# CDS

**TECHNICAL MEMORANDUM NO. CIT-CDS 95-009**
**March, 1995**

# Motion from "X" by Compensating "Y"*

*Stefano Soatto* and *Pietro Perona*

**Control and Dynamical Systems**
**California Institute of Technology**
**Pasadena, CA 91125**

# 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

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

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.