Optimization Criteria, Sensitivity and Robustness of Motion and Structure Estimation

Jana Košecká[1], Yi Ma[2], and Shankar Sastry[2]

[1] Computer Science Department, George Mason University, Fairfax, VA 22030
[2] EECS Department, University of California at Berkeley, Berkeley, CA 94720-1772
kosecka@cs.gmu.edu, {mayi, sastry}@eecs.berkeley.edu

Abstract. The prevailing efforts to study the standard formulation of motion and structure recovery have been recently focused on issues of sensitivity and robustness of existing techniques. While many cogent observations have been made and verified experimentally, many statements do not hold in general settings and make a comparison of existing techniques difficult. With an ultimate goal of clarifying these issues we study the main aspects of the problem: the choice of objective functions, optimization techniques and the sensitivity and robustness issues in the presence of noise.

We clearly reveal the relationship among different objective functions, such as "(normalized) epipolar constraints", "reprojection error" or "triangulation", which can all be be unified in a new " optimal triangulation" procedure formulated as a constrained optimization problem. Regardless of various choices of the objective function, the optimization problems all inherit the same unknown parameter space, the so called "essential manifold", making the new optimization techniques on Riemanian manifolds directly applicable.

Using these analytical results we provide a clear account of sensitivity and robustness of the proposed linear and nonlinear optimization techniques and study the analytical and practical equivalence of different objective functions. The geometric characterization of critical points of a function defined on essential manifold and the simulation results clarify the difference between the effect of bas relief ambiguity and other types of local minima leading to a consistent interpretations of simulation results over large range of signal-to-noise ratio and variety of configurations. [1]

1 Introduction

While the geometric relationships governing the motion and structure recovery problem have been long understood, the robust solutions in the presence of noise are still sought. New studies of sensitivity of different algorithms, search for intrinsic local minima and new algorithms are still subject of great interest.

The seminal work of Longuet-Higgins [9] on the characterization of the so called *epipolar constraint*, enabled the decoupling of the structure and motion

[1] This work is supported by ARO under the MURI grant DAAH04-96-1-0341

B. Triggs, A. Zisserman, R. Szeliski (Eds.): Vision Algorithms'99, LNCS 1883, pp. 166–183, 2000.

problems and led to the development of numerous linear and nonlinear algorithms for motion estimation (see [14,7,21] for overviews). The appeal of linear algorithms which use the epipolar constraint (in the discrete case [21,7,9,14] and in the differential case [6,13]) is the closed form solution to the problem which, in the absence of noise, provides true estimate of the motion. However, a further analysis of linear techniques revealed an inherent bias in the translation estimates [6,7]. The sensitivity studies of the motion estimation problem have been done both in an analytical [1,18] and experimental setting [19] and revealed the superiority of the nonlinear optimization schemes over the linear ones. Numerous nonlinear optimization schemes differed in the choice of objective functions [23], different parameterizations of the unknown parameter space [22,23,5] and means of initialization of the iterative schemes (e.g. monte-carlo simulations [21,17], or linear techniques [6]). In most cases, the underlying search space has been parameterized for computational convenience instead of being loyal to its intrinsic geometric structure. Algebraic manipulation of intrinsic geometric relationships typically gave rise to different objective functions, making the comparison of the performance of different techniques inappropriate and often obstructing the key issues of the problem. The goal of this paper is to evaluate intrinsic difficulties of the structure and motion recovery problem in the presence of *large* levels of noise, in terms of intrinsic local minima, bias, sensitivity and robustness. This evaluation is done with respect to the choice of objective function and optimization technique, in the simplified two-view, point-feature scenario. The main contributions presented in this paper are summarized briefly below:

1. We present a new optimal triangulation procedure and show that it can be formulated as an iterative two step constrained optimization: Motion estimation is formulated as optimization on the essential manifold and is followed by additional well conditioned minimization of two Raleigh quotients for estimating the structure. The procedure clearly reveals the relationship between existing objective functions used previously and exhibits superior (provable) convergence properties. This is possible thanks to the intrinsic nonlinear search schemes on the essential manifold, utilizing Riemanian structucture of the unknown parameter space.
2. We demonstrate analytically and by extensive simulations how the choice of the objective functions and configurations affects the sensitivity and robustness of the estimates, making a clear distinction between the two. We both observe and geometrically characterize how the patterns of critical points of the objective function change with increasing levels of noise for general configurations. We show the role of linear techniques for initialization and detection of these incorrect local minima. Further more we utilize the second order information to characterize the nature of the bas relief ambiguity and rotation and translation confounding for special class of "sensitive" motions/configurations.

Based on analytical and experimental results, we will give a clear profile of the performance of different algorithms over a large range of signal-to-noise ratio, and under various motion and structure configurations.

2 Optimization on the Essential Manifold

Suppose the camera motion is given by $(R, S) \in SE(3)$ (the special Euclidean group) where R is a rotation matrix in $SO(3)$ (the special orthogonal group) and $S \in \mathbb{R}^3$ is the translation vector. The intrinsic geometric relationship between two corresponding projections of a single 3D point in two images p and q (in homogeneous coordinates) then gives the so called *epipolar constraint* [9]:

$$p^T R\widehat{S}q = \tag{1}$$

where $\widehat{S} \in \mathbb{R}^{3\times 3}$ is defined such that $\widehat{S}v = S \times v$ for all $v \in \mathbb{R}^3$. Epipolar constraint decouples the problem of motion recovery from that of structure recovery. The first part of this paper will be devoted to recovering motion from directly using this constraint or its variations. In Section 3, we will see how this constraint has to be adjusted when we consider recovering motion and structure simultaneously.

The entity of our interest is the matrix $R\widehat{S}$ in the epipolar constraint; the so called *essential matrix*. The *essential manifold* is defined to be the space of all such matrices, denoted by $\mathcal{E} = \{R\widehat{S} \mid R \in SO(3), \widehat{S} \in so(3)\}$, where $SO(3)$ is a Lie group of 3×3 rotation matrices, and $so(3)$ is the Lie algebra of $SO(3)$, *i.e.*, the tangent plane of $SO(3)$ at the identity. $so(3)$ then consists of all 3×3 skew-symmetric matrices. The problem of motion recovery is equivalent to optimizing functions defined on the so called *normalized essential manifold*:

$$\mathcal{E}_1 = \{R\widehat{S} \mid R \in SO(3), \widehat{S} \in so(3), \frac{1}{2}tr(\widehat{S}^T\widehat{S}) = 1\}.$$

Note that $\frac{1}{2}tr(\widehat{S}^T\widehat{S}) = S^TS$. In order to formulate properly the optimization problem, it is crucial to understand the Riemannian structure of the normalized essential manifold. In our previous work we showed [11] that the space of essential matrices can be identified with the unit tangent bundle of the Lie group $SO(3)$, *i.e.*, $T_1(SO(3))$[2]. Further more its Riemannian metric g induced from the bi-invariant metric on $SO(3)$ is the same as that induced from the Euclidean metric with $T_1(SO(3))$ naturally embedded in $\mathbb{R}^{3\times 4}$. $(T_1(SO(3)), g)$ is the product Riemannian manifold of $(SO(3), g_1)$ and $(\mathbb{S}^2, g_2)$ with g_1 and g_2 canonical metrics for $SO(3)$ and $\mathbb{S}^2$ as Stiefel manifolds. Given this Riemannian structure of our unknown parameter space, we showed [13] that one can generalize Edelman *et al*'s methods [3] to the product Riemannian manifolds and obtain intrinsic geometric Newton's or conjugate gradient algorithms for solving such an optimization problem. Given the epipolar constraint, the problem of motion recovery R, S from a given set of image correspondences $p_i, q_i \in \mathbb{R}^3, i = 1, \ldots, N$, in the presence of noise can be naturally formulated as a minimization of the

[2] However, the unit tangent bundle $T_1(SO(3))$ is not exactly the normalized essential manifold $\mathcal{E}_1$. It is a double covering of the normalized essential space $\mathcal{E}_1$, *i.e.*, $\mathcal{E}_1 = T_1(SO(3))/\mathbb{Z}^2$ (for details see [11]).

following objective function:

$$F(R,S) = \sum_{i=1}^{N} (p_i^T R \widehat{S} q_i)^2 \tag{2}$$

for $p_i, q_i \in \mathbb{R}^3$, where $F(R,S)$ is a function defined on $T_1(SO(3)) \cong SO(3) \times \mathbb{S}^2$ with $R \in SO(3)$ represented by a 3×3 rotation matrix and $S \in \mathbb{S}^2$ a vector of unit length in $\mathbb{R}^3$. Due to the lack of space below we present only a summary of the Newton's algorithm for optimization of the above objective function on the essential manifold. Please refer for more details to [13] for this particular objective function and to [3] for the details of the Newton's or other conjugate gradient algorithms for general Stiefel or Grassmann manifolds.

Riemannian Newton's algorithm for minimizing $F(R,S)$:

1. *At the point (R,S),*
   - *Compute the gradient $G = (F_R - RF_R^T R,\ F_S - SF_S^T S)$,*
   - *Compute $\Delta = -\ Hess^{-1} G$.*
2. *Move (R,S) in the direction Δ along the geodesic to* $(\exp(R,\Delta_1),\ \exp(S,\Delta_2))$.
3. *Repeat if $\|G\| \geq \epsilon$ for pre-determined $\epsilon > 0$.*

$F_R(F_S)$ is a derivative of the objective function $F(R,S)$ with respect to its parameters.

The basic ingredients of the algorithm is the computation of the gradient and Hessian whose explicit formulas can be found in [13]. These formulas can be alternatively obtained by directly using the explicit expression of geodesics on this manifold. On $SO(3)$, the formula for the geodesic at R in the direction $\Delta_1 \in T_R(SO(3)) = R_*(so(3))$ is $R(t) = \exp(R, \Delta_1 t) = R \exp \widehat{\omega} t = R(I + \widehat{\omega} \sin t + \widehat{\omega}^2(1 - \cos t))$, where $t \in \mathbb{R}, \widehat{\omega} = R^T \Delta_1 \in so(3)$. The last equation is called the *Rodrigues' formula* (see [16]). $\mathbb{S}^2$ (as a Stiefel manifold) also has very simple expression of geodesics. At the point S along the direction $\Delta_2 \in T_S(\mathbb{S}^2)$ the geodesic is given by $S(t) = \exp(S, \Delta_2 t) = S \cos \sigma t + U \sin \sigma t$, where $\sigma = \|\Delta_2\|$ and $U = \Delta_2/\sigma$, then $S^T U = 0$ since $S^T \Delta_2 = 0$. Using these formulae for geodesics , we can calculate the first and second derivatives of $F(R,S)$ in the direction $\Delta = (\Delta_1, \Delta_2) \in T_R(SO(3)) \times T_S(\mathbb{S}^2)$. The explicit formula for the Hessian obtained in this manner plays an important role for sensitivity analysis of the motion estimation [1] as we will point out in the second part of the paper. Furthermore, using this formula, we have shown [13] that the conditions when the Hessian is guaranteed non-degenerate are the same as the conditions for the linear 8-point algorithm having a unique solution; whence the Newton's algorithm has quadratic rate of convergence.

2.1 Minimizing Normalized Epipolar Constraints

Although the epipolar constraint (1) gives the only necessary (depth independent) condition that image pairs have to satisfy, motion estimates obtained from minimizing the objective function (2) are not necessarily statistically or geometrically optimal for the commonly used noise model of image correspondences. In

general, in order to get less biased estimates, we need to *normalize* (or weight) the epipolar constraints properly, which has been initially observed in [22]. In this section, we will give a brief account of these normalized versions of epipolar constraints. In the perspective projection case[3], coordinates of image points p and q are of the form $p = (p^1, p^2, 1)^T \in \mathbb{R}^3$ and $q = (q^1, q^2, 1)^T \in \mathbb{R}^3$. Suppose that the actual measured image coordinates of N pairs of image points are: $p_i = \tilde{p}_i + x_i, \quad q_i = \tilde{q}_i + y_i$ for $i = 1, \ldots, N$, where $\tilde{p}_i$ and $\tilde{q}_i$ are ideal (noise free) image coordinates, $x_i = (x_i^1, x_i^2, 0)^T \in \mathbb{R}^3$, $y_i = (y_i^1, y_i^2, 0)^T \in \mathbb{R}^3$ and $x_i^1, x_i^2, y_i^1, y_i^2$ are independent Gaussian random variables of identical distribution $N(0, \sigma^2)$. Substituting p_i and q_i into the epipolar constraint (1), we obtain:

$$p_i^T R\widehat{S}q_i = x_i^T R\widehat{S}\tilde{q}_i + \tilde{p}_i^T R\widehat{S}y_i + x_i^T R\widehat{S}y_i.$$

Since the image coordinates p_i and q_i usually are magnitude larger than x_i and y_i, one can omit the last term in the equation above. Then $p_i^T R\widehat{S}q_i$ are independent random variables *approximately* of Gaussian distribution $N(0, \sigma^2(\|\widehat{e}_3 R\widehat{S}q_i\|^2 + \|p_i^T R\widehat{S}\widehat{e}_3\|^2))$, where $e_3 = (0, 0, 1)^T \in \mathbb{R}^3$. If we assume the *a prior* distribution of the motion (R, S) is uniform, the maximum *a posterior* (MAP) estimates of (R, S) is then the global minimum of the objective function:

$$F_s(R, S) = \sum_{i=1}^{N} \frac{(p_i^T R\widehat{S}q_i)^2}{\|\widehat{e}_3 R\widehat{S}q_i\|^2 + \|p_i^T R\widehat{S}\widehat{e}_3\|^2} \tag{3}$$

for $p_i, q_i \in \mathbb{R}^3, (R, S) \in SO(3) \times \mathbb{S}^2$. We here use F_s to denote the *statistically normalized* objective function associated with the epipolar constraint. This objective function is also referred in the literature under the name *gradient criteria* or *epipolar improvement*. Therefore, we have $(R, S)_{MAP} \approx \arg\min F_s(R, S)$. Note that in the noise free case, F_s achieves zero, just like the unnormalized objective function F of equation (2). Asymptotically, MAP estimates approach the unbiased minimum mean square estimates (MMSE). So, in general, the MAP estimates give less biased estimates than the unnormalized objective function F. Note that F_s is still a function defined on the manifold $SO(3) \times \mathbb{S}^2$. Another commonly used criteria to recover motion is to minimize the geometric distances between image points and corresponding epipolar lines. This objective function is given as:

$$F_g(R, S) = \sum_{i=1}^{N} \frac{(p_i^T R\widehat{S}q_i)^2}{\|\widehat{e}_3 R\widehat{S}q_i\|^2} + \frac{(p_i^T R\widehat{S}q_i)^2}{\|p_i^T R\widehat{S}\widehat{e}_3^T\|^2} \tag{4}$$

for $p_i, q_i \in \mathbb{R}^3, (R, S) \in SO(3) \times \mathbb{S}^2$. We here use F_g to denote this *geometrically normalized* objective function. Notice that, similar to F and F_s, F_g is also a function defined on the essential manifold and can be minimized using the given Newton's algorithm. As we know from the differential case [12], the normalization has no effect when the translational motion is in the image plane, *i.e.*, the

[3] The spherical projection case is similar and is omitted for simplicity.

unnormalized and normalized objective functions are in fact equivalent. For the discrete case, we have similar claim [8]. Therefore in such case the normalization will have very little effect on motion estimation as will be verified by simulation.

3 Optimal Triangulation

Note that, in the presence of noise, for the motion (R, S) recovered from minimizing the unnormalized or normalized objective functions F, F_s or F_g, the value of the objective functions is not necessarily zero. Consequently, if one directly uses p_i and q_i to recover the 3D location of the point to which the two images p_i and q_i correspond, the two rays corresponding to p_i and q_i may not be coplanar, hence may not intersect at one 3D point. Also, when we derived the normalized epipolar constraint F_s, we ignored the second order terms. Therefore, rigorously speaking, it does not give the exact MAP estimates. Under the assumption of Gaussian noise model, in order to obtain the optimal (MAP) estimates of camera motion and a consistent 3D structure reconstruction, in principle we need to solve the following optimal triangulation problem: Seek camera motion (R, S) and points $\tilde{p}_i, \tilde{q}_i \in \mathbb{R}^3$ on the image plane such that they minimize the distance from p_i and q_i:

$$F_t(R, S, \tilde{p}_i, \tilde{q}_i) = \sum_{i=1}^{N} \|\tilde{p}_i - p_i\|^2 + \|\tilde{q}_i - q_i\|^2 \tag{5}$$

subject to the conditions: $\tilde{p}_i^T R\widehat{S}\tilde{q}_i = 0, \quad \tilde{p}_i^T e_3 = 1, \quad \tilde{q}_i^T e_3 = 1$ for $i = 1, \ldots, N$. We here use F_t to denote the objective function for triangulation. This objective function is also referred in literature as the reprojection error. Unlike [4], we do not assume a known essential matrix $R\widehat{S}$. Instead we seek $\tilde{p}_i, \tilde{q}_i$ *and* (R, S) which minimize the objective function F_t given by (5). The objective function F_t then implicitly depends on the variables (R, S) through the constraints. Clearly, the optimal solution to this problem is exactly equivalent to the optimal MAP estimates of both motion *and* structure. Using Lagrangian multipliers, we can convert the minimization problem to an unconstrained one:

$$\min_{R,S,\tilde{p}_i,\tilde{q}_i} \sum_{i=1}^{N} \|\tilde{p}_i - p_i\|^2 + \|\tilde{q}_i - q_i\|^2 + \lambda_i \tilde{p}_i^T R\widehat{S}\tilde{q}_i + \beta_i(\tilde{p}_i^T e_3 - 1) + \gamma_i(\tilde{q}_i^T e_3 - 1).$$

The necessary conditions for minima of this objective function are:

$$2(\tilde{p}_i - p_i) + \lambda_i R\widehat{S}\tilde{q}_i + \beta_i e_3 = 0 \tag{6}$$

$$2(\tilde{q}_i - q_i) + \lambda_i \widehat{S}^T R^T \tilde{p}_i + \gamma_i e_3 = \tag{7}$$

From necessary conditions we get $\tilde{p}_i$, $\tilde{q}_i$. Substituting these and λ_i obtained from (6) back to into F_t we get:

$$F_t(R, S, \tilde{p}_i, \tilde{q}_i) = \sum_{i=1}^{N} \frac{(p_i^T R\widehat{S}\tilde{q}_i + \tilde{p}_i^T R\widehat{S}q_i)^2}{\|\widehat{e}_3 R\widehat{S}\tilde{q}_i\|^2 + \|\tilde{p}_i^T R\widehat{S}\widehat{e}_3^T\|^2} \tag{8}$$

and alternatively using (7) for λ_i instead, we get:

$$F_t(R, S, \tilde{p}_i, \tilde{q}_i) = \sum_{i=1}^{N} \frac{(p_i^T R \widehat{S} \tilde{q}_i)^2}{\|\widehat{e}_3 R \widehat{S} \tilde{q}_i\|^2} + \frac{(\tilde{p}_i^T R \widehat{S} q_i)^2}{\|\tilde{p}_i^T R \widehat{S} \widehat{e}_3^T\|^2}. \tag{9}$$

Geometrically, both expressions of F_t are the distances from the image points p_i and q_i to the epipolar lines specified by $\tilde{p}_i, \tilde{q}_i$ and (R, S). Equations (8) and (9) give explicit formulae of the residue of $\|\tilde{p}_i - p_i\|^2 + \|\tilde{q}_i - q_i\|^2$ as p_i, q_i being triangulated by $\tilde{p}_i, \tilde{q}_i$. Note that the terms in F_t are normalized *crossed epipolar constraints* between p_i and $\tilde{q}_i$ or between $\tilde{p}_i$ and q_i. These expressions of F_t can be further used to solve for (R, S) which minimizes F_t. This leads to the following iterative scheme for obtaining optimal estimates of both motion and structure, without explicitly introducing scale factors (or depths) of the 3D points.

Optimal Triangulation Algorithm Outline: *The procedure for minimizing F_t can be outlined as follows:*

1. *Initialize $\tilde{p}_i^*(R, S), \tilde{q}_i^*(R, S)$ as p_i, q_i.*
2. **Motion:** *Update (R, S) by minimizing $F_t^*(R, S) = F_t(R, S, \tilde{p}_i^*(R, S), \tilde{q}_i^*(R, S))$ given by (8) or (9) as a function defined on the manifold $SO(3) \times \mathbb{S}^2$.*
3. **Structure (Triangulation):** *Solve for $\tilde{p}_i^*(R, S)$ and $\tilde{q}_i^*(R, S)$ which minimize the objective function F_t (5) with respect to (R, S) computed in the previous step.*
4. *Back to step 2 until updates are small enough.*

At step 3, for a fixed (R, S), $\tilde{p}_i^*(R, S)$ and $\tilde{q}_i^*(R, S)$ can be computed by minimizing the distance $\|\tilde{p}_i - p_i\|^2 + \|\tilde{q}_i - q_i\|^2$ for each pair of image points. Let $t_i \in \mathbb{R}^3$ be the normal vector (of unit length) to the (epipolar) plane spanned by $(\tilde{q}_i, S)$. Given such a t_i, $\tilde{p}_i$ and $\tilde{q}_i$ are determined by:

$$\tilde{p}_i(t_i) = \frac{\widehat{e}_3 t_i' {t'}_i^T \widehat{e}_3^T p_i + \widehat{t'}_i^T \widehat{t'}_i e_3}{e_3^T \widehat{t'}_i^T \widehat{t'}_i e_3}, \quad \tilde{q}_i(t_i) = \frac{\widehat{e}_3 t_i t_i^T \widehat{e}_3^T q_i + \widehat{t}_i^T \widehat{t}_i e_3}{e_3^T \widehat{t}_i^T \widehat{t}_i e_3} \tag{10}$$

where $t_i' = R t_i$. Then the distance can be explicitly expressed as:

$$\|\tilde{q}_i - q_i\|^2 + \|\tilde{p}_i - p_i\|^2 = \|q_i\|^2 + \frac{t_i^T A_i t_i}{t_i^T B_i t_i} + \|p_i\|^2 + \frac{{t'}_i^T C_i t_i'}{{t'}_i^T D_i t_i'},$$

where

$$\begin{aligned} A_i &= I - (\widehat{e}_3 q_i q_i^T \widehat{e}_3^T + \widehat{q}_i \widehat{e}_3 + \widehat{e}_3 \widehat{q}_i), \quad B_i = \widehat{e}_3^T \widehat{e}_3 \\ C_i &= I - (\widehat{e}_3 p_i p_i^T \widehat{e}_3^T + \widehat{p}_i \widehat{e}_3 + \widehat{e}_3 \widehat{p}_i), \quad D_i = \widehat{e}_3^T \widehat{e}_3 \end{aligned}. \tag{11}$$

Then the problem of finding $\tilde{p}_i^*(R, S)$ and $\tilde{q}_i^*(R, S)$ becomes one of finding t_i^* which minimizes the function of a sum of two *singular Rayleigh quotients*:

$$\min_{t_i^T S = 0, t_i^T t_i = 1} V(t_i) = \frac{t_i^T A_i t_i}{t_i^T B_i t_i} + \frac{t_i^T R^T C_i R t_i}{t_i^T R^T D_i R t_i}. \tag{12}$$

This is an optimization problem on a unit circle $\mathbb{S}^1$ in the plane orthogonal to the vector S. If $n_1, n_2 \in \mathbb{R}^3$ are vectors such that S, n_1, n_2 form an orthonormal basis of $\mathbb{R}^3$, then $t_i = \cos(\theta)n_1 + \sin(\theta)n_2$ with $\theta \in \mathbb{R}$. We only need to find θ^* which minimizes the function $V(t_i(\theta))$. From the geometric interpretation of the optimal solution, we also know that the global minimum θ^* should lie between two values: θ_1 and θ_2 such that $t_i(\theta_1)$ and $t_i(\theta_2)$ correspond to normal vectors of the two planes spanned by (q_i, S) and $(R^T p_i, S)$ respectively (if p_i, q_i are already triangulated, these two planes coincide). Therefore, in our approach the local minima is no longer an issue for triangulation, as oppose to the method proposed in [4]. The problem now becomes a simple bounded minimization problem for a scalar function and can be efficiently solved using standard optimization routines (such as "fmin" in Matlab or the Newton's algorithm). If one properly parameterizes $t_i(\theta)$, t_i^* can also be obtained by solving a 6-degree polynomial equation, as shown in [4] (and an approximate version results in solving a 4-degree polynomial equation [21]). However, the method given in [4] involves coordinate transformation for each image pair and the given parameterization is by no means canonical. For example, if one chooses instead the commonly used parameterization of a circle $\mathbb{S}^1$: $\sin(2\theta) = \frac{2\lambda}{1+\lambda^2}$, $\cos(2\theta) = \frac{1-\lambda^2}{1+\lambda^2}$, $\lambda \in \mathbb{R}$, then it is straightforward to show from the Rayleigh quotient sum (12) that the necessary condition for minima of $V(t_i)$ is equivalent to a 6-degree polynomial equation in λ.[4] The triangulated pairs $(\tilde{p}_i, \tilde{q}_i)$ and the camera motion (R, S) obtained from the minimization automatically give a consistent (optimal) 3D structure reconstruction by two-frame stereo.

In the expressions of F_t given by (18) or (19), if we simply approximate $\tilde{p}_i, \tilde{q}_i$ by p_i, q_i respectively, we may obtain the normalized versions of epipolar constraints for recovering camera motion. Although subtle difference between F_s, F_g and F_t has previously been pointed out in [23], our approach discovers that all these three objective functions can be unified in the same optimization procedure – they are just slightly different approximations of the same objective function F_t^*. Practically speaking, using either normalized objective function F_s or F_g, one can already get camera motion estimates which are very close to the optimal ones. This will be demonstrated by extensive simulations in the next section.

4 Critical Values and Ambiguous Solutions

We devote the remainder of this paper to study of the robustness and sensitivity of motion and structure estimation problem in the presence of large levels of noise. We emphasize here the role of the linear techniques for initialization and utilize the characterization of the space of essential matrices and the intrinsic optimization techniques on the essential manifold. The focus of our robustness

[4] Since there is no closed form solution to 6-degree polynomial equations, directly minimizing the Rayleigh quotient sum (12) avoids unnecessary transformations hence can be much more efficient.

study deals with the appearance of new local minima. Like any nonlinear system, when increasing the noise level, new critical points of the objective function can be introduced through bifurcation. Although in general an objective function could have numerous critical points, numbers of different types of critical points have to satisfy the so called *Morse inequalities*, which are associated to topological invariants of the underlying manifold (see [15]). Key to this study is the computation of the *Euler characteristic* $\chi(M)$ of the underlying manifold $SO(3) \times \mathbb{RP}^2$ which is in this case 0; $\chi(SO(3) \times \mathbb{RP}^2) = 0$. Euler characteristic is equal to $\sum_{\lambda=0}^{n}(-1)^{\lambda} D_\lambda$, where D_λ is the dimension of the λ^{th} homology group $H_\lambda(M, \mathbb{K})$ of M over any field $\mathbb{K}$, the so called λ^{th} *Betti number*. In our case $D_\lambda = 1, 2, 3, 3, 2, 1$ for $\lambda = 0, 1, 2, 3, 4, 5$ types of critical points respectively. For details of this computation see [13]. Among all the critical points, those belonging to type 0 are called (local) *minima*, type n are (local) *maxima*, and types 1 to $n-1$ are *saddles*. From the above computation any Morse function defined on $SO(3) \times \mathbb{RP}^2$ must have all three kinds of critical values. The nonlinear search algorithms proposed in the above are trying to find the global minimum of given objective functions. We study the effect of initialization by linear techniques and apperance of new critical points on different slices of the nonlinear objective function which we can be easily visualized. The choice of the section is determined by the estimate of rotation where the nonlinear algorithm converged by initialization of the linear algorithm.

Rewriting the epipolar constraint as $p_i^T E q_i = 0, i = 1, \ldots, N$, minimizing the objective function F is (approximately) equivalent to the following least square problem $\min \|Ae\|^2$, where A is a $N \times 9$ matrix function of entries of p_i and q_i, and $e \in \mathbb{R}^9$ is a vector of the nine entries of E. Then e is the (usually one dimensional) null space of the 9×9 symmetric matrix $A^T A$. In the presence of noise, e is simply chosen to be the eigenvector corresponding to the least eigenvalue of $A^T A$. At a low noise level, this eigenvector in general gives a good initial estimate of the essential matrix. However, at a certain high noise level, the smallest two eigenvalues may switch roles, as do the two corresponding eigenvectors – topologically, a bifurcation as shown in Figure 2 occurs. This phenomena is very common in the motion estimation problem: at a high noise level, the translation estimate may suddenly change direction by roughly 90^o, especially in the case when translation is parallel to the image plane. We will refer to such estimates as the *second eigenmotion*. A similar situation for the differential case and small field of view has previously been reported in [2].

Figure 1 and 2 demonstrate such a sudden appearance of the second eigenmotion. They are the simulation results of the proposed nonlinear algorithm of minimizing the function F_s for a cloud of 40 randomly generated pairs of image correspondences (in a field of view 90^o, depth varying from 100 to 400 units of focal length.). Gaussian noise of standard deviation of 6.4 or 6.5 pixels is added on each image point (image size 512×512 pixels). To make the results comparable, we used the same random seeds for both runs. The actual rotation is 10^o about

the Y-axis and the actual translation is along the X-axis.[5] The ratio between translation and rotation is 2.[6] In the figures, "+" marks the actual translation, "$*$" marks the translation estimate from linear algorithm (see [14] for detail) and "$\circ$" marks the estimate from nonlinear optimization. Up to the noise level of 6.4 pixels, both rotation and translation estimates are very close to the actual motion. Increasing the noise level further by 0.1 pixel, the translation estimate suddenly switches to one which is roughly 90^o away from the actual translation. Geometrically, this estimate corresponds to the second smallest eigenvector of the matrix $A^T A$ as we discussed before. Topologically, this estimate corresponds to the local minimum introduced by a bifurcation as shown by Figure 2. Clearly, in Figure 1, there is 1 maximum, 1 saddle and 1 minimum on $\mathbb{RP}^2$; in Figure 2, there is 1 maximum, 2 saddles and 2 minima. Both patterns give the Euler characteristic of $\mathbb{RP}^2$ as 1. Rotation is fixed at the estimate from nonlinear algorithm. The errors are expressed in terms of canonical metric on $SO(3)$ for rotation and in terms of angle for translation.

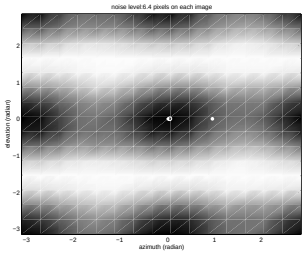

Fig. 1. Value of objective function F_s for all S at noise level 6.4 pixels. Estimation errors: 0.014 in rotation estimate and 2.39^o in translation estimate.

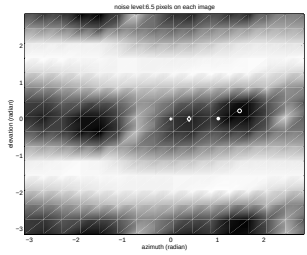

Fig. 2. Value of objective function F_s for all S at noise level 6.5 pixels. Estimation errors: 0.227 in rotation estimate and 84.66^o in translation estimate.

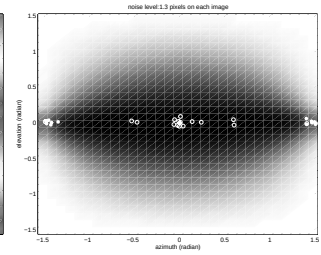

Fig. 3. Bas relief ambiguity. FOV is 20^o, points depths vary from 100 to 150 units of focal length, rotation magnitude is 2^o, T/R ratio is 2. 20 runs with noise level 1.3 pixels.

From the Figure 2, we can see that the the second eigenmotion ambiguity is even more likely to occur (at certain high noise level) than the other local minimum marked by "$\diamond$" in the figure which is a legitimate estimate of the actual one. These two estimates always occur in pair and exist for general configuration even when both the FOV and depth variation are sufficiently large. We propose a way for resolving the second eigenmotion ambiguity by linear algorithm which is used for initialization. An indicator of the configuration being close to critical

[5] We here use the convention that Y-axis is the vertical direction of the image and X-axis is the horizontal direction and the Z-axis coincides with the optical axis of the camera.

[6] Rotation and translation magnitudes are compared with respect to the center of the cloud of 3D points generated.

is the ratio of the two smallest eigenvalues of $A^T A$ σ_9 and σ_8 . By using both eigenvectors v_9 and v_8 for computing the linear motion estimates, the one which satisfies the positive depth constraint by larger margin (i.e. larger number of points satisfies the positive depth constraint) leads to the motion estimates closer to the true one (see [8] for details).

This second eigenmotion effect has a quite different interpretation as the one which was previously attributed to the bas relief ambiguity. The bas relief effect is only evident when FOV and depth variation is small, but the second eigenmotion ambiguity may show up for general configurations. Bas relief estimates are statistically meaningful since they characterize a sensitive direction in which translation and rotation are the most likely to be confound. The second eigenmotion, however, is not statistically meaningful: it is an effect of initialization which with increasing noise level causes a perturbation to a different slice of the objective function with a different topology of the residual. This effect occurs only at a high noise level and this critical noise level gives a measure of the *robustness* of linear initialization of the given algorithm. For comparison, Figure 3 demonstrates the effect of the bas relief ambiguity: the long narrow valley of the objective function corresponds to the direction that is the most sensitive to noise.[7] Translation is along the X-axis and rotation around the Y-axis. The (translation) estimates of 20 runs, marked as "$\circ$", give a distribution roughly resembling the shape of this valley – the actual translation is marked as "+"in the center of the valley which is covered by circles.

5 Experiments and Sensitivity Analysis

In this section, we clearly demonstrate by experiments the relationship among the linear algorithm (as in [14]), nonlinear algorithm (minimizing F), normalized nonlinear algorithm (minimizing F_s) and optimal triangulation (minimizing F_t). Due to the nature of the second eigenmotion ambiguity (when not corrected), it gives statistically meaningless estimates. Such estimates should be treated as "outliers" if one wants to properly evaluate a given algorithm and compare simulation results. We will demonstrate that seemingly conflicting statements in the literature about the performance of existing algorithms can in fact be given a *unified* explanation if we systematically compare the simulation results with respect to a *large range* of noise levels (as long as the results are statistically meaningful).

The following simulations were carried out with the points in general configuration and camera parameters described in Section 4. All nonlinear algorithms are initialized by the estimates from the standard 8-point linear algorithm (see [14]), instead of from the ground truth. The criteria for all nonlinear algorithms to stop are: (a) The norm of gradient is less than a given error tolerance, which

[7] This direction is given by the eigenvector of the Hessian associated with the smallest eigenvalue.

usually we pick as 10^{-8} unless otherwise stated;[8] and (b) The smallest eigenvalue of the Hessian matrix is positive.[9]

Axis Dependency Profile It has been well known that the sensitivity of the motion estimation depends on the camera motion. However, in order to give a clear account of such a dependency, one has to be careful about two things: 1. The signal-to-noise ratio and 2. Whether the simulation results are still statistically meaningful while varying the noise level. Figure 4, 5, 6 and 7 give simulation results of 100 trials for each combination of translation and rotation ("T-R") axes, for example, "X-Y" means translation is along the X-axis and the rotation axis is the Y-axis. Rotation is always 10^o about the axis and the T/R ratio is 2. In the figures, "linear" stands for the standard 8-point linear algorithm; "nonlin" is the Riemannian Newton's algorithm minimizing the epipolar constraints F, "normal" is the Riemannian Newton's algorithm minimizing the normalized epipolar constraints F_s.

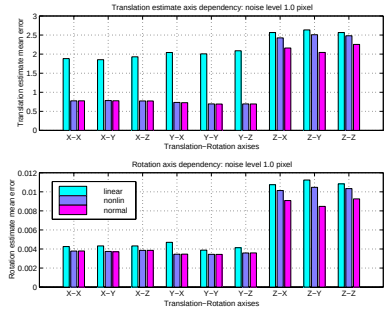

Fig. 4. Axis dependency: estimation errors in rotation and translation at noise level 1.0 pixel. T/R ratio = 2 and rotation = 10^o.

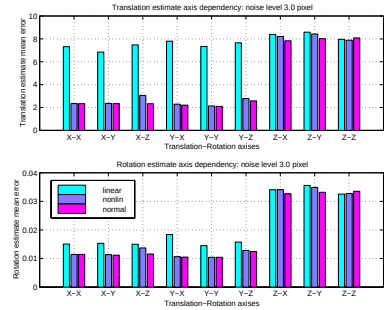

Fig. 5. Axis dependency: estimation errors in rotation and translation at noise level 3.0 pixels. T/R ratio = 2 and rotation = 10^o.

By carefully comparing the simulation results in Figure 4, 5, 6 and 7, we can draw the following conclusions:

1. **Optimization Techniques (linear vs. nonlinear)**
   (a) Minimizing F in general gives better estimates than the linear algorithm at low noise levels (Figure 4 and 5). At higher noise levels, this is no longer true (Figure 6 and 7), due to the more global nature of the linear technique.
   (b) Minimizing the normalized F_s in general gives better estimates than the linear algorithm at moderate noise levels (all figures).

[8] Our current implementation of the algorithms in Matlab has a numerical accuracy at 10^{-8}.

[9] Since we have the explicit formulae for Hessian, this condition would keep the algorithms from stopping at saddle points.

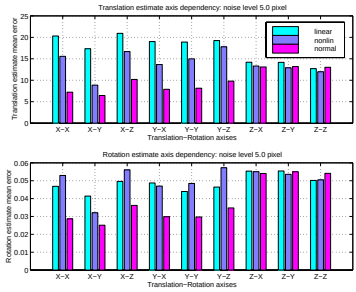

Fig. 6. Axis dependency: estimation errors in rotation and translation at noise level 5.0 pixel. T/R ratio = 2 and rotation = 10^o.

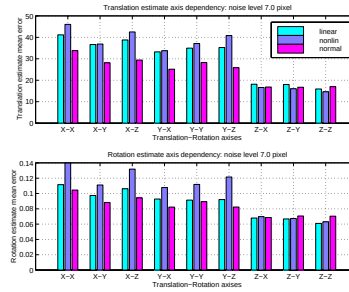

Fig. 7. Axis dependency: estimation errors in rotation and translation at noise level 7.0 pixels. T/R ratio = 2 and rotation = 10^o.

2. **Optimization Criteria (F vs. F_s)**
   (a) At relatively low noise levels (Figure 4), normalization has little effect when translation is parallel to the image plane; and estimates are indeed improved when translation is along the Z-axis.
   (b) However, at moderate noise levels (Figure 5, 6 and 7), when translation is along the Z-axis, little improvement can be gained by minimizing F_s instead of F; however, when translation is parallel to the image plane, F is more sensitive to noise and minimizing the statistically less biased F_s consistently improves the estimates.
3. **Axis Dependency**
   (a) All three algorithms are the most robust to the increasing of noise when the translation is along Z. At moderate noise levels (all figures), their performances are quite close to each other.
   (b) Although, at relatively low noise levels (Figure 4, 5 and 6), estimation errors seem to be larger when the translation is along the Z-axis, estimates are in fact much less sensitive to noise and more robust to increasing of noise in this case. The larger estimation error in case of translation along Z-axis is because the displacements of image points are smaller than those when translation is parallel to the image plane, thus the signal-to-noise ratio is in fact smaller.
   (c) At a noise level of 7 pixels (Figure 7), estimation errors seem to become smaller when the translation is along Z-axis. This is due to the fact that, at a noise level of 7 pixels, the second eigenmotion ambiguity already occurs in some of the trials when the translation is parallel to the image plane.

The second statement about the axis dependency supplements the observation given in [20]. In fact, the motion estimates are both robust and less sensitive to increasing of noise when translation is along the Z-axis. For a fixed base line, high noise level results resemble those for a smaller base line at a moderate noise level. Figure 7 is therefore a generic picture of the axis dependency profile for the differential or small base-line case (for more details see [12]).

Non-iterative vs. Iterative In general, the motion estimates obtained from directly minimizing the normalized epipolar constraints F_s or F_g are already

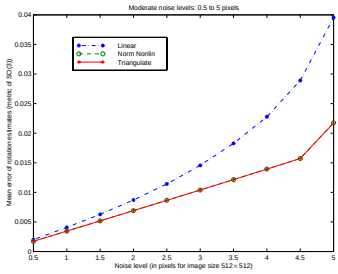

Fig. 8. Estimation errors of rotation (in canonical metric on $SO(3)$). 50 trials, rotation 10 degree around Y-axis and translation along X-axis, T/R ratio is 2. Noises range from 0.5 to 5 pixels.

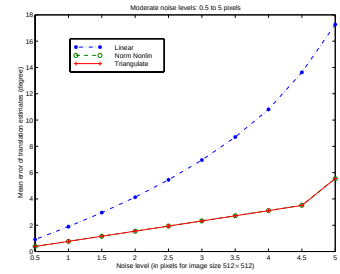

Fig. 9. Estimation errors of translation (in degree). 50 trials, rotation 10 degree around Y-axis and translation along X-axis, T/R ratio is 2. Noises range from 0.5 to 5 pixels.

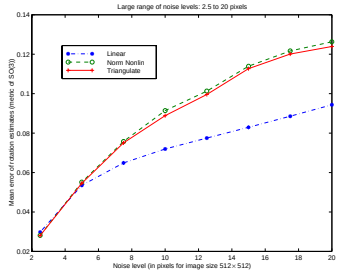

Fig. 10. Estimation errors of rotation (in canonical metric on $SO(3)$). 40 points, 50 trials, rotation 10 degree around Y-axis and translation along Z-axis, T/R ratio is 2. Noises range from 2.5 to 20 pixels.

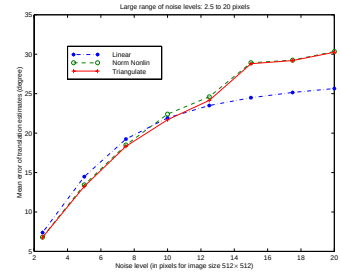

Fig. 11. Estimation errors of translation (in degree). 40 points, 50 trials, rotation 10 degree around Y-axis and translation along Z-axis, T/R ratio is 2. Noises range from 2.5 to 20 pixels.

very close to the solution of the optimal triangulation obtained by minimizing F_t iteratively between motion and structure. It is already known that, at low noise levels, the estimates from the non-iterative and iterative schemes usually differ by less than a couple of percent [23].

By comparing the simulation results in Figures 8, 9, 10 and 11 we can draw the following conclusions:

1. Although the iterative optimal triangulation algorithm usually gives better estimates (as it should), the non-iterative minimization of the normalized epipolar constraints F_s or F_g gives motion estimates with only a few percent larger errors for all range of noise levels. The higher the noise level, the more evident the improvement of the iterative scheme is.
2. Within moderate noise levels, normalized nonlinear algorithms consistently give significantly better estimates than the standard linear algorithm, especially when

the translation is parallel to the image plane. At very high noise levels, the performance of the standard linear algorithm, out performs nonlinear algorithms. This is due to the more global nature of the linear algorithm. However, such high noise levels are barely realistic in real applications.

For low level Gaussian noises, the iterative optimal triangulation algorithm gives the MAP estimates of the camera motion and scene structure, the estimation error can be shown close to the theoretical error bounds, such as the Cramer-Rao bound. This has been shown experimentally in [21]. Consequently, minimizing the normalized epipolar constraints F_s or F_g gives motion estimates close to the error bound as well.

6 Discussions and Future Work

Although previously proposed algorithms already have good performance in practice, the geometric concepts behind them have not yet been completely revealed. The non-degeneracy conditions and convergence speed of those algorithms are usually not explicitly addressed. Due to the recent development of optimization methods on Riemannian manifolds, we now can have a better mathematical understanding of these algorithms, and propose new geometric algorithms or filters, which exploit the intrinsic geometric structure of the motion and structure recovery problem. As shown in this paper, regardless of the choice of different objectives, the problem of optimization on the essential manifold is common and essential to the optimal motion and structure recovery problem. Furthermore, from a pure optimization theoretic viewpoint, most of the objective functions previously used in the literature can be unified in a single optimization procedure. Consequently, "minimizing (normalized) epipolar constraints", "triangulation", "minimizing reprojection errors" are all different (approximate) versions of the same simple optimal triangulation algorithm.

In this paper, we have studied in detail the problem of recovering a discrete motion (displacement) from image correspondences. Similar ideas certainly apply to the differential case where the rotation and translation are replaced by angular and linear velocities respectively [13]. One can show that they all in fact minimize certain normalized versions of the differential epipolar constraint. We hope the Riemannian optimization theoretic viewpoint proposed in this paper will provide a different perspective to revisit these schemes. Although the study of the proposed algorithms is carried out in a calibrated camera framework, due to a clear geometric connection between the calibrated and uncalibrated case [10], the same approach and optimization schemes can be generalized with little effort to the uncalibrated case as well. Details will be presented in future work.

References

1. K. Danilidis. *Visual Navigation*, chapter "Understanding Noise Sensitivity in Structure from Motion". Lawrence Erlbaum Associates, 1997.

2. K. Danilidis and H.-H. Nagel. Analytical results on error sensitivity of motion estimation from two views. *Image and Vision Computing*, 8:297–303, 1990.
3. A. Edelman, T. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Analysis Applications*, to appear.
4. R. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–57, 1997.
5. B. Horn. Relative orientation. *International Journal of Computer Vision*, 4:59–78, 1990.
6. A. D. Jepson and D. J. Heeger. Linear subspace methods for recovering translation direction. *Spatial Vision in Humans and Robots, Cambridge Univ. Press*, pages 39–62, 1993.
7. K. Kanatani. *Geometric Computation for Machine Vision*. Oxford Science Publications, 1993.
8. J. Košecká, Y. Ma, and S. Sastry. Optimization criteria, sensitivity and robustness of motion and structure estimation. In *Vision Algorithms Workshop, ICCV*, pages 9–16, 1999.
9. H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.
10. Y. Ma, J. Košecká, and S. Sastry. A mathematical theory of camera self-calibration. *Electronic Research Laboratory Memorandum, UC Berkeley*, UCB/ERL M98/64, October 1998.
11. Y. Ma, J. Košecká, and S. Sastry. Motion recovery from image sequences: Discrete viewpoint vs. differential viewpoint. In *Proceeding of European Conference on Computer Vision, Volume II, (also Electronic Research Laboratory Memorandum M98/11, UC Berkeley)*, pages 337–53, 1998.
12. Y. Ma, J. Košecká, and S. Sastry. Linear differential algorithm for motion recovery: A geometric approach. *Submitted to IJCV*, 1999.
13. Y. Ma, J. Košecká, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *submitted to IJCV*, 1999.
14. S. Maybank. *Theory of Reconstruction from Image Motion*. Springer-Verlag, 1993.
15. J. Milnor. *Morse Theory*. Annals of Mathematics Studies no. 51. Princeton University Press, 1969.
16. R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC press Inc., 1994.
17. S. Soatto and R. Brockett. Optimal and suboptimal structure from motion. *Proceedings of International Conference on Computer Vision*, to appear.
18. M. Spetsakis. Models of statistical visual motion estimation. *CVIPG: Image Understanding*, 60(3):300–312, November 1994.
19. T. Y. Tian, C. Tomasi, and D. Heeger. Comparison of approaches to egomotion computation. In *CVPR*, 1996.
20. J. Weng, T.S. Huang, and N. Ahuja. Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(5):451–475, 1989.
21. J. Weng, T.S. Huang, and N. Ahuja. *Motion and Structure from Image Sequences*. Springer Verlag, 1993.
22. J. Weng, T.S. Huang, and N. Ahuja. Optimal motion and structure estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):864–84, 1993.
23. Z. Zhang. Understanding the relationship between the optimization criteria in two-view motion analysis. In *Proceeding of International Conference on Computer Vision*, pages 772–77, Bombay, India, 1998.

Discussion

Kenichi Kanatani: You compare your method with other techniques, but in my view what you should really do is compare it with the theoretical accuracy bound, the lower bound beyond which accuracy can't be improved. For the problems you have described so far it is very easy to derive this bound.

Jana Košecká: Theoretical accuracy is usually expressed in terms of the Cramér-Rao bound, but there's an alternative way to look at it. If one bases the optimization on the epipolar constraint, it turns out that no matter what you do, half of the variance always gets absorbed by the structure. You can not do better than that — the error along the epipolar line gets absorbed by the structure, so you can only improve the error perpendicular to the epipolar line. One can even consider this as an alternative means of putting some lower bound on the estimates using these kind of techniques. Also, Weng, Huang and Ahuja [21] already did the comparision with the theoretical bound. Rather than repeating this analysis, we preferred to give a complementary viewpoint.