Optimal Structure From Motion: Local Ambiguities and Global Estimates

Stefano Soatto† and Roger Brockett‡

† Washington University, One Brookings dr., St. Louis - MO 63130, soatto@ee.wustl.edu
and Dipartimento di Matematica ed Informatica, Università di Udine, Italy
‡ Harvard University, 29 Oxford st., Cambridge MA 02138, brockett@hrl.harvard.edu

Abstract

We present an analysis of SFM from the point of view of noise. This analysis results in an algorithm that is provably convergent and provably optimal with respect to a chosen norm. In particular, we cast SFM as a nonlinear optimization problem and define a bilinear projection iteration that converges to fixed points of a certain cost-function. We then show that such fixed points are "fundamental", i.e. intrinsic to the problem of SFM and not an artifact introduced by our algorithm. We classify and characterize geometrically local extrema, and we argue that they correspond to phenomena observed in visual psychophysics. Finally, we show under what conditions it is possible - given convergence to a local extremum - to "jump" to the valley containing the optimum; this leads us to suggest a representation of the scene which is invariant with respect to such local extrema.

1 Introduction

The problem of "Structure From Motion" (SFM) deals with extracting three-dimensional information about the environment from the motion of its projection onto a two-dimensional surface. We restrict our attention to a *point-wise* representation of the world. Despite being a rudimentary model, it allows us to touch upon some important issues in SFM that have been addressed only marginally in the past.

After 20 years of work in SFM, we can safely say that the *geometry* of the problem is fairly well understood (for the case of feature-points), and nicely summarized in the forthcoming book of Faugeras and Luong [5]. At the same time, the performance of many of the algorithms has been demonstrated on (more or less controlled) sequences of real images, and this has led many in the Computer Vision community to conclude that SFM has been solved. On the other hand, we are witnessing the frustration of many others, especially engineers, who are implementing existing algorithms for SFM in real-world situations, and observe that they behave in a way that is much different from their declared performance. This has led some to conclude that SFM is too difficult a problem to be solved in full generality even in the case of point features [9].

We believe that such a diversity of feelings comes from the fact that, while the geometry has been studied extensively, the issue of noise has been touched upon only in a superficial way (with some notable exceptions upon which we will comment later).

Relation to previous work

This paper relates to many previous works in SFM, and some of the relationships are pointed out throughout the paper. In particular, Weng et al. [18] test various general-purpose nonlinear optimization techniques on cost functions that include the epipolar constraint as well as the average 2-norm of the image-measurements. Cramèr-Rao bounds are evaluated, and an extensive set of simulation experiments compares existing linear algorithms against the optimal. In [18], iterative optimization schemes are initialized using a linear algorithm (such as a variation of the 8-point algorithm), and therefore the process can be viewed as the *optimal refinement* of existing SFM algorithms. Under conditions in which the linear algorithms do not give a satisfactory answer, Weng et al. only guarantee convergence to a local minimum, with no indication as to how this is related to the optimum.

Recently, Szeliski and Kang have also addressed the issue of optimal estimation of SFM and characterization of ambiguities [14, 15]. They use the Levemberg-Marquardt algorithm to find the extrema of the sum of reprojection errors. They perform an analysis of the bas-relief ambiguity based on the singularities of the Hessian of the information matrix. However, their results are essentially local, since they rely on a linearized model and assume small measurement errors.

The inherent ambiguities in SFM have been studied before also by Adiv [1], Young and Chellappa [20] (who also address the aperture problem), Oliensis [9] (who gives a provably convergent algorithm), and Spetsakis and Aloimonos [13] (who also propose an optimal algorithm). Our paper extends the above results: When other optimal algorithms converge to their global minimum, the solution is identical to the one we obtain. However, we achieve useful results for noises that are one order of magnitude higher than what commonly handled in the literature, without imposing restrictions on the initial conditions. We also classify and interpret geometrically local extrema, and we observe that they correspond to phenomena observed in visual psychophysics. We show how it is possible, under certain conditions, to obtain the global solution to SFM given convergence to a local extremum.

2 Spherical Least-Squares

Suppose we are given p unit-norm vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbf{S}^2$ and an unknown transformation $\mathbf{a} \in \mathbb{R}^3$ that acts on each $\mathbf{x}_i,\ i = 1 \ldots p$ via the cross-product $\mathbf{a} \times \mathbf{x}_i \in T_{\mathbf{x}_i}\mathbf{S}^2$

(the tangent plane to the unit-sphere $\mathbf{S}^2$ at $\mathbf{x}_i$). Suppose further that we can measure each transformed vector up to an unknown scaling factor $\lambda_i \in \mathbb{R}_+$: $\mathbf{y}_i = \mathbf{a} \times \mathbf{x}_i \lambda_i + \mathbf{n}_i \quad i = 1 \dots p$ where each $\mathbf{n}_i$ represents the uncertainty (or error) in the measurement $\mathbf{y}_i$, and hence $\mathbf{n}_i \in T_{\mathbf{x}_i}\mathbf{S}^2$. Now, given a number p of vectors $\mathbf{x}_i, i = 1 \dots p$ and the corresponding measurements $\mathbf{y}_i$, we want to find the transformation $\mathbf{a} \in \mathbb{R}^3$ and the scales $\lambda = [\lambda_1, \dots, \lambda_p]^T \in \mathbb{R}^p_+$ that minimize the norm of the uncertainty $\mathbf{n}$:

$$\min_{\mathbf{a}\in\mathbb{R}^3, \lambda\in\mathbb{R}^p_+} \sum_{i=1}^{p} \|\mathbf{n}_i\|^2_{w_i} \quad \text{subject to } \mathbf{y}_i = \mathbf{a} \times \mathbf{x}_i \lambda_i + \mathbf{n}_i \quad \in T_{\mathbf{x}_i}\mathbf{S}^2 \tag{1}$$

where w_i indicate the weights chosen for the components of the cost function. In the next three subsections we will analyze three distinct cases. Obviously, in the absence of noise ($\mathbf{n} = 0$) the three problems of minimizing the cost function above under the different choices of weights have the same exact solution.

In this and the following sections we use the "hat" operator to indicate a skew-symmetric matrix $\widehat{\mathbf{x}} \in so(3)$ that performs the cross product between two vectors $\mathbf{x}$ and $\mathbf{v}$: $\widehat{\mathbf{x}}\mathbf{v} = \mathbf{x} \times \mathbf{v}$. Since in all equations the parameters $\mathbf{a}$ and λ appear as a product, it is clear that they can only be recovered up to a common scale. Therefore, we choose to normalize $\mathbf{a}$, so that $\|\mathbf{a}\| = 1$, although any other choice for the normalization would do. Using this notation, we can formulate the "Spherical Least-Squares Problem" **(SLS)** as

$$\mathbf{a}_{opt}, \lambda_{opt} = \arg \min_{\mathbf{a}\in\mathbf{S}^2, \lambda\in\mathbb{R}^p_+} \sum_{i=1}^{p} \|\mathbf{y}_i + \widehat{\mathbf{x}}_i \mathbf{a} \lambda_i\|^2_{w_i}. \tag{2}$$

We refer to the sum above as the *cost function* of SLS.

2.1 Weighted Spherical Least Squares

In this section we consider $w_i(\mathbf{a}) \doteq \|\mathbf{a} \times \mathbf{x}_i\| \in [0, 1]$.

Claim 2.1 *Given $p > 3$ points $\mathbf{x}_i \in \mathbf{S}^2$, $i = 1 \dots p$ and the corresponding measurements $\mathbf{y}_i$, under general position conditions the SLS problem (2) with $w_i = w_i(\mathbf{a})$ admits a unique solution $\mathbf{a}_{opt}$ and λ_{opt} up to a sign. If we define M to be the symmetric 3×3 matrix $M = \sum_{i=1}^{p} \mathbf{y}_i {\mathbf{y}_i}^T$, then the optimal unit-norm solution $\mathbf{a}_{opt}$ is the eigenvector of M corresponding to its smallest eigenvalue: $\mathbf{a}_{opt} = \mathbf{v}_{min}(M)$ and the optimal scales λ_{opt} are obtained from $\mathbf{a}_{opt}$ via*

$$\lambda_{i\,opt} = -\frac{{\mathbf{y}_i}^T \widehat{\mathbf{x}}_i \mathbf{a}_{opt}}{\mathbf{a}_{opt}^T \widehat{\mathbf{x}}_i^2 \mathbf{a}_{opt}} \quad i = 1 \dots p. \tag{3}$$

Proof: *Consider the cost function of SLS, defined in equation (2). For any given $\mathbf{a}$, the λ that minimizes it is given, as a function of $\mathbf{a}$, by*

$$\lambda_i(\mathbf{a}) = -(\widehat{\mathbf{x}}_i \mathbf{a})^\dagger \mathbf{y}_i = -\frac{\mathbf{a}^T \widehat{\mathbf{x}}_i \mathbf{y}_i}{\mathbf{a}^T {\widehat{\mathbf{x}}_i}^2 \mathbf{a}} \quad i = 1 \dots p. \tag{4}$$

Once we substitute $\lambda(\mathbf{a})$ back into (2), the components of the cost function become $\|(\widehat{\mathbf{x}}_i \mathbf{a})^\perp \mathbf{y}_i\|$. We now use the fact that $\mathbf{x}^\perp = -\widehat{\mathbf{x}}^2$ to replace $(\widehat{\mathbf{x}}_i \mathbf{a})^\perp \mathbf{y}_i$ with $\frac{(\widehat{\mathbf{x}_i}\mathbf{a})}{\|\widehat{\mathbf{x}_i}\mathbf{a}\|} \times \mathbf{y}_i$, ending up with minimizing for $\mathbf{a}$ the cost function $\sum_i \frac{\|(\widehat{\mathbf{x}_i}\mathbf{a}) \times \mathbf{y}_i\|^2_{w_i}}{\|\widehat{\mathbf{x}_i}\mathbf{a}\|^2}$. Since $(\widehat{\mathbf{x}}_i \mathbf{a})\times = \mathbf{a}{\mathbf{x}_i}^T - \mathbf{x}_i \mathbf{a}^T$, and $\mathbf{y}_i$ is orthogonal to $\mathbf{x}_i$, we can further simplify the SLS problem, that becomes

$$\arg \min_{\mathbf{a}\in\mathbf{S}^2} \sum_{i=1}^{p} \|\mathbf{x}_i \mathbf{a}^T \mathbf{y}_i\|^2 = \mathbf{a}^T \left(\sum_{i=1}^{p} \mathbf{y}_i {\mathbf{y}_i}^T \right) \mathbf{a}. \tag{5}$$

It is immediate to see that the least-squares unit-norm solution for $\mathbf{a}$ is given by the eigenvector of M corresponding to the smallest eigenvalue. Note that, since M is symmetric, the eigenvalues are real and positive, and the eigenvectors are unit-norm orthogonal vectors. Once the optimal $\mathbf{a}$ has been computed, the corresponding optimal λ can be obtained as $\lambda_i(\mathbf{a})$ from equation (4). Note that there is a sign ambiguity in $\mathbf{a}$, that reflects onto the sign of the vector λ.

2.2 Balanced Spherical Least Squares

Let us assume for a moment that the weights w_i are all identical to 1. We can still follow the procedure outlined in the proof of claim 2.1, but rather than ending up with minimizing the cost function $\sum_{i=1}^{p} \|\mathbf{x}_i \mathbf{a}^T \mathbf{y}_i\|^2$ in (5), we have $\sum_{i=1}^{p} \frac{\|\mathbf{x}_i \mathbf{a}^T \mathbf{y}_i\|^2}{\|\mathbf{x}_i \times \mathbf{a}\|^2}$, which can be re-written as

$$\sum_{i=1}^{p} \frac{< \mathbf{y}_i, \mathbf{a} >^2}{\|\mathbf{x}_i \times \mathbf{a}\|^2}. \tag{6}$$

If we assume that $\mathbf{x}_i$ span a small solid angle (compared to the full sphere), then it makes sense to talk about an *average* direction $\bar{\mathbf{x}}$. If we weight each point with $w_i = \frac{\|\mathbf{x}_i \times \mathbf{a}\|}{\|\bar{\mathbf{x}} \times \mathbf{a}\|}$, the cost function to be minimized becomes $\frac{\mathbf{a}^T M \mathbf{a}}{\mathbf{a}^T N \mathbf{a}}$, where $M \doteq \sum_{i=1}^{p} \mathbf{y}_i \mathbf{y}_i^T$ and $N \doteq \widehat{\bar{\mathbf{x}}}^2$. The solution for this problem has to do with Singular Rayleigh quotients and is derived in [11]. We summarize the result in the following

Claim 2.2 *The solution $\mathbf{a}_{opt}$ for the problem of minimizing the cost function (6) is obtained by minimizing the corresponding Singular Rayleigh Quotient $\frac{\mathbf{a}^T M \mathbf{a}}{\mathbf{a}^T N \mathbf{a}}$. The solution is given by the eigenvector of the matrix $M_s \doteq M - \frac{M \bar{\mathbf{x}} \bar{\mathbf{x}}^T M}{\bar{\mathbf{x}}^T M \bar{\mathbf{x}}}$ relative to the matrix N corresponding to the smallest non-zero eigenvalue.*

2.3 Unweighted Spherical Least Squares

When we choose all weights w_i to be identically equal to one, the problem of Spherical Least Squares can be reduced, following the proof of claim 2.1, to minimizing (6) subject to the constraints $< \mathbf{y}_i, \mathbf{x}_i >= 0$ and $\|\mathbf{x}_i\| = 1$. In the presence of noise, we have not found a closed-form optimal solution for the Unweighted Spherical Least-Squares problem. However, the following simple iteration can be easily proven to be contractive and therefore to converge to a fixed point:

$$\mathbf{a}_{k+1} = \mathrm{pr}_{\mathbf{S}^2} \left(I - H^{-1} D^T \right) \mathbf{a}_k \tag{7}$$

where $D = \sum \frac{\mathbf{y}\mathbf{y}^T (\mathbf{a}^T \widehat{\mathbf{x}}^2 \mathbf{a}) - \widehat{\mathbf{x}}^2 (\mathbf{a}^T \mathbf{y}\mathbf{y}^T \mathbf{a})}{(\mathbf{a}^T \widehat{\mathbf{x}}^2 \mathbf{a})^2}$, H is $\sum \frac{\mathbf{y}\mathbf{y}^T - 2\mathbf{a}\mathbf{a}^T \widehat{\mathbf{x}}^2 + \mathbf{a}^T \mathbf{y}\mathbf{y}^T \mathbf{a} \widehat{\mathbf{x}}^2 - 2\mathbf{a}\mathbf{a}^T \mathbf{y}\mathbf{y}^T}{(\mathbf{a}^T \widehat{\mathbf{x}}^2 \mathbf{a})^2} - 4 \frac{-\mathbf{y}\mathbf{y}^T \mathbf{a}\mathbf{a}^T \widehat{\mathbf{x}}^2 - \widehat{\mathbf{x}}^2 \mathbf{a}\mathbf{a}^T \mathbf{y}\mathbf{y}^T \mathbf{a}\mathbf{a}^T \widehat{\mathbf{x}}^2}{(\mathbf{a}^T \widehat{\mathbf{x}}^2 \mathbf{a})^3}$, and $\mathrm{pr}_{\mathbf{S}^2}$ denotes renormalization (projection onto the sphere). Of course this iteration is only guaranteed to converge to a local solution of the Unweighted SLS. However, we have noticed that using a Weighted SLS or a Balanced SLS as an initialization step usually places the iteration in the basin of attraction of the global minimum. In particular, for small levels of noise (up to 50% of the average signal), we have observed that both the Weighted SLS and

the Balanced SLS approximate well the solution of the unweighted SLS, and therefore running an iteration of the type above improves only marginally the solution. For higher noise levels we have observed that when the deviation of the $\mathbf{x}_i$ is large, the solution to the Weighted SLS provides an accurate initialization, while when deviation of the $\mathbf{x}_i$ is small (on the order of 40^o or less), the solution of a Balanced SLS is more accurate. In both cases, however, the iteration converges in a few steps (less than 10).

3 Structure From Motion

Consider a variation of the Spherical Least-Squares problem of equation (2), where we add to the cost function the affine term $\widehat{\mathbf{x}}_i^2\mathbf{b}$, with $\mathbf{b} \in \mathbb{R}^3$ unknown: $\min_{\mathbf{a}\in\mathbf{S}^2,\mathbf{b}\in\mathbb{R}^3\lambda\in\mathbb{R}_+^p} r_0(\mathbf{a},\mathbf{b},\lambda)$ where we define

$$r_0(\mathbf{a},\mathbf{b},\lambda) = \sum_{i=1}^{p} \|\mathbf{y}_i + \widehat{\mathbf{x}}_i\mathbf{a}\lambda_i - \widehat{\mathbf{x}}_i^2\mathbf{b}\|^2 \tag{8}$$

as the *cost function* of SFM, and we neglect the subscripts w_i that indicate the choice of weights. This problem can be interpreted as that of estimating the direction of translation $\mathbf{a} \in \mathbf{S}^2$, the rotational velocity $\mathbf{b} \in \mathbb{R}^3$ and the depth $1/\lambda_i$ $i = 1 \dots p$ of a number p of moving points in 3-D, from the noisy projection of their velocity onto the retina, modeled as a unit-sphere [7]. The estimation criterion is to minimize the (weighted) norm of the "reprojection error".

In the special case $\mathbf{b} = 0$ (no rotation), (8) reduces to a SLS problem. Depending upon the weights chosen, we have shown in section 2 how to obtain a (local) solution for the direction of translation $\mathbf{a}$ and the scaling parameters (inverse depths) λ_i.

In the presence of rotation $\mathbf{b} \neq 0$, the problem defined by equation (8) is no longer a standard SLS problem. Many in the Computer Vision literature have proposed methods to "undo" the rotation, by either warping the image [12, 10, 3], by applying a linear transformation to the measured data [16, 7], or by estimating a rotational velocity assuming small translation [9]. In all cases, however, the transformation acts on the noise as well as on the data, therefore spoiling the goal of achieving an optimal estimate. Here we take a different approach, that is somewhat more aware of noise and results in an optimal (although non-linear) estimator.

Following the derivation of the solution of SLS in claim 2.1, we can transform the problem of SFM into

$$\arg \min_{\mathbf{a}\in\mathbf{S}^2,\mathbf{b}\in\mathbb{R}^3} \sum_{i=1}^{p} \|\mathbf{x}_i\mathbf{a}^T(\mathbf{y}_i - \widehat{\mathbf{x}}_i^2\mathbf{b})\|^2 \tag{9}$$

Now, for any given $\mathbf{b}$, the $\mathbf{a}$ that minimizes the norm of the cost function of SFM in (8) can be obtained by solving the SLS problem (2) with $\mathbf{y}_i$ being substituted by $\tilde{\mathbf{y}}_i = \mathbf{y}_i - \widehat{\mathbf{x}}_i^2\mathbf{b}$. In the same fashion – given $\mathbf{a}$ – the $\mathbf{b}$ that minimizes the norm of the $\mathbf{n}$ is obtained immediately from (9) as

$$\mathbf{b}(\mathbf{a}) = \left(\sum_{i=1}^{p} \widehat{\mathbf{x}}_i^2\mathbf{a}\mathbf{a}^T\widehat{\mathbf{x}}_i^2\right)^{-1} \sum_{i=1}^{p} \widehat{\mathbf{x}}_i^2\mathbf{a}\mathbf{a}^T\mathbf{y}_i. \tag{10}$$

Therefore, the "conditional" problems of estimating $\mathbf{a}$ given $\mathbf{b}$ – or $\mathbf{b}$ given $\mathbf{a}$ – are particularly simple. Based on the simplicity of the conditional problems, one may be tempted to try the following **Bilinear Projection Algorithm (BPA)**:

- let $k = 0$ and choose any initial value for $\mathbf{b} \in \mathbb{R}^3$
- iterate the following:
  - $\mathbf{a}_k = \arg\min_{\mathbf{a}\in\mathbf{S}^2} \sum_{i=1}^{p} \|\mathbf{x}_i\mathbf{a}^T(\mathbf{y}_i - \widehat{\mathbf{x}}_i^2\mathbf{b}_k)\|^2$
  - $\mathbf{b}_{k+1} = \left(\sum_{i=1}^{p} \widehat{\mathbf{x}}_i^2\mathbf{a}_k\mathbf{a}_k^T\widehat{\mathbf{x}}_i^2\right)^{-1} \sum_{i=1}^{p} \widehat{\mathbf{x}}_i^2\mathbf{a}_k\mathbf{a}_k^T\mathbf{y}_i.$
  - $k = k + 1$

We are implicitly excluding the case $\mathbf{a} = 0$, for that case can be easily detected and treated separately (see [11]). It is straightforward to prove that this iteration converges "somewhere". It is less obvious to make sure that the iteration converges to some meaningful local extrema. Luckily we have the following

Claim 3.1 *Given $p > 5$ points in general position, the Bilinear Projection Algorithm converges to a local extremum of the bilinear cost function* **(BCF)** $r(\mathbf{a},\mathbf{b}) \doteq \sum_{i=1}^{p} \|\mathbf{x}_i\mathbf{a}^T(\mathbf{y}_i - \widehat{\mathbf{x}}_i^2\mathbf{b})\|^2$. *Extrema of the bilinear cost function r are in one-to-one correspondence with those of the cost function of SFM in (8).*

In order to prove the claim we need to establish that the bilinear iteration does not introduce "phantom" stationary points to the cost function. This is guaranteed by the following:

Lemma 3.1 *Let $\psi(\mathbf{a})$ be the p-dimensional vector with i-th component $\mathbf{x}_i\mathbf{a}^T\mathbf{y}_i$, and ϕ the $p \times 3$ matrix with i-th row equal to $\mathbf{x}_i\mathbf{a}^T\widehat{\mathbf{x}}_i^2$. Then $r(\mathbf{a},\mathbf{b}) = \|\psi(\mathbf{a}) - \phi(\mathbf{a})\mathbf{b}\|$; define $r_2(\mathbf{a}) \doteq \|\phi(\mathbf{a})^\perp\psi(\mathbf{a})\|$ [1]. Furthermore, assume that ϕ has constant rank $\rho = 3$ in some open subset Ω of $\mathbb{R}^p$. If $\mathbf{a}^*$ is a critical point (or a global minimizer) of r_2, and $\mathbf{b}^* \doteq \phi^\dagger(\mathbf{a}^*)\psi(\mathbf{a}^*)$, then $(\mathbf{a}^*,\mathbf{b}^*)$ is a critical point (or a global minimizer) of r and $r_2(\mathbf{a}^*) = r(\mathbf{a}^*,\mathbf{b}^*)$. If $(\mathbf{a}^*,\mathbf{b}^*)$ is a global minimizer of r, then $\mathbf{a}^*$ is a global minimizer of r_2 and $r_2(\mathbf{a}^*) = r(\mathbf{a}^*,\mathbf{b}^*)$. † denotes the (least-squares) pseudo-inverse.*

Proof: *This is a a special case of theorem 2.1 on page 416 of [6], with the simple extension of allowing the affine term ψ to depend on $\mathbf{a}$.*

Proof of claim 3.1: *The Bilinear Projection Algorithm is a Gauss-Newton iteration for the cost function $r(\mathbf{a},\mathbf{b})$. The lemma guarantees that the iteration performed by alternating the variables $\mathbf{a}$ and $\mathbf{b}$ has the same fixed points of the iteration performed simultaneously on $\mathbf{a}$ and $\mathbf{b}$. The first part of the claim follows from standard properties of Gauss-Newton iterations. That such extrema correspond to those of the original cost function in (8) follows by applying the lemma again to the SLS problem.*

The above claim guarantees that, following the BPA outlined in this paragraph, we do not introduce spurious solutions to the problem of SFM. Therefore, it remains to be established whether the *original problem of SFM*, as formulated in (8), *admits spurious local solutions* in the presence of noise $\mathbf{n}$. This is the subject of the next section.

[1] Here the notation $\phi^\perp$ stands for the projector operator onto the orthogonal complement of the range space of ϕ, defined as $\phi^\perp = I - \phi\phi^\dagger$, where † denotes the pseudo-inverse.

4 "Bas-relief ambiguity", "rubbery motion percept" and a robust representation of shape

In order to detect and classify the local extrema associated with the cost function of SFM, one can start by setting up a random sampling simulation. Note that, by virtue of lemma 3.1, it is equivalent to check any of the cost functions $r_0(\mathbf{a}, \mathbf{b}, \lambda)$, $r(\mathbf{a}, \mathbf{b})$ or $r_2(\mathbf{a})$, defined in (8) and in lemma 3.1 respectively. This makes the random search particularly favorable for $r_2(\mathbf{a})$, since it only depends upon 2 parameters, and can therefore be visualized, as we will see in section 5. We first observe that local extrema tend to cluster in a small number of groups (eight), and then give analytical explanation for the geometric/noise configurations that give raise to such local minima.

The "bas-relief" ambiguity

One of the common complaints to SFM algorithms from perspective is that they become unreliable in the presence of small fields of view, and when the rotational component of motion is "confused" with the translational component (see for instance [18]). It is our goal in this section to formalize this concept and establish that this effect is "intrinsic", and not an artifact of the algorithm. We will also discuss what information can be robustly recovered under these conditions.

When an object occupies a small portion of the visual field, $\mathbf{x}_i$ tend to be similar. We call their average direction $\bar{\mathbf{x}}$. Rotation about an axis orthogonal to the line of sight passing through the object corresponds to $\mathbf{a}$ and $\mathbf{b}$ being orthogonal to each other and both orthogonal to $\bar{\mathbf{x}}$. We will now show that these conditions, in presence of noise, give rise to a minimum of the cost function (8).

Claim 4.1 *Let* $\mathbf{y}_i = -\widehat{\mathbf{x}}_i\mathbf{a}\lambda_i + \widehat{\mathbf{x}}_i^2\mathbf{b} + \mathbf{n}_i$, *and* $\bar{\mathbf{x}} \perp \mathbf{a} \perp \mathbf{b}$. *Furthermore, let* $\mathbf{b} = \mathbf{b}_0 + \delta\mathbf{b}$, *with* $\|\delta\mathbf{b}\| \cong \|\mathbf{n}\|$, *the average norm of the measurement error. Then the cost function* $r_0(\mathbf{a}, \mathbf{b}, \lambda) \doteq \sum_{i=1}^{p} \|\mathbf{y}_i + \widehat{\mathbf{x}}_i\mathbf{a}\lambda_i - \widehat{\mathbf{x}}_i^2\mathbf{b}\|^2$ *has a local extremum at* $\tilde{\mathbf{b}} = \mathbf{b}_0$, $\tilde{\lambda} = \lambda - \bar{\lambda}$, *and* $\tilde{\mathbf{a}} = \mathbf{v}_{min}(\sum_{i=1}^{p} \mathbf{y}_i\mathbf{y}_i{}^T)$, *where the bar denotes the average, and* $\mathbf{v}_{min}$ *denotes the eigenvector corresponding to the smallest eigenvalue.*

Proof *Without loss of generality, assume* $\mathbf{b}_0 = 0$, *so that* $\mathbf{b} = \delta\mathbf{b}$ *is of size comparable to the noise:* $\|\mathbf{b}\| \cong \|\mathbf{n}\|$. *Since* $\bar{\mathbf{x}} \perp \mathbf{b}$ *and* $\mathbf{x}_i \in \mathbf{S}^2$, *we have* $\|\widehat{\mathbf{x}}_i^2\mathbf{b}\| \cong \|\mathbf{b}\| \quad \forall i = 1 \ldots p$ *and therefore the terms* $\widehat{\mathbf{x}}_i^2\mathbf{b}$ *are comparable with the noise* $\mathbf{n}_i$, *and they can be lumped as a bias into* $\tilde{\mathbf{n}}_i = \widehat{\mathbf{x}}_i^2\mathbf{b} + \mathbf{n}_i$. *The value of* $\mathbf{b}$ *that minimizes the norm of* $\tilde{\mathbf{n}}$ *is* $\tilde{\mathbf{b}} = 0$, *and the corresponding* $\tilde{\mathbf{a}}$ *is obtained as the solution to the SLS problem (2), e.g.* $\tilde{\mathbf{a}} = \mathbf{v}_{min}(\sum_{i=1}^{p} \mathbf{y}_i\mathbf{y}_i{}^T)$. *In order to evaluate the corresponding* $\tilde{\lambda}$, *we observe that* $\lambda_i = -(\widehat{\mathbf{x}}_i\mathbf{a})^\dagger\mathbf{y}_i + (\widehat{\mathbf{x}}_i\mathbf{a})^\dagger\widehat{\mathbf{x}}^2\mathbf{b} \cong \lambda_i^0 + (\widehat{\mathbf{x}}\mathbf{a})^\dagger\widehat{\mathbf{x}}^2\mathbf{b} = \lambda_i^0 + \bar{\lambda}$ *where* λ_i^0 *are obtained assuming* $\mathbf{b} = 0$ *and* $\bar{\lambda}$ *is the average of the scales* λ_i. *Therefore, the scales corresponding to the local solution* $\mathbf{b} = 0$ *are the zero-mean version of the original ones* $\tilde{\lambda}_i \cong \lambda_i - \bar{\lambda}$.

Remark 4.1 *From the claim we can conclude that, in presence of high noise levels, a portion of the rotational velocity* $\mathbf{b}$ *can be confused with noise and compensated by a "bias" in the translational velocity* $\mathbf{a}$, *and the corresponding inverse depths* λ_i *are offset towards the origin. This statement confirms the observations of Weng et al. [18], although they attribute the effect to the geometry of the epipolar constraint. The consequence of claim 4.1 seems in contrast with the observations of [17], that the axis of rotation has no impact on the bias of the estimate of translation. However, in order for the bas-relief effect to show, the conditions of claim 4.1 must be met, in particular the aperture angle must be small, the noise level must be high and the algorithm must be initialized far away from the true solution. Some of these conditions were not explored in the experimental setup of [17], and therefore the effect was not observed.*

Other extrema

In the proof of claim 4.1 we have computed the local minimum $\tilde{\mathbf{a}}$ associated with $\tilde{\mathbf{b}} = 0$ as the solution of the corresponding weighted SLS problem. Such a solution is the eigenvector of the matrix M (defined in claim 2.1) corresponding to its smallest eigenvalue. In the absence of noise, such eigenvalue is 0. For small noise levels, it still is distinctively smaller than the remaining two. However, in the presence of large noise, the eigenvalues become comparable and therefore the actual solution of the SLS problem can be any of the three eigenvectors of M. This indeed happens – for large noise levels – and it accounts for three of the extrema of the cost function found experimentally.

Detecting local minima and switching between extrema

We have established that there are (at least) two extrema of the bilinear cost function (BCF) for $\mathbf{b}$: one corresponding to the true solution, and one corresponding to the bas-relief ambiguity $\tilde{\mathbf{b}} = 0$. Correspondingly, there are three extrema for $\mathbf{a}$, the eigenvectors of the matrix M defined in claim 2.1. Out of these three extrema, there is a minimum, a saddle, and a maximum, as a consequence of the analysis in section 2. It is possible to detect whether a stationary point $\tilde{\mathbf{a}}$ is a local minimum or not. In fact, given $\tilde{\mathbf{a}}$, we can compute its orthogonal complement (the remaining two eigenvectors of M), and compute the corresponding residual. We then just choose the eigenvector that carries the smallest residual. By doing so, we rule out 4 local extrema (2+2), and we are left with 2 possibilities: $\tilde{\mathbf{b}} = \mathbf{b}$, or $\tilde{\mathbf{b}} \cong 0$, the bas-relief ambiguity.

As it turns out, this situation can also be detected easily, even without knowing the noise level $\|\mathbf{n}\|$ (which gives a lower bound on the residual). In fact, the correct $\mathbf{b}$ leads to reconstructed scales λ that are positive, while in the bas-relief ambiguity they are offset towards the origin and, as long as not all λ_i are equal (i.e. when the structure is a perfect fronto-parallel plane), some $\tilde{\lambda}_i$ will be negative, as a consequence of claim 4.1. Note that even in the latter case, which corresponds to the bas-relief ambiguity, it is still possible to retrieve a useful representation of shape, for $\tilde{\lambda}$ can be re-scaled by choosing ρ so that $\tilde{\lambda} + \rho > 0$, which leads to the *bas-relief*. These scaled parameters can be chosen to effectively re-initialize the algorithm, as we will see in the experimental section.

"Rubbery motion" percept

An interesting phenomenon that has been observed in psychophysical experiments occurs under the same conditions of the bas-relief ambiguity. However, instead of rotational velocity being underestimated, it is perceived as being the opposite of the true one.

In order to analyze this phenomenon from the point of view of noise, let us consider the same conditions of the bas-relief ambiguity as expressed in claim 4.1, and assume that $\tilde{\mathbf{b}} = -\mathbf{b}$, and $\tilde{\lambda}_i = -\lambda_i^0 - \bar{\lambda}$. In order for this to be a legitimate local extremum, as a consequence of lemma 3.1, there must exist some $\tilde{\mathbf{a}}$ that makes the noise $\tilde{\mathbf{n}}_i$ small, where $\mathbf{y}_i = -\widehat{\mathbf{x}}_i\tilde{\mathbf{a}}\tilde{\lambda}_i + \widehat{\mathbf{x}}_i^2\tilde{\mathbf{b}} + \tilde{\mathbf{n}}_i$. If we substitute the expressions for $\tilde{\lambda}_i$ and $\tilde{\mathbf{b}}$ we get $\mathbf{y}_i \cong \widehat{\mathbf{x}}_i\tilde{\mathbf{a}}\lambda_i^0 - (\widehat{\bar{\mathbf{a}}\bar{\mathbf{x}}})(\widehat{\bar{\mathbf{a}}\bar{\mathbf{x}}})^\dagger\widehat{\bar{\mathbf{x}}}^2\mathbf{b} - \widehat{\bar{\mathbf{x}}}^2\mathbf{b} + \tilde{\mathbf{n}}_i$. From that expression it is possible to see that, under the assumptions of the bas-relief ambiguity, $\tilde{\mathbf{a}} = -\mathbf{a}_0$, where $\mathbf{a}_0$ is the solution to the SLS problem obtained by assuming $\mathbf{b} = 0$. In fact, in that case we have $\mathbf{y}_i \cong \widehat{\mathbf{x}}_i\tilde{\mathbf{a}}\lambda_i^0 - (\widehat{\bar{\mathbf{a}}\bar{\mathbf{x}}})^\perp\widehat{\bar{\mathbf{x}}}^2\mathbf{b} + \tilde{\mathbf{n}}_i$ but $\widehat{\bar{\mathbf{x}}}^2\mathbf{b} \cong \mathbf{b}$ and $\widehat{\bar{\mathbf{a}}\bar{\mathbf{x}}} \cong \frac{\mathbf{b}}{\|\mathbf{b}\|}$, so that $(\widehat{\bar{\mathbf{a}}\bar{\mathbf{x}}}) \times (\widehat{\bar{\mathbf{x}}}^2\mathbf{b}) = 0$ and the second term in the previous expression is negligible. We have therefore

$$\mathbf{y}_i \cong \widehat{\mathbf{x}}_i\tilde{\mathbf{a}}\lambda_i^0 + \tilde{\mathbf{n}}_i. \tag{11}$$

The value of $\tilde{\mathbf{a}}$ that minimizes the sum of the norms of $\tilde{\mathbf{n}}_i$ is obtained as the solution of the SLS problem associated to (11).

Remark 4.2 *Note that this solution, which we call the "rubbery motion" effect, is* not *just the correct solution with the flipped sign, for that would correspond to* $\tilde{\lambda}_i = -\lambda_i$*, while here we have* $\tilde{\lambda}_i = \lambda_i - 2\lambda_i^0$.

A robust representation of shape

The averaged inverse depth is invariant under the bas-relief ambiguity. The rubbery motion percept consists in a sign change of the averaged inverse depth. Therefore, the averaged inverse depth represents shape up to a global scaling factor and a sign, and it is invariant under the bas-relief ambiguity and the rubbery motion percept.

5 Experiments

How much noise is too much noise?

Typical feature-tracking/optical-flow algorithms declare accuracy in locating corresponding feature-points in the order of 0.1 pixels std [2]. It is our experience that this is indeed the case for about 30% of the feature-points extracted automatically according to a SSD (Sum of Squared Differences) criterion. However, for 70% of the features, a more realistic figure for the localization error is 1 pixel std. Now consider a camera with a 30^o field of view and an imaging sensor of 512×512 pixels, translating forward at $0.5m/s$. Depending upon the scene being viewed, the average norm of the flow vectors on the imaging sensor is in the order of 1-2 pixels (for a 15 frames/second capture rate). Therefore, an error in the order of 1 pixel corresponds to 50%-100% of the measurements. Consider again the camera just described, but now looking at an object that is $2m$ ahead of the camera and rotating about an axis passing through its centroid at $1^o/s$. In this case the average norm of flow vectors is 0.1 pixels/frame, and 1 pixel error corresponds to an intolerable 1000%.

Therefore, even the use of the most accurate feature-tracker does not dispense us from dealing with noise in scenarios that are very often encountered in real-world situations.

Sample tests on real images

Since the algorithm described in section 3 is optimal by construction, we should expect it to work at least as well as any other algorithm. Most real image sequences available do not have a reliable ground-truth, and therefore a fair comparison is impossible. Later in this section, however, we report the results of a *simulation* to compare the optimal algorithm versus ones based upon epipolar geometry and linear subspace constraints.

Here we just report the use of the algorithm on a real image sequence for the sake of example. We have chosen as a sample experiment a "box-sequence" that was available on the Web in Matlab format with calibration data (figure 1); the motion pattern is the one leading to the bas-relief ambiguity. The box rotates about a vertical axis at a rate of $3^o/frame$, which is a fairly large motion. Under these conditions even the 8-point algorithm of Longuet-Higgins works.

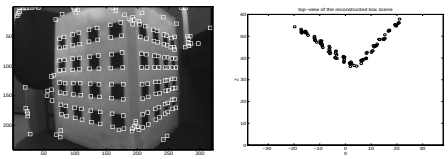

Figure 1: **Box scene** *A box rotates about a vertical axis at* $3^o/frame$*. The top-view of the reconstructed scene is shown on the right in normalized units (units of translational velocity). No reliable ground-truth is available.*

Sample of convergence behavior during simulations

In figure 2 we show a typical case where the iteration converges to the global minimum. The residual decreases up to the level of the noise (left), and both the parameters $\mathbf{a}, \mathbf{b}$ (center) and the scales λ (right) are within bounds from the true values.

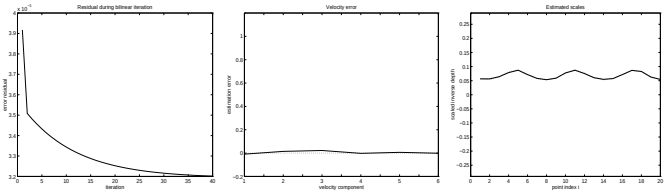

Figure 2: **Convergence to the global minimum** *(left) residual cost function, (center) parameter estimation error, (right) estimated scales. Ground truth is in dotted lines (if you can see them, that is).*

However, sometimes the residual stabilizes at a level different from the noise level, as in figure 3 (top-row left), but both the parameters (center) and the scales (right) are quite far from their true values, indicated in dotted lines. This is a clear sign

that the algorithm has converged to a local extremum. However, if we check all three eigenvectors of the matrix M and the scales λ that they generate (3 middle-row), we see that one of them (center) produces an estimate that corresponds to the averaged version of the correct scales. Therefore, there has been a switch of the eigenvalues of M. We can now switch to the solution for $\mathbf{a}$ and λ corresponding to the eigenvector that generates the smallest residual, and use that as initial condition for a second run of the algorithm, that converges in 5 iterations (3 bottom-row). Figure 4 shows convergence to the bas-relief ambiguity.

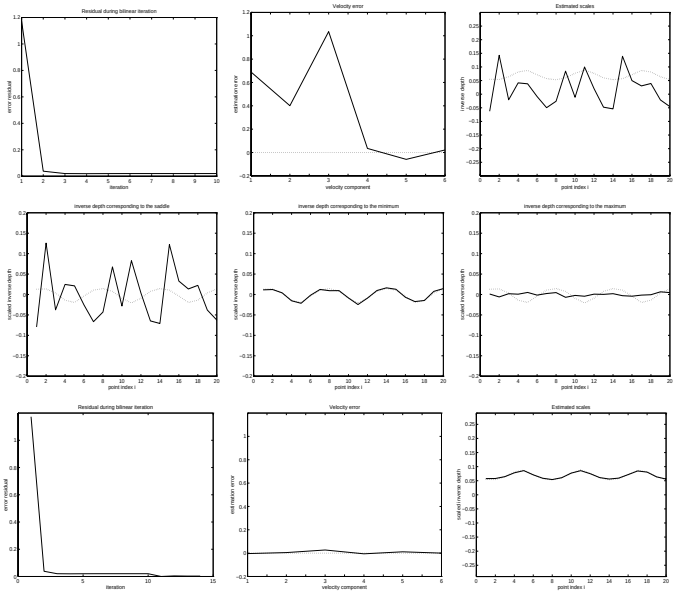

Figure 3: **Convergence to a local extremum** *(top) the residual stabilizes (left), but the parameter error (center) and scales (right) are far away from their true values (in dotted lines). The normalized scales corresponding to the 3 eigenvectors of M, plotted in the middle row, show that one of them corresponds to the correct estimate. We may then switch to the correct solution and re-initialize the algorithm, that converges to the correct solution within 5 steps (bottom row).*

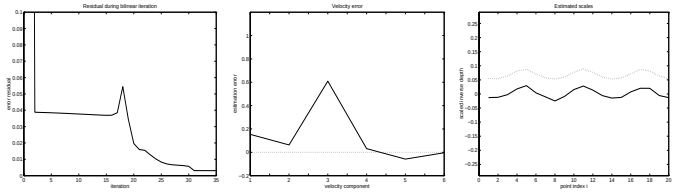

Figure 4: **Convergence to the bas-relief ambiguity** *The residual stabilizes (left), the parameters (center) are far away from the true values, but the estimated scales (right) are an averaged version of the true ones.*

Predicted behavior based on the analysis

Based upon the analysis carried out in section 3, we know that local extrema of the original function of SFM r in (8) are in correspondence with the local extrema of the reduced function $r_2(\mathbf{a})$. Since $\|\mathbf{a}\| = 1$, we can represent $\mathbf{a}$ in spherical coordinates and plot the cost function r_2 (fig. 5). The motion is a fixating one similar to that of the box experiment (figure 1). In addition to the global minimum, corresponding to the coordinates $(0, \pi/2)$ (azimuth, elevation), we expect a maximum and a saddle in the orthogonal direction. These are showed in figure 5 (top left). The saddle coincides with the singularity of the spherical coordinates: in fact, the two lines $(-\pi/2, \alpha)$, and $(\pi/2, \alpha)$ correspond to a point on the sphere. The rubbery interpretation corresponds to the local minimum diametrically opposed to the true motion (figure 5 top right). the local extrema corresponding to the bas-relief ambiguity are reported in figure 5 (bottom left), while in (bottom right) we show the location of the extrema corresponding to the "rubbery" bas-relief ambiguity. We expect that our simulations will show convergence to some or all of these local extrema.

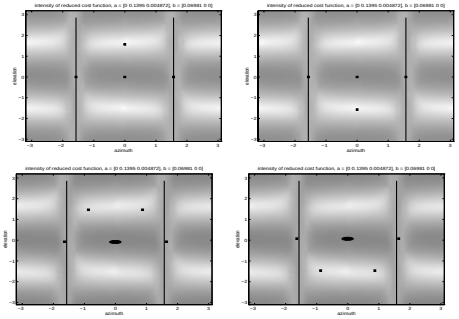

Figure 5: **Predicted extrema for fixating motion** *Local extrema are plotted as black asterisks superimposed to the residual of the cost function r_2 (top left). We also show the local extrema corresponding to the rubbery interpretation (top right), the bas-relief ambiguity (bottom left), and the rubbery bas-relief ambiguity (bottom right).*

Experimental trials

We have considered $p = 20$ points on a volume of side $1m$ centered $2m$ from the center of projection. The points have images on the unit sphere. First we have considered forward translation at $0.2m/frame$, and noise levels of 0.1, 1 and 10 pixels std, corresponding to 4%, 40% and 400% of the measurements respectively. For each noise level we have performed 200 trials. In figure 7 we show the plot of the residual of the cost function superimposed to the point where the Bilinear Projection iteration converged (a black asterisk). On the second column, the same points have been checked against local minima, and the global minimum has been chosen correctly in all cases. The situation is very different for a fixating motion. We have considered the same situation just described, but where the cloud of dots rotates of $1^o/frame$ about an axis passing through the centroid. In this case we have considered noise of 0.05, 0.5 and 5 pixels std, corresponding to 2%, 20% and 200% of the measurements. Already at 2% noise we notice that the algorithm converges to local minima corresponding to both the saddle, the bas-relief ambiguity, and the rubbery motion perception. If we check for local minima, however, we can get to the correct estimate in all 200 trials (right). For 200% noise, however, the rubbery motion perception becomes stable, so that about 40% of the trials return the rubbery motion solution even after correction.

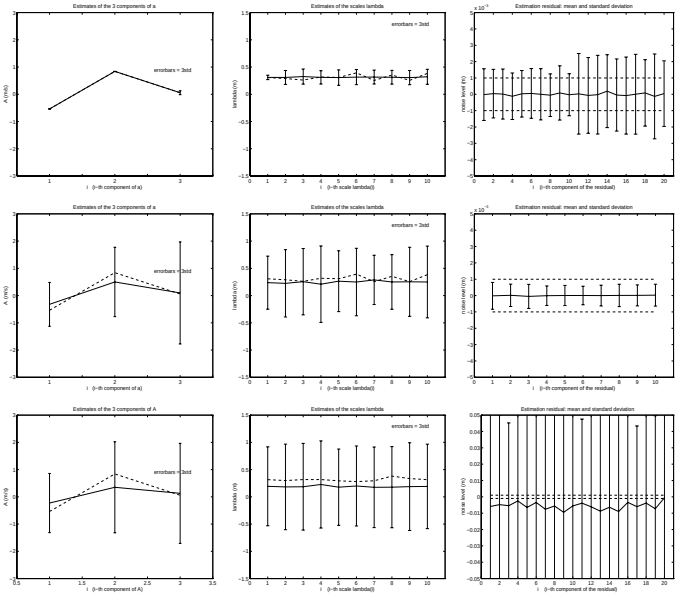

Figure 6: **Comparison with epipolar geometry and linear subspace methods** *(Top row) the estimates of translation (left) and the scales (center) obtained with the BPA are plotted with errorbars for 100 trials of the experiment. Noise is 50% of the measurements. The residual (right) is small, but it can be no smaller than the noise level, plotted in dashed lines. (Center row) algorithms based on the epipolar constraint perform considerably worse (left and center), despite the fact that the residual is smaller (right). The fact that the residual is smaller than the noise level is indeed a consequence of the fact that the epipolar constraint does not minimize the reprojection error. Linear subspace methods (bottom row) exhibit a bias both in the estimates (left and center) and in the residual (right).*

Comparison with other algorithms

In this section we compare the optimal algorithm proposed in section 3 with other approaches based upon epipolar geometry [4], and upon linear subspace methods [7]. We use as a representative of the first class the algorithm described in [19], and as a representative of linear subspace methods the one in [16]. We consider $p = 20$ points distributed uniformly in a cube of side $1m$ centered at $2m$, rotating at $1^o/frame$, with measurements on the unit sphere corrupted by 50% noise. In figure 6 we show the result of 100 trials. The upshot is, not surprisingly, that the optimal algorithm works better. The outcome of the experiments for linear subspace methods show that the estimates are biased, as reported in [16].

6 Conclusions

The assumption of "small noise" is often illegitimate in conditions normally encountered in real-world experiments with SFM. Therefore, SFM needs to be addressed from the point of view of noise. We have proposed a provably convergent algorithm (the bilinear projection iteration), characterized the set of local extrema, given a geometric interpretation and proven that they are intrinsic to SFM, and proposed a representation of shape that is invariant to local minima.

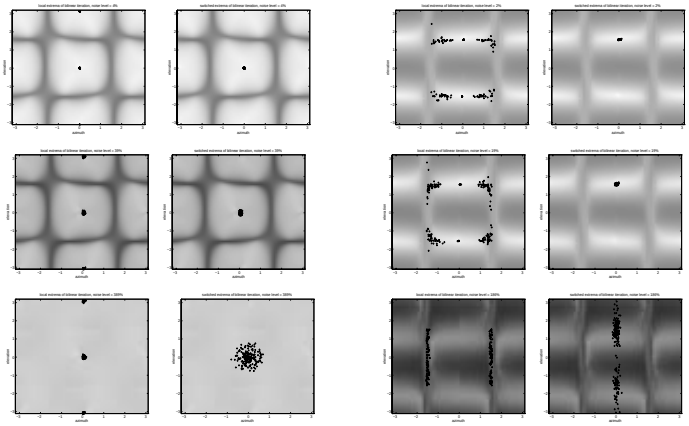

Figure 7: **Forward translation (left)** *The residual of the cost function r_2 is superimposed to the fixed points of the Bilinear Projection iteration before (left) and after (right) correction for local extrema. Noise is 4% (top), 40% (center) and 400% (bottom) of the measurements. Convergence to the valley of the global optimum is achieved in all 200 trials.* **Bas-relief ambiguity (right)** *The Bilinear Projection algorithm converges to local minima corresponding to both the bas-relief ambiguity and the rubbery interpretation (left). For noises of 2% (top row) and 20% (center row), checking for local extrema is sufficient to achieve the correct solution in all 200 trials. For 200% noise (bottom), the rubbery interpretation is stable in 30% of the trials.*

Acknowledgements We thank J. Oliensis and C. Tomasi for comments.

References

[1] G. Adiv. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 11(5), 1989.

[2] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *Int. J. of Computer Vision*, 12(1):43–78, 1994.

[3] J. Bergen, R. Kumar, P. Anandan, and M. Irani. Representation of scenes from collections of images. In *Proc. of the IEEE Workshop on Visual Scene Representation*, Boston, June 1995.

[4] O. D. Faugeras. *Three Dimensional Vision, a geometric viewpoint.* MIT Press, 1993.

[5] O. D. Faugeras and Q. T. Luong. in preparation, 1997.

[6] G. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least-squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10 (2):413–432, 1973.

[7] A. Jepson and D. Heeger. Linear subspace methods for recovering rigid motion. *Spatial Vision in Humans and Robots, Cambridge University Press*, 1992.

[8] K. Kanatani. Statistical Optimization for Geometric Computation. *Unpublished manuscript* 1997.

[9] J. Oliensis. Provably correct algorithms for multi-frame structure from motion. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1996.

[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] S. Soatto and R. Brockett. Optimal Structure From Motion. *Technical Report*, April 1997, revised Oct. 1997. `http://ee.wustl.edu/~soatto/papers/SFM-long.ps`.

[12] 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.

[13] M. Spetsakis and J. Aloimonos. Optimal motion estimation. *Proc. of the IEEE workshop on visual motion*, Irvine, March 1989.

[14] R. Szeliski and S. Kang. Recovering 3D shape and motion from images using nonlinear least squares. *J. Visual Communication and Image Representation*, vol.5 n. 1, 1994.

[15] R. Szeliski and S. Kang. Shape ambiguities in structure from motion. *IEEE Trans. on Patt. An. and Mach. Intell.*, vol.19 n. 5 , 1997.

[16] I. Thomas and E. Simoncelli. Linear Structure from Motion. *Technical Report IRCS 94-26*, University of Pennsylvania, 1994.

[17] T. Tian, C. Tomasi, and D. Heeger. Comparison of approaches to egomotion computation. In *proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1996.

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

[19] J. Weng, T. S. 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.

[20] G. Young and R. Chellappa. Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(10), 1992.