Cyclic optimization algorithms for simultaneous structure and motion recovery in computer vision

Seok Lee \textsuperscript{a}; Frank Chongwoo Park \textsuperscript{b}

\textsuperscript{a} Computing and Intelligence Lab, Samsung Advanced Institute of Technology, SEOUL, Korea
\textsuperscript{b} Mechanical and Aerospace Engineering Dept., Seoul National University, Seoul, Korea

First Published on: 28 January 2008

To cite this Article: Lee, Seok and Park, Frank Chongwoo (2008) ’Cyclic optimization algorithms for simultaneous structure and motion recovery in computer vision’, Engineering Optimization, 40:5, 403 — 419

To link to this article: DOI: 10.1080/03052150701804654

URL: http://dx.doi.org/10.1080/03052150701804654

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Cyclic optimization algorithms for simultaneous structure and motion recovery in computer vision

Seok Lee\textsuperscript{a} and Frank Chongwoo Park\textsuperscript{b}

\textsuperscript{a}Computing and Intelligence Lab, Samsung Advanced Institute of Technology, Gyeonggi-Do 446-712, Korea; \textsuperscript{b}Mechanical and Aerospace Engineering Dept. Seoul National University, Seoul 151-742, Korea

(Received 15 December 2006; final version received 21 September 2007)

With few exceptions, most previous approaches to the structure from motion (SFM) problem in computer vision have been based on a decoupling between motion and depth recovery, usually via the epipolar constraint. This article offers closed-form cyclic optimization algorithms for the simultaneous recovery of motion and depth in the discrete SFM problem. Cyclic coordinate descent (CCD) algorithms in which each stage admits closed-form solutions are developed for two widely used fitting criteria: the geometric error in one image, and the reprojection error criterion. As a by-product, analytic gradients that can be used in descent-based optimization methods are also obtained. The computational efficiency, statistical consistency, noise robustness, and accuracy of the algorithms are assessed via experiments with synthetic image data.

Keywords: computer vision; structure from motion; cyclic coordinate descent

1. Introduction

Extracting information about the three-dimensional shape and motion of a moving rigid object, typically from a time sequence of images obtained from one or more fixed cameras, is referred to in the vision literature as the structure from motion (SFM) problem, or alternatively as the shape and motion recovery problem. One of the most fundamental problems in computer vision, its origins can be traced to some of the earliest works in computer vision (\textit{e.g.} Ullman 1979, Longuet-Higgins 1981), and befittingly there exists an enormous literature on the subject (\textit{e.g.} Faugeras and Luong, Maybank 1993, Hartley and Zisserman 2000, 2001; see also the reviews by Aggarwal and Nandhakumar 1988, Weng \textit{et al.} 1993).

Because of the richness and complexity of the shape and motion recovery problem, as exemplified by the wide range of choices available for, for example, the image projection model, feature types, the number of available views, models of the environment, and the types of estimation algorithms used, as well as its immediate relevance due to recent advances in computing and video technology, it continues to be a fundamental and active topic of vision research.
The discrete version of the shape and motion recovery problem addressed in this article makes the following assumptions: a single fixed camera with a perspective projection model and, for each pair of consecutive images, the availability of image plane point correspondence information for a fixed number of feature points on a single rigid object. In this setting the shape and motion recovery problem can be formulated as attempting to determine the three-dimensional motion of the object, for example, its relative position and orientation displacement between consecutive images, together with the precise three-dimensional world coordinates for each feature point.

Since the seminal work of Longuet-Higgins (1981), most approaches to the shape and motion recovery problem have, depending on the type of available measurements, been framed as least-squares fitting problems subject to the epipolar constraint. From a practical perspective, the attractiveness of the epipolar approach seems to stem from the fact that shape and motion recovery can be decoupled. However, it has recently been pointed out that this decoupling can cause statistical bias in the translation estimates (Jepson and Heeger 1993, Kanatani 1993, Daniilidis and Spetsakis 1996) and more generally lead to statistically inconsistent estimators (Zhang and Tomasi 2002), increase the sensitivity to noise and quantization (Weng et al. 1993), and cause motion estimates to become unstable over long time sequences (McLauchlan and Murray 1995).

These observations have renewed interest in methods that do not explicitly rely on the epipolar constraint, but rather address the structure and motion recovery problem in its original coupled form. Prior to this recent awareness of consistency and bias issues, one of the first studies to simultaneously recover discrete structure and motion from two or more views was that of Szeliski and Kang (1994), who employed a Levenberg–Marquardt algorithm to minimize a weighted square distance between observed and predicted feature coordinates. The main disadvantage of this approach is that it is computationally much more intensive than epipolar methods; not only is the problem dimension much greater, but because the algorithms also iteratively optimize over the structure and motion parameters, they tend to converge more slowly. At a more subtle level, the perceived lack of geometric elegance vis-à-vis epipolar approaches also seems to have been a factor in the tendency to initially dismiss the Szeliski and Kang’s approach.

In this article it is shown that the simultaneous shape and motion recovery problem, as formulated by Szeliski and Kang (1994), can in fact admit efficient numerical solutions. The main contribution is a set of efficient algorithms for two widely used discrete SFM criteria: the coupled least-squares geometric error described by Szeliski and Kang (1994), and the re-projection error criterion. While both involve simultaneous optimization over structure and motion parameters, it turns out that the embedded conditional problems, i.e. optimizing with respect to a certain set of parameters while keeping the remains fixed and iterating this procedure over the entire set of parameters (these approaches are usually referred to in the optimization literature as cyclic optimization algorithms (Luenberger 1989)), lead to easily computed closed-form solutions. In particular, no line search procedure is required in the proposed algorithms. These solutions also produce, as a useful by-product, analytical gradient formulae that can be effectively used in more traditional non-linear optimization methods, such as that of Szeliski and Kang (1994), or any of the traditional descent methods based on line search, such as steepest descent, or quasi-Newton methods that approximate the inverse Hessian.

These advantages can ultimately be traced to the fact that the simultaneous shape and motion recovery problem as formulated here leads to linear objective functions on the rotation group $SO(3)$ of the form $\text{Tr}(AR)$, where $R \in SO(3)$, and $A \in \mathbb{R}^{3 \times 3}$ is a given non-singular matrix, which have analytical formulae for the minima and maxima. In contrast, epipolar constraint-based approaches typically require the minimization of a quadratic function on $SO(3)$ of the form $\text{Tr}(RAR^TB)$, where $A, B \in \mathbb{R}^{3 \times 3}$ are given symmetric positive semi-definite matrices.
This article demonstrates, via experiments with synthetic images, that these algorithms are not only computationally efficient, but also statistically consistent, and lead to more robust structure and motion estimates than those obtained with the widely used linear epipolar-based method described by Maybank (1993). After formulating the various criteria in section 2, the respective closed-form algorithms are derived in section 3. Experimental results are presented in section 4.

2. Problem formulation

2.1. Notation and Definitions

A perspective projection model is assumed, in which the camera image plane is a unit distance away from its lens centre, and the $z$-axis of the camera frame is aligned along the optical axis (Figure 1). Thus the image plane, denoted $I$, corresponds to the $x$–$y$ plane.

If a point $x$ on a rigid body is projected to the point $p$ in the image plane via perspective projection, one has the relation $p = x/\lambda$, where $\lambda$ is the $z$-component of $x$; $p$ thus is always of the form $(p_x, p_y, 1)$. Under a finite rigid-body displacement, the point $x$ is displaced to $y = Rx + t$, where $R \in SO(3)$ is a rotation matrix and $t \in \mathbb{R}^3$ is a translation vector. Denoting the $z$-component of $y$ by $\gamma$, and letting $q$ be the perspective projection of $y$ onto the image plane ($q$, like $p$, is always of the form $(q_x, q_y, 1)$), one has as before $q = y/\gamma$. Therefore the relation $y = Rx + t$ can be written equivalently as

$$q\gamma = Rp\lambda + t.$$  

(1)

Observe that, for given values of $p$ and $q$, both $(R, t, \lambda, \gamma)$ and $(R, ct, c\lambda, c\gamma)$ will satisfy the above equations for any non-zero scalar $c$. This implies the well-known fact that unique recovery of $(R, t, \lambda, \gamma)$ is not possible; a one-parameter family of solutions will exist. (Put another way, ‘any object moving in front of a camera will be indistinguishable from one that is twice as big, twice

![Figure 1. The perspective projection camera model.](image-url)
as far, and moving twice as fast’.) Following standard practice, the scale ambiguity is eliminated from the formulation by requiring $\|t\| = 1$, and treating $\|t\| = 0$ as a special case. As a reminder, it is noted that the 
\textit{discrete epipolar constraint} (Longuet–Higgins 1981) is expressed as

$$p^T R[t] q = 0$$

(2)

where it is common practice to eliminate the scale ambiguity from the formulation by requiring $\|t\| = 1$.

\subsection*{2.2. Fitting criteria}

Denote the $n$ image point correspondence measurement pairs by $\{(p_1, q_1), \ldots, (p_n, q_n)\}$, and the matrices $P, Q, T \in \mathbb{R}^{3 \times n}$ by

$$P = \begin{bmatrix} p_1 & \ldots & p_n \end{bmatrix}$$

(3)

$$Q = \begin{bmatrix} q_1 & \ldots & q_n \end{bmatrix}$$

(4)

$$T = \begin{bmatrix} t & \ldots & t \end{bmatrix}$$

(5)

where $t \in \mathbb{R}^3$ denotes the translational component of the rigid-body displacement. For the reasons cited earlier, the constraint $\|t\| = 1$ is imposed. $\Lambda$ and $\Gamma$ are further defined as the following $n \times n$ diagonal depth matrices:

$$\Lambda = \text{Diag}\{\lambda_1, \ldots, \lambda_n\}$$

(6)

$$\Gamma = \text{Diag}\{\gamma_1, \ldots, \gamma_n\}.$$  

(7)

Finally, $\| \cdot \|$ denotes the standard Euclidean norm in the case of vectors, \textit{i.e.} $\|x\|^2 = x^T x$ for $x \in \mathbb{R}^n$, and the Frobenius norm in the case of matrices, \textit{i.e.} $\|A\|^2 = \text{Tr}(A^T A)$.

\subsection*{2.2.1. Geometric distance error in one image}

For the least-squares error in image space, the objective function is given by

$$\min J(R, t, \lambda, \gamma) = \sum_{i=1}^{n} \|q_i - \frac{1}{\gamma_i}(\lambda_i R p_i + t)\|^2$$

(8)

which can be rewritten in matrix notation as

$$\min J(R, T, \Lambda, \Gamma) = \|Q - (R P \Lambda + T) \Gamma^{-1}\|^2.$$  

(9)

The underlying physical assumption behind this criterion is that noise affects only the $q_i$ measurements, and that the noise is of additive isotropic type in the image space. This is consistent with the usual modelling of pixel noise in images.
2.2.2. Reprojection error criterion

To formulate the reprojection error criterion, first define

$$\hat{P} = [\hat{p}_1 \cdots \hat{p}_n] \in \mathbb{R}^{3 \times n}$$  \hspace{1cm} (10)

$$\hat{Q} = [\hat{q}_1 \cdots \hat{q}_n] \in \mathbb{R}^{3 \times n}$$  \hspace{1cm} (11)

where $\hat{p}_i, \hat{q}_i \in I$ are optimization parameters to be determined. The objective function is given by

$$\min J(R, t, \hat{P}, \lambda, \hat{Q}, \gamma) = \sum_{i=1}^{n} \| p_i - \hat{p}_i \|^2 + \| q_i - \hat{q}_i \|^2$$  \hspace{1cm} (12)

subject to the constraint $\hat{q}_i \gamma_i = R \hat{p}_i \lambda_i + t$, $i = 1, \ldots, n$. The above can be rewritten in matrix notation as

$$\min J(R, T, \hat{P}, \Lambda, \hat{Q}, \Gamma) = \| P - \hat{P} \|^2 + \| Q - \hat{Q} \|^2$$  \hspace{1cm} (13)

subject to $\hat{Q} \Gamma = R \hat{P} \Lambda + T$. The underlying physical assumption behind this criterion is that noise affects both the $p_i$ and $q_i$ measurements, and that the noise is of the additive isotropic type in the image space.

3. Cyclic optimization algorithms

Given an objective function $L(x), x \in \mathbb{R}^n$, the cyclic coordinate descent (CCD) method optimizes $L(x)$ by sequentially minimizing with respect to each of the components $x_i$, i.e. $L(x)$ is first minimized with respect to $x_1$ while keeping the remaining $x_i$s fixed, followed by $x_2, x_3$, and so forth. Variations of this method are collectively referred to as coordinate descent methods (Luenberger 1989). While it is difficult to make precise mathematical statements about the convergence behaviour of CCD algorithms, it is generally accepted that they have poorer convergence than traditional descent methods; a general rule of thumb is $n - 1$ iterations for one iteration of steepest descent.

CCD methods become attractive when the coupling between variables becomes weak, and when conditional minimizations with respect to each of the coordinates is particularly simple. It should be noted that schemes for improving the convergence of CCD methods have been proposed in the literature, for example the Gauss–Southwell method (Luenberger 1989) and successive over-relaxation schemes (Briggs 2000). While not identified as such, CCD algorithms have appeared in various forms throughout the vision literature (e.g. Soatto and Brockett 2000).

3.1. Convergence analysis

For the proposed objective functions, the CCD method involves cyclically optimizing over $SE(3)$ (which is further partitioned into $SO(3)$ and $\mathbb{R}^3$), $\mathbb{R}^n$, and $\mathbb{R}^n$. Convergence of the CCD method can usually be established following Zangwill’s general approach (Zangwill 1969) for the convergence analysis of algorithms (also described by Luenberger 1989). The convergence of CCD algorithms for quadratic functions on $SE(3) \cong SO(3) \times \mathbb{R}^3$ is shown by Gwak et al. (2004), and it is straightforward to extend these results to show convergence of the proposed CCD algorithms – only a distance metric on the search space needs to be constructed. Moreover, as shown below, for the proposed objective functions each optimization over a component space produces a unique minimum that can be characterized analytically. Not only is this attractive from a computational
THEOREM 1 (Global Convergence Theorem) Let \( A \) be an algorithm on \( X \), and suppose that, given \( x_0 \), the sequence \( \{x\}_{k=0}^{\infty} \) is generated satisfying \( x_{k+1} \in A(x_k) \). Let a solution set \( \Gamma \in X \) be given, and suppose that

(i) all points \( x_k \) are contained in a compact set \( S \in X \)
(ii) there is a continuous function \( Z \) on \( X \) such that
   (a) if \( x \notin \Gamma \), then \( Z(y) < Z(x) \) for all \( y \in A(x) \)
   (b) if \( x \in \Gamma \), then \( Z(y) \leq Z(x) \) for all \( y \in A(x) \)
(iii) the mapping \( A \) is closed at points outside \( \Gamma \).

Then the limit of any convergent subsequence of \( x_k \) is a solution.

An important corollary of the Global Convergence Theorem is that if, under the conditions of the theorem, \( \Gamma \) consists of a single point \( \bar{x} \), the sequence \( x_k \) converges to \( \bar{x} \). With the above preliminaries the objective function \( L : SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \), denoted \( L(R, t, \alpha, \beta) \), is considered. The following mappings are first defined.

DEFINITION 1 (i) \( C^1 : SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \times so(3) \) is a point-to-point mapping defined by

\[
C^1(R, t, \alpha, \beta) = (R, t, \alpha, \beta, \Omega)
\]

where \( \Omega \) corresponds to the search direction along the \( SO(3) \) component at \( (R, t, \alpha, \beta) \), expressed as an element of \( so(3) \) by left-translation.

(ii) \( S^1 \) is a mapping defined on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \times so(3) \) that assigns to each point \( (R, t, \alpha, \beta, \Omega) \in SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \times so(3) \) the following set in \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \):

\[
S^1(R, t, \alpha, \beta, \Omega) = \left\{ (R_+, t, \alpha, \beta) | R_+ = Re^{\Omega t} (0 \leq t \leq 2\pi), \left. L(R_+, t, \alpha, \beta) = \min_{0 \leq t \leq 2\pi} L(Re^{\Omega t}, t, \alpha, \beta) \right\}.
\]

\( S^1 \) is a set-valued mapping since in some cases there may be many values of \( R_+ \) yielding the minimum (with \( b, \alpha, \beta \) fixed).

(iii) \( S^2 \) is a mapping defined on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \) that assigns to each point \( (R, t, \alpha, \beta) \in SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \) the following set in \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \):

\[
S^2(R, t, \alpha, \beta) = \left\{ (R, t, \alpha, \beta) | \nabla_b L = 0 \text{ and } \frac{\partial^2 L}{\partial b^2} > 0 \right\}.
\]

\( S^2 \) is a set-valued mapping since in some cases there may be many values of \( b_+ \) yielding the minimum (with \( R, \alpha, \beta \) fixed).

(iv) \( S^3 \) is a mapping defined on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \) that assigns to each point \( (R, t, \alpha, \beta) \in SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \) the following set in \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \):

\[
S^3(R, t, \alpha, \beta) = \left\{ (R, t, \alpha_+, \beta) | \nabla_\alpha L = 0 \text{ and } \frac{\partial^2 L}{\partial \alpha^2} > 0 \right\}.
\]

\( S^3 \) is a set-valued mapping since in some cases there may be many values of \( \alpha_+ \) yielding the minimum (with \( R, t, \beta \) fixed).
(v) \( S^4 \) is a mapping defined on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \) that assigns to each point \((R, t, \alpha, \beta) \in SE(3) \times \mathbb{R}^n \times \mathbb{R}^n\) the following set in \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n\):

\[
S^4(R, t, \alpha, \beta) = \left\{ (R, t, \alpha, \beta_+) | \nabla_{\beta} L = 0 \text{ and } \frac{\partial^2 L}{\partial \beta^2} > 0 \right\}.
\]

(18)

Again, \( S^4 \) is a set-valued mapping since in some cases there may be many values of \( \beta_+ \) yielding the minimum (with \( R, t, \) and \( \alpha \) fixed).

**Theorem 2** (Global Convergence of Cyclic Coordinate Descent on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \)) Let \( SE(3) \) have a left-invariant metric space structure, and consider a differentiable objective function \( L(R, t, \alpha, \beta) \) on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \). Define the solution set

\[
\Gamma = \{(R, t, \alpha, \beta) | \nabla_R L = \nabla_b L = \nabla_a L = \nabla_\beta L = 0\}
\]

(19)

where \( \nabla_R \) denotes the gradient of \( L \) (regarding other variables as fixed) with respect to the natural Riemannian metric on \( SO(3) \). Let \( A \) be an algorithm on \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \) defined as the composition of five maps \( A = S^4 \circ S^3 \circ S^2 \circ S^1 \circ C^1 \), where \( C^1 \) is assumed to be continuous, and \( S^1, S^2, S^3, \) and \( S^4 \) are assumed to yield unique minimum points. \( A \) is also assumed to be restricted to points on a compact set of \( SE(3) \times \mathbb{R}^n \times \mathbb{R}^n \). Then the algorithm \( A \) is globally convergent.

**Proof**  \( S^1 \), which is just a line search over the doubly infinite line, is clearly closed. \( S^2, S^3, \) and \( S^4 \) are continuous point-to-point mappings, and since points are restricted to a compact set, the composite map \( A \) is also closed. The objective function \( L(R, t, \alpha, \beta) \) also serves as a continuous descent function for \( A \) with respect to \( \Gamma \); this is because the mappings \( S^2, S^3, \) and \( S^4 \) yield a decrease, and because \( S^1 \) yields a decrease or, by the uniqueness assumption, it cannot change position. The theorem now follows by invoking the general Global Convergence Theorem.

3.2. Linear objective functions on \( SO(3) \)

It is well-known (Wahba 1965, Nadas 1978, Markley 1988, Brockett 1989, Li et al. 1998, Park et al. 2000) that the minimizer and maximizer of the function

\[
J(R) = \text{Tr}(AR),
\]

(20)

where \( R \in SO(3) \) and \( A \in \mathbb{R}^{3 \times 3} \) is a given arbitrary non-singular matrix, are respectively given by

\[
R_{\text{max}} = \begin{cases} V \text{Diag}^[1, 1, 1] \ U^T, & \text{det}(A) > 0 \\ V \text{Diag}^[1, 1, -1] \ U^T, & \text{det}(A) < 0 \end{cases}
\]

(21)

\[
R_{\text{min}} = \begin{cases} V \text{Diag}[-1, -1, 1] \ U^T, & \text{det}(A) > 0 \\ V \text{Diag}[-1, -1, -1] \ U^T, & \text{det}(A) < 0 \end{cases}
\]

(22)

where \( U, V \in SO(3) \) are obtained from the singular value decomposition \( A = U \Sigma V^T \) (see Cohen (1996) for computational details).
3.3. Least squares geometric distance error

To simplify notation, the inverse depth parameter \( \eta_i = \gamma_i^{-1}, i = 1, \ldots, n \), and \( N = \text{Diag}\{\eta_1, \ldots, \eta_n\} \in \mathbb{R}^{n \times n} \) are introduced. The objective function is then given by

\[
\min J(R, T, \Lambda, \Gamma) = \| Q\Gamma - RP\Lambda - T \|^2. \tag{23}
\]

Minimization is now performed repeatedly over each of the variables, keeping the other variables fixed.

- **Minimize with respect to** \( R \) Holding the other variables constant, the linear function \( \text{Tr}(AR) \) is minimized with respect to \( R \in SO(3) \), where

\[
A = P\Lambda N(NT^T - Q^T). \tag{24}
\]

- **Minimize with respect to** \( t \) Holding the other variables constant, minimize

\[
J(T) = \text{Tr}[TN^2T^T + 2(Q - RP\Lambda N)NT^T] \tag{25}
\]

which can be further simplified to

\[
J(t) = (\eta^T \eta)t^T t - 2c^T t \tag{26}
\]

where

\[
c^T = [1 \cdots 1] N(Q - RP\Lambda N)^T. \tag{27}
\]

Since it can be assumed that \( \eta^T \eta > 0 \), the unique minimum is given by \( t^* = c/\eta^T \eta \). If the constraint \( \|t\| = 1 \) is further imposed, the minimum can be expressed as \( t^* = c/\|c\| \).

- **Minimize with respect to** \( \Lambda \) Holding the other variables constant, minimize

\[
J(\Lambda) = \text{Tr}[\Lambda G\Lambda - 2H\Lambda] \tag{28}
\]

where \( G = (PN)^T(PN) \) and \( H = N(Q - TN)^T RP \). The above can be simplified to

\[
J(\lambda) = \sum_{i=1}^{n} \eta_i^2 \|p_i\|^2 \lambda_i^2 - 2h_{ii} \lambda_i \tag{29}
\]

where \( h_{ii} \) denotes the \( i \)th diagonal entry of \( H \). Note that as long as each \( \eta_i^2 \|p_i\|^2 > 0 \), there exists a unique minimum given by \( \lambda_i = (\eta_i \|p_i\|)^{-2}h_{ii}, i = 1, \ldots, n \). Clearly situations in which \( \|p_i\| = 0 \) are highly unlikely to occur, while \( \eta_i = 0 \) implies an infinite depth estimate \( \gamma_i \) for feature \( i \), an equally unlikely event.

- **Minimization with respect to** \( N \) Holding the other variables constant, first minimize

\[
J(\eta) = \text{Tr}[N^TGN - 2HN] \tag{30}
\]

where

\[
G = (P\Lambda + R^T T)^T(P\Lambda + R^T T) \quad \text{and} \quad H = Q^T(RP\Lambda + T). \tag{31}
\]

As long as \( p_i \lambda_i + R^T t \neq 0, g_{ii} \) is guaranteed to be greater than zero. It can be seen that \( p_i \lambda_i + R^T t = 0 \) will be zero only if the body frame displacement as seen from the final frame, \(-R^T t\), is equal to \( x_i = p_i \lambda_i \). The unique minimum \( \eta_i \) is given by \( \eta_i^* = h_{ii}/g_{ii} \), so that \( \gamma_i^* = g_{ii}/h_{ii} \).
3.4. Reprojection error criterion

Using matrix notation and substituting the constraints into the objective function, one obtains

\[
\min J(R, T, \hat{P}, \lambda, \hat{Q}, \gamma) = \|P - \hat{P}\|^2 + \|Q - R\hat{P}\Lambda N - TN\|^2.
\]  

(32)

Minimization over each of the variables is now performed separately.

- **Minimize with respect to** \(R\) **Holding the other variables constant**, minimize \(\text{Tr}(AR)\) over \(R \in SO(3)\), where

\[
A = \hat{P}\Lambda N(NT^T - QT).
\]  

(33)

Observe that this is the same as the corresponding \(SO(3)\) objective function for the image space least-squares error, but with \(P\) now replaced by \(\hat{P}\).

- **Minimize with respect to** \(t\) **Holding the other variables constant**, minimize

\[
J(t) = (\eta^T \eta)t^T t - 2c^T t
\]  

(34)

where

\[
c^T = [1 \cdots 1] N(Q - R\hat{P}\Lambda N)^T.
\]  

(35)

Like the image space least-squares error, the minimum is given by \(t^* = c/\eta^T \eta\) in the general case, and \(t^* = c/\|c\|\) if the additional constraint \(\|t\| = 1\) is imposed.

- **Minimize with respect to** \(\Lambda\) **Holding the other variables constant**, minimize

\[
J(\lambda) = \sum_{i=1}^{n} \eta_i^2 p_i \|\hat{\pi}_i\|^2 \lambda_i^2 - 2h_{ii} \lambda_i
\]  

(36)

where \(h_{ii}\) denotes the \(i\)th diagonal entry of \(H = N(Q - TN)^T R\hat{P}\). Like the image space least-squares error, the minimum is given by \(\lambda_i = h_{ii}(\eta_i/\|\hat{\pi}_i\|)^{-2}, i = 1, \ldots, n\).

- **Minimization with respect to** \(N\) **Holding the other variables constant**, minimize the following with respect to \(\eta\):

\[
J(\eta) = \sum_{i=1}^{n} \eta_i^2 g_{ii} - 2h_{ii} \eta_i
\]  

(37)

where \(g_{ii}\) and \(h_{ii}\) are the \(i\)th diagonal entries of \(G = (\hat{P}\Lambda + R^T T)(\hat{P}\Lambda + R^T T)\) and \(H = \hat{Q}^T (R\hat{P}\Lambda + T)\), respectively. Like the image space least-squares error, the minimum \(\eta_i\) is given by \(\eta_i^* = h_{ii}/g_{ii}\), so that \(\gamma_i^* = g_{ii}/h_{ii}\).

- **Minimization with respect to** \(\hat{P}\) **Holding the other variables constant**, minimize

\[
J(\hat{P}) = \text{Tr}[\hat{P}[I + (\Lambda N)(\Lambda N)^T]\hat{P}^T + 2(\Lambda N^2 T^T R - \Lambda N Q^T - P^T)\hat{P}].
\]  

(38)

Define \(H = P + R^T (Q - TN)\Lambda D\eta \in \mathfrak{H}_{3 \times n}\), and represent the columns of \(H\) by

\[
H = [h_1 \cdots h_n]
\]  

(40)

where each component of \(h_i \in \mathfrak{H}_3\) is given by \(h_i = (h_{ix}, h_{iy}, h_{iz})\). Further denote the components of each \(\hat{p}_i \in I\) by \(\hat{p}_i = (\hat{p}_{ix}, \hat{p}_{iy}, 1)\). The objective function can then be expanded to

\[
J(\hat{P}) = \sum_{i=1}^{n} (1 + \lambda_i^2 \eta_i^2)(\hat{p}_{ix}^2 + \hat{p}_{iy}^2) - 2(h_{ix} \hat{p}_{ix} + h_{iy} \hat{p}_{iy})
\]  

(41)
and the minimum is given by
\[
\begin{bmatrix}
\hat{p}_{ix}^* \\
\hat{p}_{iy}^
\end{bmatrix} = \begin{bmatrix}
 h_{ix} \\
 h_{iy}
\end{bmatrix} \cdot \frac{1}{1 + \lambda_i^2 \eta_i^2}.
\] (42)

4. Experimental results

This section investigates the performance of the proposed algorithms with synthetic image data. The focus is on two aspects: (i) the noise sensitivity and statistical consistency of the CCD algorithms compared with standard epipolar-based methods, and (ii) the computational performance of the CCD algorithms compared with a non-linear iterative optimization algorithm of the type proposed by Szeliski and Kang (1994).

As reported in the recent literature, certain epipolar-constraint-based methods lead to statistically inconsistent estimators as a result of transformations that produce non-isotropic noise models. It is first verified, via simulations, that the proposed algorithm does in fact yield statistically consistent estimators. Specifically, the way in which bias and variance errors in the motion (rotation and translation) and shape (depth) are affected not only by the noise levels, but also by the number of feature points used is examined.

The widely used singular value decomposition-based linear algorithm described by Maybank (1993) is used. For the epipolar estimator, and the Levenberg–Marquardt algorithm described by Szeliski and Kang (1994) is used for the iterative non-linear optimizer. The simulations are performed on a 3.0 GHz Pentium 4 personal computer with 1 gigabyte memory. The algorithms are implemented in Visual C++, and IMSL numerical routines are called from the same programming environment.

Unless specified otherwise, the simulations are performed under the following conditions: 50 points are randomly generated from a uniform distribution over a cube with range \([0, 120] \times [0, 120] \times [60, 120]\). Each point is first rotated about the y-axis by 0.5 rad, followed by a translation of 10 units along the z-axis, and then projected to a 256 \(\times\) 256 pixel image plane of focal length 1. A zero-mean two-pixel standard deviation Gaussian noise is added to each displaced feature point. The convergence criteria for the optimization algorithms are set such that the absolute value of the objective function between the present and previous iterations differs by less than \(10^{-6}\). Each simulation trial is repeated 100 times, each time with a newly generated set of 50 feature points, and the mean and standard deviation obtained over the 100 trials is used to assess overall algorithm performance.

4.1. Noise sensitivity analysis

Figures 2 and 3 compare the noise sensitivity of the two CCD algorithms with the linear epipolar based method. The x-axis represents the standard deviation of the Gaussian noise (in pixels), while the y-axis represents the difference between the actual and estimated optimal motion. The y-axis in Figure 2 represents the average error obtained over 100 simulation trials, and the y-axis in Figure 3 represents the associated standard deviation. The rotation and translation errors are measured as \(\| \log(R_{\text{real}}^T R_{\text{estimated}}) \|\) and \(\|t_{\text{real}} - t_{\text{estimated}}\|\), respectively where \(\log(\cdot)\) denotes the matrix logarithm and \(\|\cdot\|\) denotes the Euclidean two-norm (and its induced norm for matrices). As is evident from the figures, the errors are greatest for the epipolar method, followed by the reprojection error, and the least-squares geometric distance error in one image. In particular, the noise sensitivity of the epipolar-based method is considerably larger than that for either of the two proposed algorithms.
4.2. Statistical consistency analysis

As noted by Zhang and Tomasi (2002), a lack of statistical consistency typically indicates an inappropriate underlying noise model, so that increasing the number of feature point measurements fails to increase the accuracy of the estimates. This simulation examines how the error of the motion and depth estimates behaves as the number of feature points is gradually increased from 10 to 100.
Figures 4 and 5 show the motion and depth errors as a function of the number of feature points. The motion errors are measured in terms of the previously described rotation and translation errors, while the depth errors are described in terms of the mean square error between $\lambda_i$ and $\gamma_i$. As is evident from the figures, the estimated values show a clear tendency to approach the actual values as the number of feature points is increased. The variance also clearly tends to zero with an increase in the number of feature points.
Figure 4. Motion estimation error as a function of number of feature points.

Figure 6 plots the root mean square error of the depth estimates as a function of the number of feature points for the linear epipolar method as well as for the two proposed algorithms. As is evident from the figure, the epipolar-based method shows only a slight decrease in depth error relative to the other two methods, even where the number of feature points is increased from 10 to 100. In terms of absolute accuracy, the reprojection error clearly performs best, while the least-squares and epipolar methods perform similarly as the number of feature points approaches 100.
4.3. Convergence performance

This section compares the computational performance of the proposed CCD algorithm with that of a Levenberg–Marquardt non-linear optimization algorithm. Szeliski and Kang (1994) applied a Levenberg–Marquardt algorithm to the objective function (8), with analytical formulae for the gradients provided. Figure 7 and Table 1 show the computation times for each algorithm as a function of the number of the feature points, which are varied from 10 to 200. Given that
the Levenberg–Marquardt algorithm is highly efficient, the proposed CCD algorithm performs quite well, particularly as the number of feature points is increased. The convergence times are comparable up to 50 feature points, but diverge significantly as the number of feature points is increased further; in the case of 200 feature points, the convergence rate for the proposed CCD algorithm is approximately 10 times faster than that of the Levenberg–Marquardt algorithm.
5. Conclusion

A pair of computationally efficient CCD algorithms for simultaneously estimating the motion and depth parameters in the discrete structure from motion problem have been presented. Compared with some existing epipolar-based methods, the two algorithms lead to statistically consistent estimators, and are also more robust to noise. This result is not altogether surprising given that one is now directly solving the original coupled form of the structure from motion problem, whereas epipolar methods implicitly transform the original noise model into a potentially non-isotropic (and physically invalid) form.

Extensive simulation studies also suggest that the CCD algorithm is computationally more efficient than methods which attempt to minimize the criterion via the Levenberg–Marquardt method, with significantly greater performance improvements as the number of feature points is increased.

Acknowledgements

This research was supported in part by the Korea Research Foundation under the 2001 visiting faculty program, and by IAMD-SNU.

References


