"MFm": 3-D Motion From 2-D Motion Causally Integrated Over Time Part I: Theory

Alessandro Chiuso Stefano Soatto

Washington University, One Brookings Dr. 1127, St. Louis - MO 63130
tel. (314)935-7340, fax. (314)935-7550, email soatto@essrl.wustl.edu

Keywords: Visual motion, structure from motion, active and real-time vision, shape, geometry.

Abstract

We introduce a provably minimal and stable nonlinear filter for causally estimating the three-dimensional motion of a scene in real time from a sequence of two-dimensional images. This problem is subject to fundamental tradeoffs between the the ease in solving the so-called "correspondence problem" and the robustness of the resulting algorithm. This paper and its companion [7] aim at addressing this fundamental tradeoff. We contend that it is possible to integrate visual information over time, hence achieving a global estimate of 3-D motion, while maintaining the correspondence problem local. Among the obstacles we encounter is the fact that individual points tend to become occluded during motion, while novel points become visible.

Although there exists no finite-dimensional optimal solution to this problem, it is our goal to provide algorithms that work in practice as well as in theory. Our contributions towards this end can be summarized into four parts. We first study (1) the conditions that are necessary in order to be able to causally reconstruct structure and motion. While this problem has been addressed before, we give a novel and simpler proof that provides geometric insight and an explicit characterization of the ambiguities. We then (2) prove, for the first time, uniform observability of motion and structure; this result is crucial for the (3) proof of stability of the algorithm that we propose. In passing, we show (4) how the conditions we impose on our models are tight: imposing either more or less results in either a biased or an unstable filter. This paper is concerned with theory. In a companion to the present paper [7] we describe a complete real-time *implementation* of the filter, which includes an approach to causally handle *occlusions* and experiments on long sequences of real images.

Our approach is related to several previous contributions presented in the literature, although we compensate for their shortcomings which include handling occlusions and overcoming stability problems due to the sub-minimality of the models previously employed.

1 Introduction

We are interested in using vision as a sensor for machines to interact with the environment by moving, tracking, manipulating objects etc. In order to do so, a machine must be able to estimate its three-dimensional (3-D) motion relative to the scene and – to an extent that depends upon the application – the shape of the scene.

Inferring the three-dimensional shape of a moving scene from its (two-dimensional) images is one of the classical problems of Computer Vision, known by the name of "Shape From Motion" (SFM). It is somewhat of an unfortunate name, since what needs to be inferred from the images, along with scene shape, is (3-D) motion itself; over the years, research in SFM seems to have crystallized on describing scenes as a collection of *point-features*, hardly a meaningful representation of the complexity of the visible world. Perhaps a more appropriate name for this problem – as someone suggested – would be "(3-D) Motion From (2-D) Motion"

or, for the fond of the acronym, MFm. Ironically, this paper follows the crowd by using a point-wise representation of the environment. Our apology is that – in this paper – we care more for estimating 3-D motion than shape; a crude representation of the latter comes as a byproduct. Furthermore, even for such a simplistic model as a set of point features, reconstructing shape is remarkably difficult. Impossible indeed – in a sense – as we argue in section 1.2.

SFM (or MFm) is subject to fundamental tradeoffs. As we articulate in section 1.3, when the so-called "baseline" is long, estimating relative orientation is simple, provided that image-motion is given (the infamous "correspondence problem"). However, solving the correspondence problem is appallingly difficult, for it amounts to a global matching problem – all too often solved by hand – which spoils the possibility of use in real-time control systems. When the images are collected closely in time, on the other hand, the correspondence problem becomes an easy local variational problem. However, estimating 3-D motion becomes rather difficult since – on small motions – the noise in the image overwhelms the feeble information contained in the 2-D motion of the features. This paper and its companion [7] aim at addressing this fundamental tradeoff. We contend that it is possible to integrate visual information over time, hence achieving a global estimate of 3-D motion, while maintaining the correspondence problem local. Among the obstacles we encounter is the fact that individual points tend to become occluded during motion, while novel points become visible.

While we show how information can be integrated causally over time, we have to tone down our hopes of being able to do so *optimally*, for there exists no finite-dimensional optimal solution to this problem. Therefore, we have to resort to *approximations*. However, "approximate" does not mean "approximative": it is our goal to provide algorithms that work in practice as well as in theory, in the sense of being provably stable and efficient. This paper aims at setting a small step in this direction. Our contributions can be summarized into four parts. On the *observability* of shape and motion, we provide a novel and considerably simpler proof of the (previously known) global observability, but we also for the first time prove uniform observability. We use it to characterize the *minimal realization* of the model, and describe its geometric properties. These results are crucial for proving the *stability* of the estimation algorithm that we propose (a *nonlinear filter*). Finally, in a companion to the present paper [7], we offer a complete real-time *implementation* of the algorithms, which includes an approach to causally handle *occlusions*.

1.1 A first formalization of the problem

Consider an N-tuple of points in the three-dimensional Euclidean space, represented as a matrix

$$\mathbf{X} \doteq \left[\begin{array}{cccc} \mathbf{X}^1 & \mathbf{X}^2 & \dots & \mathbf{X}^N \end{array} \right] \in \mathbb{R}^{3 \times N} \tag{1}$$

and let them move under the action of a rigid motion[1]

$$\{g(\tau)\}_{\tau \in [0,t]} \in SE(3). \tag{2}$$

Associated to each motion $g(t)$ there is a velocity[2] $\widehat{v}(t) \in se(3)$. Under such velocity, motion evolves according to[3]

$$g(t+1) = \exp(\widehat{v}(t))g(t) \qquad \widehat{v}(t) \in se(3). \tag{4}$$

[1] $SE(3)$ stands for the Special Euclidean group of rigid motions in $\mathbb{R}^3$, which is represented by a translation vector T and a rotation matrix R: $g = \{T, R\}$. Rotation matrices are orthogonal with unit determinant and form a group, called Special Orthogonal group: $R \in SO(3) = \{R \mid R^T R = RR^T = I\}$. The Euclidean group acts on the coordinates of each point via $g(t)\mathbf{X}^i = R(t)\mathbf{X}^i + T(t)$.

[2] Spatial velocity is represented by a vector of linear velocity V and a skew-symmetric matrix $\widehat{\omega}$ of rotational velocity. Skew-symmetric 3×3 matrices are denoted by $so(3) = \{A \in \mathbb{R}^{3\times 3} \mid A^T = -A\}$. They are isomorphic to three-dimensional vectors via the "hat" operator, which represents the cross product: $\widehat{\mathbf{X}}^i \mathbf{X}^j = \mathbf{X}^i \times \mathbf{X}^j$. In fact, for each three-dimensional vector $\mathbf{a} = [a_1,\ a_2,\ a_3]^T$ there is a unique skew-symmetric matrix $\widehat{\mathbf{a}}$ given by

$$\widehat{\mathbf{a}} = \left[\begin{array}{ccc} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{array} \right]. \tag{3}$$

Vice-versa, every skew-symmetric 3×3 matrix has this form and can thus be represented by the vector $\mathbf{a}$.

[3] The exponential can be computed in closed form using Rodrigues' formula (21).

We assume that - to an extent discussed in later sections - the *correspondence problem* is solved, that is we know which point corresponds to which in different projections (views). Equivalently, we assume that we can measure the (noisy) projection[4]

$$\mathbf{y}^i(t) = \pi\left(g(t)\mathbf{X}^i\right) + \mathbf{n}^i(t) \in \mathbb{R}^2 \quad \forall\, i = 1 \dots N \tag{5}$$

where we know the correspondence $\mathbf{y}^i \leftrightarrow \mathbf{X}^i$. Finally, by organizing the time-evolution of the configuration of points and their motion, we end up with a discrete-time, non-linear dynamical system:

$$\begin{cases} \mathbf{X}(t+1) = \mathbf{X}(t) & \mathbf{X}(0) = \mathbf{X}_0 \in \mathbb{R}^{3\times N} \\ g(t+1) = \exp(\widehat{v}(t))g(t) & g(0) = g_0 \in SE(3) \\ v(t+1) = v(t) + \alpha(t) & v(0) = v_0 \in se(3) = \mathbb{R}^6 \\ \mathbf{y}^i(t) = \pi\left(g(t)\mathbf{X}^i(t)\right) + \mathbf{n}^i(t) & \mathbf{n}^i(t) \sim \mathcal{N}(0, \Sigma_n) \end{cases} \tag{6}$$

where $\sim \mathcal{N}(M, S)$ indicates that a vector is distributed normally with mean M and covariance S. In the above system, α is the relative acceleration between the viewer and the scene. If some prior modeling information is available (for instance when the camera is mounted on a vehicle or on a robot arm), this is the place to use it. Otherwise a statistical model can be employed. In particular, one way to represent analytically the fact that no information whatsoever is available is by modeling α as a white noise process. This is what we do in this paper[5]. In principle one would like - at least for this simplified formalization of Structure From Motion - to find the "best" solution, which corresponds to the optimal estimate (in some sense) of the state of the above system $\{\mathbf{X}, g, v\}$ given a sequence output measurements (correspondences) over an interval of time. We call an algorithm that delivers the optimal estimate of the state at time t *causally* (i.e. based upon measurements up to time t) the *optimal filter*.

1.2 Critique and extensions

There are numerous reasons why the above formalization is altogether simplistic from the point of view of the vision scientist, chief the fact that the position of N points in space is hardly a satisfactory representation of the shape of a scene. Furthermore, we have assumed that the scene is a single rigid object (or that it has been *segmented* into rigid objects and we restrict the attention to one of them), and that we know the *correspondence* between points; assumptions that are all but realistic in any scene of practical interest.

However, taking the stance of the mathematician, one would like to find the richest instance of the problem that can be solved rigorously. Unfortunately, *even for the simple case just outlined, the "right" solution does not exist* [6]. Therefore, before proceeding onto formulations of the problem that capture richer visual phenomena, we feel the need to say what can be said rigorously at least on this simple instance.

From the stance of the engineer, we would like to offer an analysis that results in robust and efficient algorithms with guaranteed performance in their domain, that can be implemented in real-time and inserted into the sensing-action loop of autonomous control systems. We regard this as a useful contribution to the community, even though the system we propose will not capture the complexity of the motion of a silk gown or that of the foliage of a tree.

In passing, we remark that some of the ideas set forth in this paper can be extended to a representation of the scene where objects are surfaces chosen within a parametric class of models, and to higher-order deterministic and stochastic models for motion, as well as to different projection models, including partially calibrated ones [7].

---

[4] We take as projection model an ideal pinhole, so that $\mathbf{y} = \pi(\mathbf{X}) = \left[\begin{array}{cc} \frac{X_1}{X_3} & \frac{X_2}{X_3} \end{array}\right]^T$. This choice is not crucial and the discussion can be easily extended to other projection models (e.g. spherical, orthographic, para-perspective, etc.) including partially unknown ones (self-calibration). We do not distinguish between $\mathbf{y}$ and its projective coordinate (with a 1 appended), so that we can write $\mathbf{X} = \mathbf{y}X_3$.

[5] We wish to emphasize that this choice is not crucial towards the conclusions reached in this paper. Any other model would do, as long as the overall system is observable.

[6] In the presence of noise, the state of the model (6) can be represented as a stochastic process. In our case such a process evolves on a differentiable manifold (the product of the configuration space and the Lie group of rigid motions). The evolution of the state induces an evolution of its conditional density, which is represented by a partial differential operator known as Fokker-Planck operator. It can be shown that – under a wide range of conditions that include model (6)– there do not exist densities that are invariant under such operator, and therefore no finite-dimensional solution to the optimal filtering problem can be found (see [8]).

1.3 Tradeoffs in Structure From Motion

The distance between the centers of projection of two views is commonly referred to as "baseline" (or "parallax") between the two images. As we will see in section 2, in order to be able to recover the structure of the scene it is necessary that the baseline be non-zero. In the presence of noise, the larger the parallax the higher the "signal-to-noise ratio". For a relatively *large baseline* - such as that of sequences of snapshots taken from largely different viewpoints - *estimating structure and motion is easy.* However, solving the *correspondence problem is difficult*, if not impossible, without expert intervention of a human operator.

On the other hand, for a *small baseline* - such as the inter-frame parallax of sequences taken by a moving camera with a fast sampling rate (30-60 Hz) - the *correspondence problem is simple to solve*, at least locally in space and time (more on this later). However, *estimating structure and motion is quite difficult* since the effects of noise are overwhelming (up to 1000% of the average signal in sequences commonly encountered in real life) [7]. A meaningful scheme for time integration of visual information must result in an effective increase of the baseline, while using local information from each frame.

No matter how one chooses to increase the baseline in order to bypass the tradeoff with correspondence, however, one inevitably runs into deeper problems, namely the fact that individual feature points can *appear and disappear due to occlusions*, or to changes in their appearance due to specularities, changes in the light distribution, shadows etc. We discuss this issue in the companion paper [7].

1.4 Relation to previous work and organization of the paper

Since the optimal solution to the filtering problem for the model (6) does not exist, one has to resort to *approximations.* There are in the literature countless ways to approximate the optimal filter, including numerical integration, Monte Carlo methods, and the use of parametric densities (splines, sums of Gaussians etc.) (see for instance [21, 5, 16, 30, 2] and references therein). Each of these techniques, however, has shortcomings since they are for *general purpose* and therefore do not exploit the specifics of our problem. In particular, our models can reach a very high dimension for the state (in the order of several hundreds), which makes the use of numerical integration and Monte Carlo methods prohibitive for real-time processing. Furthermore, seeking a representation of the whole conditional density of the state may be an overkill: for most practical purpose one is content with a point-estimate (ideally the maximum likelihood) and a measure of the uncertainty in the estimate. For this reason, we concentrate on *wide-sense* estimation, by designing an approximate filter for the mode of the conditional density of the state and its dispersion about the mode [8].

We are interested in estimating motion so that we can use the estimates to accomplish spatial control tasks such as moving, tracking, grasping etc. In order to do so, the estimates must be provided *in real-time and causally*, while we can rely on the fact that images are taken at adjacent instants in time and the relative motion between the scene and the viewer is somewhat smooth (rather than having isolated "snapshots"). Therefore, we do not compare our algorithms with batch multi-frame approaches for Structure From Motion (such as those based upon multi-linear geometry). If one can afford the time for processing sequences of images off-line, of course a batch approach that optimizes simultaneously on all frames will perform better![9]

Our work falls within the category of causal motion and structure estimation (also referred to as "recursive", or "Kalman-filter based" methods), that has a long history. To our knowledge, Dickmanns and Gennery were the first to address the causal estimation of motion [12, 9], confined to structured environments (objects with known shape in the case of Gennery, freeways with structured shape in the case

[7] There are many heuristics to bypass this tradeoff. For instance, one could track individual feature-points from frame to frame in a sequence, but start processing them only when the baseline is "large enough" (thereby discarding information from intermediate frames). A more principled way to proceed is to *increase the baseline by integrating visual information over time.* Notice that time-integration does not mean time-averaging: if the noise is such that estimation between adjacent frames is spoiled (the residual cost being minimized is flat), their average is meaningless.

[8] It can be argued that a wide-sense approach is sensible only if the posterior density is unimodal. We have been unable to prove any general properties on the posterior; although we know that it is possible, in principle, for it to be multi-modal, in all the experiments performed we have never experienced a splitting of the mode.

[9] One may argue that batch approaches are now fast enough that they can be used for real-time processing. Our take on this issue is exposed in the next section, where we argue that speed is not the problem; robustness is, especially when images are taken at frame-rate from a moving camera and therefore the baseline is short.

of Dickmanns). The past fifteen years have seen a proliferation of recursive schemes to estimate Euclidean structure from known motion [20], motion from known structure [6, 25], or both simultaneously [22, 33, 11, 31, 26, 10, 15, 37, 14, 32, 35, 34, 17, 28, 19, 1, 4, 24, 36, 18] just to cite a few. The first attempts to prove stability of the schemes proposed are not until recent [23]. However, few of the schemes cited addresses occlusions, which make them prone to the tradeoffs described in section 1.3 and therefore hardly usable in realistic scenes where occlusions are the norm. The first attempts to handle occlusions in a causal scheme[10] came only a few years ago: McLauchlan [22] proposed a filter with variable state, that however requires a batch initialization, while Soatto and Perona [29] proposed several schemes in which the problem of occlusions was bypassed by eliminating structure from the model. Their scheme, however, had the scale factor tied to motion, rather than to shape, and therefore could not exploit the invariance of shape in order to achieve a large effective baseline. Our approach is similar in spirit to the work of Azarbayejani and Pentland [4], extended to handle occlusions. In addition, the model in [4] is sub-minimal which, as we explain in section 4.2, results in an unstable filter. Furthermore, it gives "infinite weight" to the measurements at the initial instant. We correct all the above issues here.

The first part of this study is concerned with *analysis*. We first study the conditions that are necessary in order to be able to causally reconstruct structure and motion (section 2). While the observability of structure from motion has been addressed before, we give a novel and simpler proof that provides geometric insight and an explicit characterization of the ambiguities. In section 3 we prove uniform observability for the first time; this result is crucial for the proof of stability of the algorithm that we propose in section 4.1. In passing, we show how the conditions we impose on our models are tight: imposing either more or less results in either a biased or an unstable filter (section 4.2).

The second part, which we present in a companion paper [7] is concerned with the *implementation* of a system for functioning in real time on real scenes. This paper shares general motivations with its companion [7], so that parts of the introduction are common to the two.

2 Observability

To what extent can the 3-D shape and motion of a scene be reconstructed *causally* from measurements of the motion of its projection onto the sensor? This is the subject of this section, which we start by establishing some notation that will be used throughout the rest of the paper.

2.1 Preliminaries

Let $g \in SE(3)$ indicate an element of the Euclidean group of rigid motions, represented by a translation vector $T \in \mathbb{R}^3$ and a rotation matrix $R \in SO(3)$, and let $\alpha \neq 0$ be a scalar. The *similarity group*, which we indicate by $g_\alpha \in SE(3) \times \mathbb{R}_+$ is the composition of a rigid motion and a scaling, which acts on points in $\mathbb{R}^3$ as follows: $g_\alpha(\mathbf{X}) = \alpha R\mathbf{X} + \alpha T$. We also define an action of g_α on $SE(3)$ as $g_\alpha(g') = \{\alpha RT' + \alpha T, RR'\}$ and an action on $se(3)$ as [2] $g_\alpha(v) = \{\alpha V, \widehat{\omega}\}$. The similarity group, acting on an N-tuple of points in $\mathbb{R}^3$, generates an equivalence class:

$$[\mathbf{X}] = \{\mathbf{Y} \in \mathbb{R}^{3\times N} \mid \exists\, g_\alpha \mid \mathbf{Y} = g_\alpha \mathbf{X}\} \tag{7}$$

two configurations of points $\mathbf{X}$ and $\mathbf{Y} \in \mathbb{R}^{3\times N}$ are equivalent if there exists a similarity transformation g_α that brings one onto the other: $\mathbf{Y} = g_\alpha \mathbf{X}$. Such equivalence class in (7) is called a *fiber*, and the collection of all fibers is called a *fiber bundle*, or *homogeneous space* [13]. Therefore, the similarity group organizes the space of N-tuples into a fiber bundle, which we call the *state-space bundle*: given a point $\mathbf{X}$ in $\mathbb{R}^{3\times N}$, it belongs to one and only one fiber. From any given point it is possible to move either along the fiber (via the similarity group) or across fibers. One element of each fiber is sufficient to represent it, since all other elements are just transformed versions of it via the similarity group. In order to obtain a representation of the whole bundle, however, we need a consistent way of choosing a representative for each fiber. This is called a *base* of the fiber bundle (see [13]).

[10] There are several natural ways of handling missing data in a batch approach: since they do not extend to causal processing, we do not review them here.

Consider now a discrete-time nonlinear dynamical system of the form

$$\begin{cases} \xi(t+1) = f(\xi(t)) & \xi(t_0) = \xi_0 \\ y(t) = h(\xi(t)) \end{cases} \tag{8}$$

and let $y(t; t_0, \xi_0)$ indicate the output of the system at time t, starting from the initial condition ξ_0 at time t_0. In the next section we want to characterize the states ξ that can be "reconstructed" from the measurements y. Such a characterization depends upon the structure of the system f, h and not on the measurement noise, which is therefore assumed to be absent for the purpose of analysis in this section.

Definition 1 *Consider a system in the form (8) and a point in the state-space ξ_0. We say that ξ_0 is* indistinguishable *from ξ_0' if $y(t; t_0, \xi_0') = y(t; t_0, \xi_0) \quad \forall\ t, t_0$. We indicate with $\mathcal{I}(\xi_0)$ the set of initial conditions that are indistinguishable from ξ_0.*

Definition 2 *We say that the system (8) is* observable up to a (group) transformation ψ *if $\mathcal{I}(\xi_0) = [\xi_0] \doteq \{\xi_0' \mid \exists\ \psi \mid \xi_0' = \psi(\xi_0)\}$.*

Clearly, from measurements of the output $y(t)$ over any period of time, it is possible to recover at most the equivalence class (fiber) where the initial condition belongs, that is $\mathcal{I}(\xi_0)$, but not ξ_0 itself. The only case when this is possible is when the system is observable up to the identity transformation. In this case we have that $\mathcal{I}(\xi_0) = \{\xi_0\}$ and we say that the system is *observable*.

For a generic linear time-varying system of the form

$$\begin{cases} \xi(t+1) = F(t)\xi(t) & \xi(t_0) = \xi_0 \\ y(t) = H(t)\xi(t) \end{cases} \tag{9}$$

we define the *k-observability Grammian* as $M_k(t) \doteq \sum_{i=t}^{t+k} \Phi_i^T(t) H^T(t) H(t) \Phi_i(t) \quad \forall\ i\ >\ t$ where $\Phi_t(t) = I$ and $\Phi_i(t) \doteq F(i-1) \dots F(t)$. The following definition will come handy in section 4.1:

Definition 3 *We say that the system (9) is* uniformly observable *if there exist real numbers $m_1 > 0,\ m_2 > 0$ and an integer $k > 0$ such that $m_1 I \leq M_k(t) \leq m_2 I\ \forall\ t$.*

2.2 Structure from motion is observable up to a similarity

The following theorem revisits the well-known fact that, under constant velocity, structure and motion are (causally) observable up to a (global) similarity transformation.

Proposition 1 *The model (6) where the points $\mathbf{X}$ are in general position is observable up to a similarity transformation of $\mathbf{X}$ provided that $V_0 \neq 0$. In particular, the set of initial conditions that are indistinguishable from $\{\mathbf{X}_0, g_0, v_0\}$, where $g_0 = \{T_0, R_0\}$ and $e^{\widehat{v_0}} = \{V_0, U_0\}$, is given by $\{\tilde{R}\mathbf{X}_0\alpha + \tilde{T}\alpha, \tilde{g}_0, \tilde{v}_0\}$, where $\tilde{g}_0 = \{T_0\alpha - R_0\tilde{R}^T\tilde{T}\alpha, R_0\tilde{R}^T\}$ and $e^{\widehat{\tilde{v}_0}} = \{V_0\alpha, U_0\}$ for an arbitrary α and $\tilde{T}, \tilde{R}$.*

Proof: *Consider two initial conditions $\{\mathbf{X}_1, g_1, v_1\}$ and $\{\mathbf{X}_2, g_2, v_2\}$. For them to be indistinguishable we must have $\mathbf{y}(t) = \pi(g_1(t)\mathbf{X}_1(t)) = \pi(g_2(t)\mathbf{X}_2(t)) \quad \forall\ t \geq 0$. In particular, at time $t = 0$ this is equivalent to the existence of a diagonal matrix of scalings, $A(1)$ such that $g_1(0)\mathbf{X}_1(0) = (g_2(0)\mathbf{X}_2) \cdot A(1)$, where the operator $\cdot$ performs the scaling according to $(g\mathbf{X}) \cdot A \doteq RXA + TA$. Under the assumption of constant velocity, we have that $g(t) = e^{\widehat{tv}} g(0)$, and therefore the group action g only appears at the initial time. Consequently, we drop the time index and write simply g_1 and g_2 as points in $SE(3)$. At the generic time instant t, the indistinguishability condition can therefore be written as $e^{\widehat{tv_1}} g_1 \mathbf{X}_1 = (e^{\widehat{tv_2}} g_2 \mathbf{X}_2) \cdot A(t+1)$. Therefore, given $\mathbf{X}_2, g_2, v_2$, in order to find the initial conditions that are indistinguishable from it, we need to find $\mathbf{X}_1, g_1, v_1$ and $A(k), k \geq 1$ such that, after some substitutions, we have*

$$\begin{cases} g_1\mathbf{X}_1 = (g_2\mathbf{X}_2) \cdot A(1) \\ e^{\widehat{v_1}} e^{(k-1)\widehat{v_2}} g_2\mathbf{X}_2 = \left(e^{\widehat{v_2}} e^{(k-1)\widehat{v_2}} g_2\mathbf{X}_2\right) \cdot A(k+1) \quad k \geq 1. \end{cases} \tag{10}$$

Making the representation of $SE(3)$ explicit, we write the above conditions as

$$\begin{cases} R_1\mathbf{X}_1 + \bar{T}_1 = (R_2\mathbf{X}_2 + \bar{T}_2)A(1) \\ U_1\tilde{\mathbf{X}}_k A(k) + \bar{V}_1 = U_2\tilde{\mathbf{X}}_k A(k+1) + \bar{V}_2 A(k+1) \end{cases} \tag{11}$$

where we have defined $\tilde{\mathbf{X}}_k \doteq e^{(k-1)\widehat{v_2}} g_2 \mathbf{X}_2$ *which, by the assumption of general position, is of full rank 3, and* $\bar{V}$ *denotes the rank-one matrix* $\bar{V} \doteq V I_N^T$*, where* I_N *is the* N*-dimensional vector of ones. We can rewrite the second of the equations above in a more enlightening way as follows:*

$$\tilde{\mathbf{X}}_k A(k) A^{-1}(k+1) - U_1^T U_2 \tilde{\mathbf{X}}_k = U_1^T (\bar{V}_2 A(k+1) - \bar{V}_1) A^{-1}(k+1). \tag{12}$$

The $3 \times N$ *matrix on the right hand-side has rank at most 2, while the left hand-side has rank 3, following the general-position conditions, unless* $A(k)A^{-1}(k+1) = I$ *and* $U_1^T U_2 = I$*, in which case it is identically zero. Therefore, both terms in the above equations must be identically zero. From* $U_1^T U_2 = I$ *we conclude that* $U_1 = U_2$*, while from* $A(k)A^{-1}(k+1) = I$ *we conclude that* $A(k)$ *is constant. However, the right hand-side imposes that* $\bar{V}_2 A = \bar{V}_1$*, or in vector form* $V_2 a^T = V_1 I_N^T$ *where* $A = diag\{a\}$*, which implies that* $A = \alpha I$*, i.e. a multiple of the identity. Now, going back to the first equation in (11), we conclude that* $R_1 = R_2 \tilde{R}^T$*, for any* $\tilde{R}^T \in SO(3)$*,* $X_1 = (\tilde{R} X_0 + \tilde{T})\alpha$ *for any* $\tilde{T} \in \mathbb{R}^3$*, and finally* $T_1 = (T_2 - R_1 \tilde{R}^T \tilde{T})\alpha$*, which concludes the proof.*

Remark 1 *The relevance of the above proposition for the practical estimation of Shape from Motion (where velocity is not necessarily constant) is that one can solve the problem using the above model only when velocity varies slowly compared to the sampling frequency. If, however, some information on the dynamics of the acceleration becomes available (as for instance if the camera is mounted on a support with some inertia), then the restriction on velocity can be lifted. This framework, however, will not hold if the data* $y(t)$ *are snapshots of a scene taken from sparse viewpoints.*

The following theorem states that it is possible to make the model observable by fixing the direction of three points and one depth. When we interpret the state-space as a fiber bundle under the action of the similarity group, fixating the direction of three points and one depth identifies a base of the bundle, that is a point in the similarity group. Without loss of generality (i.e. modulo a re-ordering of the states) we will assume the indices of such three points to be 1, 2 and 3. We consider a point $\mathbf{X}$ as parameterized by its direction $\mathbf{y}$ and depth ρ, so that $\mathbf{X} = \mathbf{y}\rho$.

Proposition 2 *Given the direction of three non-coplanar points,* $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3$ *and the scale of one point,* $\rho^1 > 0$*, and given vectors* ϕ^i*,* $i = 1 \dots N$*, the set of motions* $g = \{T, R\} \in SE(3)$ *and scales* $\alpha \in \mathbb{R}$ *such that*

$$\alpha R \mathbf{y}^i \rho^i + \alpha T = \phi^i \qquad \forall\, i = 1 \dots N \geq 3 \tag{13}$$

has measure zero.

Proof: *Suppose that the statement holds for* $N = 3$*, then it holds for any* $N > 3$*, as any additional equation of the form* $\phi^i = \alpha R \mathbf{y}^i \rho^i + \alpha T$ *is linear in the variable* $\mathbf{X}^i \doteq \mathbf{y}^i \rho^i$*, and therefore can be solved uniquely. Since* ${\mathbf{X}_3}^i = \rho^i$*, the latter is uniquely determined, and so is* $\mathbf{y}^i = \frac{\mathbf{X}^i}{\rho^i}$*. Therefore, we only need to prove the statement for* $N = 3$*:*

$$\begin{cases} \phi^1 = \alpha R \mathbf{y}^1 \rho^1 + \alpha T \\ \phi^2 = \alpha R \mathbf{y}^2 \rho^2 + \alpha T \\ \phi^3 = \alpha R \mathbf{y}^3 \rho^3 + \alpha T. \end{cases} \tag{14}$$

Solve the first equation for αT*,*

$$\alpha T = \phi^1 - \alpha R \mathbf{y}^1 \rho^1 \neq 0 \tag{15}$$

and substitute into the second and third equation to get

$$\begin{cases} \phi^2 - \phi^1 = \alpha R (\mathbf{y}^2 \rho^2 - \mathbf{y}^1 \rho^1) \\ \phi^3 - \phi^1 = \alpha R (\mathbf{y}^3 \rho^3 - \mathbf{y}^1 \rho^1). \end{cases} \tag{16}$$

The scale $\alpha > 0$ *can be solved for as a function of the unknown scales* ρ^2 *and* ρ^3

$$\alpha = \frac{\|\phi^2 - \phi^1\|}{\|\mathbf{y}^2 \rho^2 - \mathbf{y}^1 \rho^1\|} = \frac{\|\phi^3 - \phi^1\|}{\|\mathbf{y}^3 \rho^3 - \mathbf{y}^1 \rho^1\|} \tag{17}$$

(note that these expressions are always legitimate as a consequence of the non-coplanarity assumption). After substituting α *in equations (16), we get*

$$\begin{cases} \frac{\phi^2 - \phi^1}{\|\phi^2 - \phi^1\|} = R \frac{\mathbf{y}^2 \rho^2 - \mathbf{y}^1 \rho^1}{\|\mathbf{y}^2 \rho^2 - \mathbf{y}^1 \rho^1\|} \\ \frac{\phi^3 - \phi^1}{\|\phi^3 - \phi^1\|} = R \frac{\mathbf{y}^3 \rho^3 - \mathbf{y}^1 \rho^1}{\|\mathbf{y}^3 \rho^3 - \mathbf{y}^1 \rho^1\|}. \end{cases} \tag{18}$$

In the above equations, the only unknowns are R, ρ^2 and ρ^3. Note that, while on the left hand-side there are two fixed unit-norm vectors, on the right hand-side there are unit-norm vectors parameterized by ρ^2 and ρ^3 respectively. In particular, the right hand-side of the first equation in (18) is a vector on the unit circle of the plane spanned by $\mathbf{y}^1$ and $\mathbf{y}^2$, while the right hand-side of the second equation is a vector on the unit circle of the plane π_2 spanned by $\mathbf{y}^1$ and $\mathbf{y}^3$. By the assumption of non-coplanarity, these two planes do not coincide. We write the above equation in a more compact form as

$$\begin{cases} \phi_\nu{}^1 = R\mu_{\rho^2} \\ \phi_\nu{}^2 = R\mu_{\rho^3}. \end{cases} \tag{19}$$

Now R must preserve the angle between $\phi_\nu{}^1$ and $\phi_\nu{}^2$, which we indicate as $\widehat{\phi_\nu{}^1\phi_\nu{}^2}$, and therefore μ_{ρ^2} and μ_{ρ^2} must be chosen accordingly. If $\widehat{\phi_\nu{}^1\phi_\nu{}^2} > \pi - \widehat{\pi_1\pi_2}$, no such choice is possible. Otherwise, there exists a one-dimensional interval set of ρ^1, ρ^2 for which one can find a rotation R that preserves the angle. However, R must also preserve the cross product, so that we have

$$\phi_\nu{}^1 \times \phi_\nu{}^2 = (R\mu_{\rho^2}) \times R\mu_{\rho^3} = R\mu_{\rho^2} \times (R^T R\mu_{\rho^3}) = R(\mu_{\rho^2} \times \mu_{\rho^3}) \tag{20}$$

(note that the norm of the two cross products is the same as a consequence of the conservation of the inner product), and therefore ρ^2 and ρ^3 are determined uniquely; as a consequence, so is R, which concludes the proof.

3 Realization

In order to design a finite-dimensional approximation to the optimal filter, we need a minimal realization of (6). How to obtain it is the subject of this section.

3.1 Local coordinates

Our first step consists in characterizing the local-coordinate representation of the model (6). To this end, we represent $SO(3)$ locally in canonical exponential coordinates: let Ω be a three-dimensional real vector ($\Omega \in \mathbb{R}^3$); $\frac{\Omega}{\|\Omega\|}$ specifies the direction of rotation and $\|\Omega\|$ specifies the angle of rotation in radians. Then a rotation matrix can be represented by $\widehat{\Omega} \in so(3)$ such that $R \doteq \exp(\widehat{\Omega}) \in SO(3)$. Rodrigues' formula is a convenient way to compute the exponential:

$$\exp(\widehat{\Omega}) = I + \frac{\widehat{\Omega}}{\|\Omega\|}\sin(\|\Omega\|) + \frac{\widehat{\Omega}^2}{\|\Omega\|^2}(1 - \cos(\|\Omega\|)). \tag{21}$$

The three-dimensional coordinate $\mathbf{X}^i$ is represented by its projection onto the image plane $\mathbf{y}^i$ and its depth ρ^i, so that $\mathbf{y}^i \doteq \pi(\mathbf{X}^i) \doteq \begin{bmatrix} \frac{X_1^i}{X_3^i} & \frac{X_2^i}{X_3^i} \end{bmatrix}^T$ $\rho^i = X_3^i$. Such a representation has the advantage of decomposing the uncertainty in the measured directions $\mathbf{y}$ (low) from the uncertainty in depth ρ (high). In what follows we will not make a distinction between the homogeneous three-dimensional vector $\mathbf{y}^i \in \mathbb{R}^3/\mathbb{R}$ (with a 1 as the third element) and the corresponding two-dimensional vector $\mathbf{y}^i \in \mathbb{R}^2$, since this will be clear from the context. The model (6) in local coordinates is therefore

$$\begin{cases} \mathbf{y}_0^i(t+1) = \mathbf{y}_0^i(t) & i = 1 \dots N & \mathbf{y}_0^i(0) = \mathbf{y}_0^i \\ \rho^i(t+1) = \rho^i(t) & i = 1 \dots N & \rho^i(0) = \rho_0^i \\ T(t+1) = \exp(\widehat{\omega}(t))T(t) + V(t) & & T(0) = T_0 \\ \Omega(t+1) = Log_{SO(3)}(\exp(\widehat{\omega}(t))\exp(\widehat{\Omega}(t))) & & \Omega(0) = \Omega_0 \\ V(t+1) = V(t) + \alpha_V(t) & & V(0) = V_0 \\ \omega(t+1) = \omega(t) + \alpha_\omega(t) & & \omega(0) = \omega_0 \\ \mathbf{y}^i(t) = \pi\left(\exp(\widehat{\Omega}(t))\mathbf{y}_0^i(t)\rho^i(t) + T(t)\right) + \mathbf{n}^i(t) & & i = 1 \dots N. \end{cases} \tag{22}$$

The notation $Log_{SO(3)}(R)$ stands for Ω such that $R = e^{\widehat{\Omega}}$ and is computed by inverting Rodrigues' formula.

3.2 Minimal realization

In linear time-invariant systems one can decompose the state-space into an observable subspace and its (unobservable) complement (the so-called Kalman decomposition). In the case of our system, which is nonlinear and observable up to a group transformation, we can exploit the bundle structure of the state-space to realize a similar concept of decomposition: each base of the fiber bundle is observable, while individual fibers are not. Therefore, in order to restrict our attention to the observable component of the system, we only need to choose a base of the fiber bundle, that is a particular (representative) point on each fiber[11]. Proposition 2 suggests a way to render the model (22) observable by eliminating the states that fix the unobservable subspace.

Corollary 1 *The model*

$$\begin{cases} \mathbf{y}_0^i(t+1) = \mathbf{y}_0^i(t) & i = 4 \dots N & \mathbf{y}_0^i(0) = \mathbf{y}_0^i \\ \rho^i(t+1) = \rho^i(t) & i = 2 \dots N & \rho^i(0) = \rho_0^i \\ T(t+1) = \exp(\widehat{\omega}(t))T(t) + V(t) & & T(0) = T_0 \\ \Omega(t+1) = Log_{SO(3)}(\exp(\widehat{\omega}(t)) \exp(\widehat{\Omega}(t))) & & \Omega(0) = \Omega_0 \\ V(t+1) = V(t) + \alpha_V(t) & V(0) = V_0 & \\ \omega(t+1) = \omega(t) + \alpha_\omega(t) & \omega(0) = \omega_0 & \\ \mathbf{y}^i(t) = \pi \left(\exp(\widehat{\Omega}(t)) \mathbf{y}_0^i(t) \rho^i(t) + T(t) \right) + n^i(t) & & i = 1 \dots N. \end{cases} \tag{23}$$

which is obtained by eliminating $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3$ *and* ρ^1 *from the state of the model (22), is observable.*

The model (23) is essentially the one we are going to use in the companion paper [7] to design a nonlinear filter.

3.3 Linearization and uniform observability of the minimal realization

Proposition 3 *Let* $\xi \doteq [\rho^1 \ {\mathbf{y}^1}^T, \ \dots, \ T^T, \ \Omega^T, \ V^T, \ \omega^T]^T$ *and* $F(t) \doteq \frac{\partial f(\xi)}{\partial \xi}$, $H(t) \doteq \frac{\partial \pi(\xi)}{\partial \xi}$ *denote the linearization of the state and measurement equation in (23) respectively, and let* $N \geq 5$*: Then the linearized system is uniformly observable for all* $0 < \rho^i < \infty, \ \forall \ i = 1 \dots N$.

Proof: *Let* $k = 2$*; that there exists an* $m_2 < \infty$ *such that* $M_2(t) \leq m_2 I$ *follows from the fact that* $F(t)$ *and* $H(t)$ *are bounded for all* t*, as can be easily verified. We now want to guarantee that* $M_2(t)$ *is strictly positive definite for all* t*. To this end, it is sufficient to show that the matrix*

$$U_2(t) \doteq \begin{bmatrix} H(t) \\ H(t+1)F(t) \end{bmatrix} \tag{24}$$

has full generic rank equal to $3N + 7$ *for all values of* t*, as can be verified using a symbolic manipulation package whenever* $N \geq 5$.

4 Estimation of structure and motion

Following the derivation in previous sections, the problem of estimating the motion, velocity and point-wise structure of the scene can be converted into the problem of estimating the state of the model (23). We propose to solve the task using a nonlinear filter, properly designed to account for the observability properties of the model. The implementation, which we report in detail in the companion paper [7], results in a sub-optimal

[11] Essentially we need to fix one particular element of the similarity group, i.e. a rotation, a translation and a scale. There are several ways of doing that. For instance, we can choose to fix the initial position and orientation of the camera relative to the object (pose), and the overall scaling of the points (scale), e.g. $T(0) = 0$, $R(0) = I$, $\|\mathbf{X}\| = 1$ or $\rho^i = 1$ for some $i \in 1 \dots N$, which corresponds to the object reference frame and the camera frame being the same at the initial time instant. Under this choice, the evolution of T and R becomes a deterministic integrator with no uncertainty, so the corresponding states can be removed from the model altogether.

However, velocity is not constant, while the coordinates of the points $\mathbf{X}$ are; it is therefore preferable for the sake of filtering to choose the base of the fiber by fixing quantities related to $\mathbf{X}$ rather than to T, R.

filter, as is well known [6]. However, it is important to guarantee that the estimation error, while different from zero, remains bounded. We do so, for the first time, in the next section 4.1. To streamline the notation, we represent the model (23) as

$$\begin{cases} \xi(t+1) = f(\xi(t)) + w(t) \quad w(t) \sim \mathcal{N}(0, \Sigma_{w_0}(t)), \\ y(t) = h(\xi(t)) + n(t) \quad n(t) \sim \mathcal{N}(0, \Sigma_{n_0}(t)) \end{cases} \tag{25}$$

The filter is described by a difference equation for the state $\hat{\xi}(t)$. We call the estimation error

$$\tilde{\xi}(t) \doteq \xi(t) - \hat{\xi}(t) \tag{26}$$

and its variance at time t $P(t)$. The initial conditions for the estimator are

$$\begin{cases} \hat{\xi}(0) = \xi_0 \\ P(0) = P_0 > 0 \end{cases} \tag{27}$$

and its evolution is governed by

$$\begin{cases} \hat{\xi}(t+1) = f(\hat{\xi}(t)) + K(t)[y(t+1) - h(\hat{\xi}(t))] \\ P(t+1) = \mathcal{R}\left(P(t), F(t), H(t), \Sigma_n, \Sigma_w\right) \end{cases} \tag{28}$$

where $\mathcal{R}$ denotes the usual Riccati equation which uses the linearization of the model $\{F, H\}$ computed at the current estimate of the state, as described in [16] (see also the companion paper [7]). K is the Kalman gain.

Note that we call Σ_{n_0}, Σ_{w_0} the variance of the measurement and model noises, and Σ_n, Σ_w the tuning parameters that appear in the Riccati equation. The latter are free for the designer to choose, as described in the companion paper [7].

4.1 Stability

The aim of this section is to prove that the estimation error generated by the filter just described is bounded. In order to do so, we need a few definitions.

Definition 4 *A stochastic process $\tilde{\xi}(t)$ is said to be* exponentially bounded in mean-square *(or MS-bounded) if there are real numbers η, $\nu > 0$ and $0 < \theta < 1$ such that $E\|\tilde{\xi}(t)\|^2 \leq \eta\|\tilde{\xi}(0)\|^2\theta^t + \nu$ for all $t \geq 0$. $\tilde{\xi}(t)$ is said do be* bounded with probability one *(or bounded WP1) if $P[\sup_{t\geq 0}\|\tilde{\xi}(t)\| < \infty] = 1$.*

Definition 5 *The filter (25) is said to be* stable *if there exist positive real numbers ϵ and δ such that $\|\tilde{\xi}(0)\| \leq \epsilon$, $\Sigma_n(t) \leq \delta I$, $\Sigma_w(t) \leq \delta I \implies \tilde{\xi}(t)$* is bounded. *Depending on whether $\xi(t)$ is bounded in mean square or with probability one, we say that the filter is* "MS-stable" *or* "stable WP1".

We are now ready to state the core proposition of this section

Proposition 4 *Let $0 < \rho^i < \infty \ \forall\ i = 1 \ldots N$ and $N \geq 5$ in the model (23). Then the filter based on such a model is MS-stable and stable WP1.*

First we need a result that follows directly from corollary 5.2 of [3]:

Lemma 1 *In the filter based on the model (23), let $P_0 > 0$. Then there exist positive real numbers p_1 and p_2 such that*

$$p_1 I \leq P(t) \leq p_2 I \quad \forall\ t \geq 0. \tag{29}$$

Proof: *The proof follows from corollary 5.2 of [3], using proposition 3 on the uniform observability of the linearization of (23).*

Proof of proposition 4: *The proposition follows immediately from theorem 3.1 in [27], making use in the assumptions of the boundedness of $F(t)$, $H(t)$, lemma 1 and the differentiability of f and g when $0 < \rho^i < \infty\ \forall\ i$.*

4.2 Instability of non-minimal models

Most recursive schemes for causally reconstructing structure and motion available in the literature represent structure using only one state per point (either its depth in an inertial frame, or its inverse, or other variations on the theme). This corresponds to reducing the state of the model (23), with the states $\mathbf{y}_0^i$ substituted for the measurements $\mathbf{y}^i(0)$, which causes the model noise $n(t)$ to be non-zero-mean[12]. When the zero-mean assumption implicit in the use of the Kalman filter is violated, the filter diverges[13]. In this case we say that the model is *sub-minimal*.

On the other hand, when the model is non-minimal – such is the case when we do not force it to evolve on a base of the state-space bundle – the filter is free to wonder on the non-observable space (i.e. along the fibers of the state-space bundle), therefore causing the explosion of the variance of the estimation error along the components of the state parallel to the fibers[14].

Therefore, the minimal realization (23) enforces no more and no less than the conditions that are required for designing a stable filter (hence the name *minimal*).

5 Conclusions

The causal estimation of three-dimensional structure and motion can be posed as a nonlinear filtering problem. In this paper we have analyzed it by providing a characterization of global observability, uniform observability, minimal realization and stability of the filter.

As described in a companion paper [7], the filter has been implemented on a Personal Computer, and the implementation has been made available to the public.

The next logical steps are in two directions. On one hand to explore more meaningful representations of the environment as a collection of surfaces with certain shape emitting a certain energy distribution. On the other hand, a theoretically sound treatment of nonlinear filtering for these problem involves estimation on Riemannian manifolds and homogeneous spaces. Both are open and challenging problems in need of meaningful solutions.

Acknowledgments

This research was supported by NSF grant IIS-9876145 and ARO grant DAAD19-99-1-0139. We wish to thank Hailin Jin, Paolo Favaro and Xiaolin Feng for their skillful implementation of the algorithm described in this paper in a real-time system. We also thank Carlo Tomasi, Pietro Perona, Giorgio Picci, Ruggero Frezza, John Hauser and John Oliensis for discussions.

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 7(4):348–401, 1985.

[2] D. Alspach and H. Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Trans. on Automatic Control*, 17:439–448, 1972.

[12] In fact, if we call $n^i(0)$ the error in measuring the position of the point i at time 0, we have that $E[n^i(t)] = n^i(0)\ \forall\ t$.

[13] In order to show this, it is sufficient to use a linear model

$$\begin{array}{rcl} x(t+1) & = & Ax(t) + w(t) \\ y(t) & = & Cx(t) + n(t) \end{array} \tag{30}$$

and assume that $n(t)$ is such that $E[n(t)] = m_n \neq 0\ \forall\ t$ and let the Kalman filter based on the above model be $\hat{x}(t+1) = A\hat{x}(t) + K(t)\left(y(t) - C\hat{x}(t)\right)$ where the gain $K(t)$ has been computed as if m_n was zero. If we define the estimation error $\tilde{x} \doteq x - \hat{x}$, it is immediate to see that its evolution is given by $\tilde{x}(t+1) = (A + K(t)C)\tilde{x}(t) + K(t)m_n$ and the mean of the estimation error $m_{\tilde{x}}(t) \doteq E[\tilde{x}(t)]$ obeys $m_{\tilde{x}}(t+1) = (A + K(t)C)m_{\tilde{x}}(t) + K(t)m_n$. If the model is observable, the matrix $A + K(t)C$ is stable (has eigenvalues inside the complex unit circle), while $K(t) \neq 0$. Since the mean of the noise m_n is constant and different from zero, it follows that the mean of the estimation error $m_{\tilde{x}} \to \infty$, and therefore the filter is unstable. Similar considerations apply for the extended version of the Kalman filter for nonlinear models.

[14] Following the notation of the previous footnote, even if $m_n = 0$, the matrix $A - K(t)C$ may have eigenvectors outside the unit circle, and therefore one or more components of $m_{\tilde{x}}(t)$ can diverge to infinity.

[3] B. Anderson and J. Moore. *Optimal filtering.* Prentice-Hall, 1979.

[4] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure and focal length. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(6):562–575, 1995.

[5] A. Blake and M. Isard. *Active contours.* Springer Verlag, 1998.

[6] T. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE Trans. Pattern Anal. Mach. Intell.*, Jan. 1986.

[7] A. Chiuso, P. Favaro, H. Jin, and S. soatto. 3-d motion and structure causally integrated over time, part 2: Implementation. Technical report, submitted to ECCV, November 1999.

[8] M. C. DeLara. Finite-dimensional filters: invariance group techniques. *SIAM J. on control and optimization*, 35(3):1002–1029, 1997.

[9] E. D. Dickmanns and V. Graefe. Applications of dynamic monocular machine vision. *Machine Vision and Applications*, 1:241–261, 1988.

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

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

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

[13] Guckenheimer and Holmes. *Nonlinear oscillations, dynamical systems and bifurcations of vector fields.* Springer Verlag, 1986.

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

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

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

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

[18] J. J. Koenderink and A. J. Van Doorn. Affine structure from motion. *J. Optic. Soc. Am.*, 8(2):377–385, 1991.

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

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

[21] S. Maybank. *Theory of reconstruction from image motion*, volume 28 of *Information Sciences.* Springer-Verlag, 1992.

[22] P. McLauchlan, I. Reid, and D. Murray. Recursive affine structure and motion from image sequences. *Proc. of the 3rd Eur. Conf. Comp. Vision*, Stockholm, May 1994.

[23] J. Oliensis. Provably correct algorithms for multi-frame structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1996.

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

[25] J. Philip. Estimation of three dimensional motion of rigid objects from noisy observations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(1):61–66, 1991.

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

[27] K. Reif, S. Gunther, E. Yaz, and R. Unbenhauen. Stochastic stability of the discrete-time extended kalman filter. *IEEE Trans. on Automatic Control*, 44(4):714–728, 1999.

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

[29] S. Soatto and P. Perona. Reducing "structure from motion": a general framework for dynamic vision. part 1: modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(9):993–942, September 1998.

[30] H. Sorenson and D. Alspach. Recursive bayesian estimation using gaussian sums. *Automatica*, 7:465–479, 1971.

[31] M. Spetsakis and J. Aloimonos. A multi-frame approach to visual motion perception. *Int. J. Computer Vision 6 (3)*, 1991.

[32] R. Szeliski. Recovering 3d shape and motion from image streams using nonlinear least squares. *J. visual communication and image representation*, 1994.

[33] M. A. Taalebinezhaad. Direct recovery of motion and shape in the general case by fixation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(8):847–853, 1992.

[34] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. of Computer Vision*, 9(2):137–154, 1992.

[35] J. Weng, N. Ahuja, and T. Huang. Optimal motion and structure estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15:864–884, 1993.

[36] Z. Zhang and O. D. Faugeras. Three dimensional motion computation and object segmentation in a long sequence of stereo frames. *Int. J. of Computer Vision*, 7(3):211–241, 1992.

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