [13] provides a principled way for estimating the parameters from the above model.

5.5.1 Direct methods for plane fitting

The same fitting can be accomplished directly from image brightness derivatives. From the brightness constraint we have

$$\nabla_{\mathbf{x}} I(\mathbf{x},t)\dot{\mathbf{x}} + I_t = \nabla_{\mathbf{x}} I(\mathbf{x},t)\mathcal{C}(\mathbf{x},\mathbf{a}) \begin{bmatrix} V \\ \Omega \end{bmatrix} + I_t = \mathcal{G}(\mathbf{x},\nabla_{\mathbf{x}} I,\mathbf{a}) \begin{bmatrix} V \\ \Omega \end{bmatrix} + I_t = 0 \tag{41}$$

which can be solved again for the motion parameters and substituted back in order to get the implicit dynamic constraint

$$\mathcal{G}^{\perp}(\mathbf{x},\nabla_{\mathbf{x}} I,\mathbf{a})I_t = 0 \tag{42}$$

which depends only upon the plane parameters and the image brightness derivatives and can be fed into an Essential filter in order to estimate the plane parameters $\mathbf{a}$ recursively.

5.6 Motion from planar structure

The expression of the motion field (30) can be reinterpreted in order to formulate a constraint only on the motion parameters and not involving the plane parameters. To this end, we write the optical flow as as

$$\dot{\mathbf{x}} = \tilde{\mathcal{C}}(\mathbf{x}, V) \begin{bmatrix} \mathbf{a} \\ \Omega \end{bmatrix} \tag{43}$$

where $\tilde{\mathcal{C}}(\mathbf{x}, V) \doteq [\mathcal{A}(\mathbf{x})V\mathbf{x}^T \mid \mathcal{B}(\mathbf{x})]$. We can now follow the same procedure as in the previous section, in order to derive a constraint only on the motion components and image velocities

$$\tilde{\mathcal{C}}^{\perp}(\mathbf{x}, V)\dot{\mathbf{x}} = 0 \tag{44}$$

that can be fed into an essential filter in order to estimate the direction of translation $V \in \mathbf{S}^2$. The same procedure can be performed directly from image brightness, from the constraint

$$\tilde{\mathcal{G}}^{\perp}(\mathbf{x}, \nabla_{\mathbf{x}} I, V) I_t = 0 \tag{45}$$

where $\tilde{\mathcal{G}} \doteq \nabla_{\mathbf{x}} I \ \tilde{\mathcal{C}}$.

References

[1] 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.

[2] 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.

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

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

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

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

[7] B.K.P. Horn. Relative orientation. *Int. J. of Computer Vision*, 4:59–78, 1990.

[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] P. Anandan R. Kumar and K. Hanna. Shape recovery from multiple views: a parallax based approach. *Proc. of the Image Understanding Workshop*, 1994.

[11] 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.

[12] H. S. Sawhney. Simplifying motion and structure analysis using planar parallax and image warping. *Proc. of the Int. Conf. on Pattern Recognition*, 1994.

[13] 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.

[14] S. Soatto and P. Perona. Motion from fixation. CDS Technical report CIT-CDS-95-006, California Institute of Technology, February 1995.

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

[16] J. Weng, T. Huang, and N. Ahuja. Motion and structure from two perspective views: algorithms, error analysis and error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(5):451–476, 1989.

[17] 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.