A New Geometrical Interpretation of Trilinear Constraints

Xiaolin Feng   Jean-Yves Bouguet   Pietro Perona
MC 136-93, Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
Email: {xlfeng,bouguetj,perona}@vision.caltech.edu

Abstract

We give a new geometrical interpretation of the well-known algebraic trilinear constraints used in motion analysis from three views observation. We show that those algebraic equations correspond to depth errors appropriately weighted by a function of the relative reliability of the corresponding measurements. Therefore directly minimizing the algebraic trilinear equations, in the least squares sense, is a strategy that works well for estimating motion. In addition, we propose a new scheme for recovering the scale factor for motion estimation that is very insensitive to input noise. All our theoretical statements are supported by experimental results.

Keywords: trilinear constraints, geometry, weighted depth error.

1 Introduction

Motion analysis has attracted lots of attention in the computational vision community. In recent years, multiple views analysis has become a main stream of the research. The trilinear constraints are now considered the fundamental equations for motion analysis from three views. The associated tensor (called trifocal tensor) has already been extensively studied and its properties are described in a whole body of literature [9, 11, 6, 2, 14, 8].

However, as algebraic equations, the trilinear constraints do not have obvious geometrical interpretations. In a noiseless case, the algebraic equations are exactly satisfied for all the scene points. In presence of noise however, the residual error generated by the constraint equations may vary dramatically from point to point. In that sense, each point carries its reliability that should be appropriately accounted for when building the cost function to minimize. The philosophy is similar to applying appropriate weights to the different observation points. If the weights are not chosen properly, it is very likely that the best estimation will not be achieved. In that sense, one may think that pure algebraic constraints, taken in a least squares fashion, are not reliable to use. However in practice, trilinear constraints do give satisfying estimation results without additional weighting coefficients. The fundamental goal of this paper is to provide a geometrical interpretation of the trilinear constraints that naturally explains this fact.

We revisit the trilinear constraints from the 3D structure depth viewpoint and show that they are naturally weighted constraints enforcing depth equivalence of the points in space. A clear geometrical interpretation of the weight function is presented. We show theoretically as well as experimentally that standard trilinear algebraic cost is almost equivalent to the optimal depth matching cost. The extension of the work to the uncalibrated sequence is also given. This provides a practical explanation of why pre-normalizing the point coordinates [5, 6] is useful for better estimation performances using trilinear constraints. In addition, we propose a reliable scale propagation scheme that is very insensitive to the noise in the point coordinates.

The paper is organized as follows. Section 2 and 3 give the background and bring up the main problem that this paper tries to solve. The geometrical interpretation follows in the next section. In section 5 the scale propagation scheme is proposed. Experiments are reported in section 6. We briefly extend the work to the uncalibrated case in section 7 and followed by the conclusion.

2 Notation and Background

We adopt the tensorial notations for trilinearities as the literature in multiple views analysis usually does. The coordinates of a point are specified with superscript and named contravariant vector, i.e.,$\mathbf{u} = [u^1, u^2, \ldots]$. An element in the dual space is called a covariant vector and represented by subscripts, i.e., $\mathbf{l} = [l_1, l_2, \ldots]$. In a similar manner, a notation a^i_j denotes a matrix, the triply indexed quantities $\mathcal{T}_i^{jk}$ are named trifocal tensors. We use the usual covariant-contravariant summation convention: any index repeated in covariant and contravariant forms implies a summation over the range of index values, i.e., $u^i l_i = u^1 l_1 + u^2 l_2 + \ldots + u^n l_n$.

The three dimensional space is represented as the 3D projective space $\mathcal{P}^3$. A point $\mathbf{P}$ in Euclidean space with coordinates $[X, Y, Z]$ is represented by homogeneous 4-vector $\mathbf{x} = [X, Y, Z, 1]^T$. Similarly, the image plane is regarded as the 2D projective space $\mathcal{P}^2$

and points on the image are represented by homogeneous 3-vector coordinates $\mathbf{u} = [u^1, u^2, 1]^T$. The homogeneous property in the projective space (see [1]) means that multiplying the vector with any nonzero scale factor represents the same vector in the space. Therefore, without loss of generality, we enforce the last element of the vector to be 1 in both $\mathcal{P}^3$ and $\mathcal{P}^2$ so that the other elements exactly represent coordinates in Euclidean space. Nevertheless, for notation convenience, we will still keep a three vector parameterization for points on the image plane: $\mathbf{u} = [u^1, u^2, u^3 = 1]^T$,

In projective space, the perspective projection of points in $\mathcal{P}^3$ onto the image plane $\mathcal{P}^2$ can be described as a linear transformation by a 3×4 matrix M:

$$\lambda \mathbf{u} = M\mathbf{x} = Kg\mathbf{x} \tag{1}$$

where M is decomposed into a product of a 3×3 camera matrix K (an upper diagonal matrix containing the intrinsic camera parameters) and a 3×4 transformation matrix g that represents the rigid body transformation between the world frame and the camera frame. This transformation may be written as $g = [R, T]$ where R is a rotation matrix, and T a translation vector. Observe that the scalar λ is the Euclidean depth of $\mathbf{P}$ if both vectors $\mathbf{u}$ and $\mathbf{x}$ have their last coordinates normalized to 1.

Let us consider the projective images of 3 views, and let M, $M^{'}$ and $M^{''}$ be the associated projection matrices. A point $\mathbf{P}$ of coordinates $\mathbf{x}$ in space $\mathcal{P}^3$ will be observed on the three views at positions $\lambda\mathbf{u} = M\mathbf{x}, \lambda'\mathbf{u}' = M'\mathbf{x}$ and $\lambda''\mathbf{u}'' = M''\mathbf{x}$. It is well known that in an uncalibrated image sequence if we allow the camera matrices to be arbitrary, the scene can only be reconstructed up to a 3D projective transformation. Therefore, without loss of generality, we may assume $M = [I, 0]$ (as in [6]). In the calibrated case, that assumption corresponds to choosing the first camera reference frame as world frame. In addition, for the sake of simplicity, we denote the components of M' by $M' = [\mathbf{a}, \mathbf{v}'] = [a^i_j, {v'}^i]$ ($\mathbf{a}$ is the 3×3 left minor and $\mathbf{v}'$ is the fourth column of M') and $M'' = [\mathbf{b}, \mathbf{v}''] = [b^i_j, {v''}^i]$ respectively.

From the image projections of a point $\mathbf{P}$ onto three views $\mathbf{u}, \mathbf{u}', \mathbf{u}''$, one derives a total of 9 trilinear constraints corresponding to all possible choices of (i, j, l, m) [6, 11, 2]:

$$\begin{aligned} \mathcal{E}1^{ijlm} &= u^k L^{ijlm}_k = 0 \\ L^{ijlm}_k &= {u'}^i {u''}^m \mathcal{T}^{jl}_k - {u'}^j {u''}^m \mathcal{T}^{il}_k \\ &\quad - {u'}^i {u''}^l \mathcal{T}^{jm}_k + {u'}^j {u''}^l \mathcal{T}^{im}_k \\ &1 \le i < j \le 3 \qquad 1 \le l < m \le 3 \end{aligned} \tag{2}$$

where the trifocal tensor $\mathcal{T}^{jk}_i = a^j_i {v''}^k - {v'}^j b^k_i$. The trilinear cost functions $\mathcal{E}1^{ijlm}$ and the intermediate terms L^{ijlm}_k are conveniently defined here for future address.

3 Normalization

In real image sequences, due to noise on the image coordinate measurements, Eq.(2) will not be exactly satisfied for all the points. The state-of-art techniques to minimize the residue errors are taken at two steps: estimate the trifocal tensors $\mathcal{T}^{jk}_i$ first and then retrieve M', M'' from the tensors. The 27 coefficients of trifocal tensors are not arbitrary but satisfying a set of algebraic and geometric constraints. These constraints allow to parameterize the tensors with minimal 18 parameters [3, 8, 14, 4]. Also, M', M'' belong to a family of homographies spanned with 3 degree of freedom and cannot be uniquely determined from the tensors [10, 13], therefore we have to enforce certain constraints to retrieve the projection matrices M', M''. The properties of these quantities and details of these constraints are not the main concern in this paper, we recommend the readers to check the references if interested. Here we generally describe the constraints for $\mathcal{T}^{jk}_i$ and M', M'' as $S_\mathcal{T}$ and S_M respectively. The standard linear method to identify the unknowns is [6, 4, 3]:

$$\begin{aligned} \{\mathcal{T}^{jk}_i\}^* &= arg \min_{\mathcal{T}^{jk}_i \in S_\mathcal{T}} \sum_{\mathbf{P}} \sum_{ijlm} (\mathcal{E}1^{ijlm})^2 \\ &\longmapsto \{M', M''\}^* \in S_M \end{aligned} \tag{3}$$

The term $\sum_\mathbf{P}$ means sum over all the points $\mathbf{P}$. Most of the quantities appearing in this paper are functions of the points $\mathbf{P}$, we don't bother to specify it in their expressions. Obviously $\mathcal{E}1$ is linear with respect to $\mathcal{T}^{jk}_i$ and has simpler form. Such linear algorithm achieved satisfying results. However, there is a basic question that has not yet been addressed in the literature: normalization. It's well known that $\mathcal{E}1$ is not a normalized cost function for all the points $\mathbf{P}$, its geometrical interpretation has yet not been established. A geometrically meaningful cost function is the error measured on the image plane after reprojection [14, 3]: Since $L^{ijlm}_k, k = 1, 2, 3$ with any admissible choice of $(ijlm)$ is actually a projection of a line which goes through the point $\mathbf{P}$ in space onto the view1 image plane, the cost function will be the algebraic distance between $\mathbf{u}$ and this line:

$$\mathcal{E}2^{ijlm} = \frac{\mathcal{E}1^{ijlm}}{\sqrt{(L^{ijlm}_1)^2 + (L^{ijlm}_2)^2}} \tag{4}$$

On the image plane, $\mathcal{E}2$ is normalized and equally fair to all the points. However, unfortunately it is complicated and not linear as $\mathcal{E}1$ is. We notice that $\mathcal{E}1$ differs from $\mathcal{E}2$ only by the weight term $\sqrt{(L^{ijlm}_1)^2 + (L^{ijlm}_2)^2}$. However, this weight term varies from point to point. The interesting question rises: why practically does the unnormalized but simple and linear cost function $\mathcal{E}1$ give satisfying results?

In the following four sections, we are going to assume the camera is calibrated, that is, $M' = g', M'' =$

g'' and $\mathbf{u}, \mathbf{u}', \mathbf{u}''$ are normalized coordinates. The extension to the uncalibrated sequence will be briefly discussed in Section7.

4 Geometrical Interpretation of Trilinear Constraint

Instead of investigating the problem from the 2 dimensional image plane point of view, we revisit the trilinear constraint from the 3 dimensional structure viewpoint, and reach the interpretation for $\mathcal{E}1^{ijlm}$ and also the answer to the question of its performance.

Claim 1: The trilinear constraint Eq.(2) is a weighted depth matching constraint with a geometrically interpreted weight function.

Consider two pairs of views, (1,2) and (1,3), let us call λ_a the depth of the point $\mathbf{P}$ in view1 frame as reconstructed from the first pair, and λ_b from the second pair. When there is no noise of course $\lambda_a = \lambda_b$. Since $M = [I, 0]$, we can write the point position as $\mathbf{x} = [\lambda_a \mathbf{u}^T, 1]^T$. Substituting this expression into the second projective function gives:

$$\lambda^{'} \mathbf{u}^{'} = M'\mathbf{x} = [\mathbf{a}, \mathbf{v}'][\lambda_a \mathbf{u}^T, 1]^T \tag{5}$$

The depth of the point $\mathbf{P}$ in these first 2 views, λ_a and $\lambda^{'}$, can be estimated by the triangulation:

$$\begin{bmatrix} a^1_k u^k & -u'^1 \\ a^2_k u^k & -u'^2 \\ a^3_k u^k & -u'^3 \end{bmatrix} \begin{bmatrix} \lambda_a \\ \lambda^{'} \end{bmatrix} = - \begin{bmatrix} v'^1 \\ v'^2 \\ v'^3 \end{bmatrix} \tag{6}$$

The similar triangulation equation can be written if we consider the motion between view1 and view3:

$$\begin{bmatrix} b^1_k u^k & -u''^1 \\ b^2_k u^k & -u''^2 \\ b^3_k u^k & -u''^3 \end{bmatrix} \begin{bmatrix} \lambda_b \\ \lambda^{''} \end{bmatrix} = - \begin{bmatrix} v''^1 \\ v''^2 \\ v''^3 \end{bmatrix} \tag{7}$$

Considering all 3 views together, we may enforce the depth of the point to be identical now after triangulating it by either pair of views:

$$\mathcal{E}3 = \lambda_a - \lambda_b = 0 \tag{8}$$

Eq.(8) is called a depth matching constraint. We are very aware that such depth constraint in $3D$ space is not the optimal one for motion estimation comparing to the other pure geometrical constraints, neither is the linear method shown in Eq.(6) and (7) for triangulation [7]. However, from the probability viewpoint, we can modify the constraint to be the most reasonable and reliable one by enforcing the probabilistically optimal weight as shown later in Eq.(12). The interpretation of the trilinear constraints will be based on this viewpoint.

Each triangulation set has 3 equations, one of them is redundant. Therefore, taking any two (i,j) from Eq.(6) and any two (l,m) from Eq.(7) to calculate λ_a and λ_b, we have the depth matching equation:

$$0 = \mathcal{E}3^{ijlm} = \frac{N_{\lambda_a}}{D_{\lambda_a}} - \frac{N_{\lambda_b}}{D_{\lambda_b}} \tag{9}$$
$$= \frac{\begin{bmatrix} u'^j & -u'^i \end{bmatrix} \begin{bmatrix} v'^i \\ v'^j \end{bmatrix}}{\begin{bmatrix} -u'^j & u'^i \end{bmatrix} \begin{bmatrix} a^i_k u^k \\ a^j_k u^k \end{bmatrix}} - \frac{\begin{bmatrix} u''^m & -u''^l \end{bmatrix} \begin{bmatrix} v''^l \\ v''^m \end{bmatrix}}{\begin{bmatrix} -u''^m & u''^l \end{bmatrix} \begin{bmatrix} b^l_k u^k \\ b^m_k u^k \end{bmatrix}}$$

For convenience $N_{\lambda_a}, D_{\lambda_a}$ denote the numerator and denominator of the formula of λ_a here, similar notation for λ_b. By multiplying the product of the two denominators $D_{\lambda_a} D_{\lambda_b}$ on both sides of Eq.(9), we get exactly the trilinear constraint shown in Eq.(2). That is:

$$\mathcal{E}1^{ijlm} = \mathcal{E}3^{ijlm} D_{\lambda_a} D_{\lambda_b} = 0$$

Therefore, the trilinear cost function $\mathcal{E}1^{ijlm}$ is from the depth matching cost $\mathcal{E}3^{ijlm}$ weighted by the factor:

$$\omega_{tri}{}^{ijlm} = D_{\lambda_a} D_{\lambda_b} \tag{10}$$

The importance of the weight ω_{tri} (subscript 'tri' represents 'trilinear') only depends on its absolute value. Therefore, we can always enforce the same sign on D_{λ_a} and D_{λ_b} so that ω_{tri} takes positive value. To examine the geometrical meaning of ω_{tri}, let us first look at D_{λ_a}. The vector $\mathbf{au}$ transforms $\mathbf{u}$ from view1 into view2 so that it is represented in the same coordinate frame as the vector $\mathbf{u}'$. Taking their (i, j) components is equivalent to projecting the vectors onto the $i - j$ plane (for example, $1 - 3$ is the $X - Z$ plane) in the view2 frame. Consequently, the depth is estimated from a 2D triangle by projecting the 3D structure triangle onto this $i - j$ plane to get rid of one redundant equation. Without considering the order of three axis, we can describe these projected vectors in the view2 as: $\mathbf{u}'_2 = \begin{bmatrix} u'^i & u'^j & 0 \end{bmatrix}$ and $\mathbf{u}_2 = \begin{bmatrix} a^i_k u^k & a^j_k u^k & 0 \end{bmatrix}$. The value of D_{λ_a} is exactly the norm of the cross product of these two projected vectors:

$$|D_{\lambda_a}| = \left| \begin{bmatrix} -u'^j & u'^i \end{bmatrix} \begin{bmatrix} a^i_k u^k & a^j_k u^k \end{bmatrix}^T \right| \tag{11}$$
$$= ||\mathbf{u}'_2 \times \mathbf{u}_2|| = ||\mathbf{u}'_2|| ||\mathbf{u}_2|| \sin\theta$$

Where θ is the angle between vectors $\mathbf{u}'_2$ and $\mathbf{u}_2$. Since we project the structure triangle onto the $i - j$ plane, Eq.(11) contains both the 3D structure triangle information and the goodness of the choice of $i - j$ as the projection plane. If the projection is ill-conditioned, e.g. $i - j$ is perpendicular to the 3D triangle plane, the 3D triangle will be projected only as a line and D_{λ_a} will go to zero. If the projected triangle keeps the shape of the 3D triangle, $||\mathbf{u}'_2||$ and $||\mathbf{u}_2||$ have reasonable values, then the angle θ will represent the confidence of the triangulation. The term $\sin\theta$ changes from 0 to 1 while the directions of two vectors $\mathbf{u}'_2$ and $\mathbf{u}_2$ change from collinear to perpendicular. Intuitively the more collinear the two vectors are, the more ill-conditioned the depth triangulation is. Consequently D_{λ_a} is closer to zero by Eq.(11) and a smaller weight ω_{tri} will be added for this feature point. Same interpretation can be applied on the effect of D_{λ_b} to ω_{tri}. In the 2 view case, Spetsakis and Aloimonos [12] had similar intuitive observation that the algebraic epipolar constraint is naturally weighted by a function of points reliability in the scene.

Claim 2: The weight ω_{tri} is an approximately optimal weight function for the depth matching constraint.

Let us model the feature localization noise on the image plane as a Gaussian distribution $\mathcal{N}_u(0, \sigma_u^2)$. The variance of the triangulation for each point with projected planes $(i - j), (l - m)$ can be derived and denoted as $\sigma^2_{\lambda_a}(ij)$ and $\sigma^2_{\lambda_b}(lm)$. According to Eq.(9), the cost function $\mathcal{E}3^{ijlm}$ has the variance:

$$\sigma^2_{\mathcal{E}3^{ijlm}} = \sigma^2_{\lambda_a}(ij) + \sigma^2_{\lambda_b}(lm)$$

This variance encodes the reliability of the depth triangulation and matching. Therefore, from probability viewpoint, the constraint should be optimally weighted by $1/\sigma_{\mathcal{E}3^{ijlm}}$ to recover the motion parameters:

$$\{M', M''\}^* = arg \min_{M', M'' \in S_M} \sum_{\mathbf{P}} \sum_{ijlm} \left(\frac{\mathcal{E}3^{ijlm}}{\sigma_{\mathcal{E}3^{ijlm}}} \right)^2$$

$$\omega_{opt}{}^{ijlm} = \frac{1}{\sigma_{\mathcal{E}3^{ijlm}}} = \frac{1}{\sqrt{\sigma^2_{\lambda_a}(ij) + \sigma^2_{\lambda_b}(lm)}} \quad (12)$$

To show the relationship between trilinear weight ω_{tri} and the optimal weight ω_{opt}, we need to investigate first how D_λ reflects the term $\frac{1}{\sigma_\lambda}$ in the linear triangulation. In order to clarify this issue, we make a few approximations. The 2D projected triangulation that illustrates the trilinear constraint is shown in Fig1. u',u and T represent the projected vector of $\mathbf{u}'$, $\mathbf{au}$ and $\mathbf{v}'$ on the plane. The depth triangulation equation is $\lambda' u' = \lambda u + T$. u' and u are assumed to be unit length for all features. If $(i = (1, 2), j = 3)$ projected plane is chosen, this introduces an approximation of at most 15% error for a field of view of 60^o. Let us denote the vectors by:

$$u' = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} \qquad u = \begin{bmatrix} \cos\beta \\ \sin\beta \end{bmatrix} \qquad T = \begin{bmatrix} t \\ 0 \end{bmatrix}$$

Where $(\alpha, \beta) \in (0, 2\pi)$ representing all possible camera orientations. The depth λ can be easily obtained: $\lambda = \frac{t \sin\alpha}{\sin(\beta - \alpha)}$, and our goal is to check the relationship between its denominator $D = \sin(\beta - \alpha)$ and its reliability. Let us assume that Gaussian noise $\mathcal{N}(0, \sigma^2)$ is added to both α and β. This assumption transfers the noise model $\mathcal{N}_u(0, \sigma_u^2)$ on the image plane to the rotation angle, which is reasonable since readers can verify, for the 60^o angle of view, σ varies only in the range $[0.75\sigma_u, \sigma_u]$. The reliability of λ can be represented by:

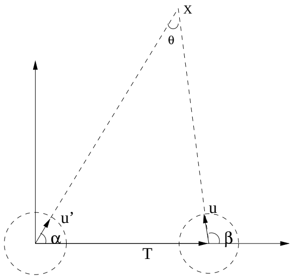

Figure 1: 2D Triangulation Illustration

$$\frac{1}{\sigma_\lambda} = \frac{\sin^2(\beta - \alpha)}{t\sigma\sqrt{\sin^2\alpha\cos^2(\beta - \alpha) + \sin^2\beta}} \quad (13)$$

We pick the constants $t = 4, \sigma = 0.01rad$. Eq.(13) intuitively tells us that $1/\sigma_\lambda$ strongly related to D^2. The objective of this reliability weight is to 'reject' unreliable points by assigning to them smaller weight. Observe that the maximum of its denominator is only $\sqrt{2}t\sigma$. Therefore, for unreliable points, the cause of their reliability quantity $1/\sigma_\lambda \to 0$ must be $D^2 \to 0$.

Since D is a function of $\theta = \beta - \alpha$, we pick 2 independent angles β and θ to represent the triangles. Fig2 shows in a form of a 3D mesh plot the relation between $D^2 = \sin^2\theta$, β and $1/\sigma_\lambda$ for the triangles in the range $\theta \in (0, \pi/2), \beta \in (\theta, \pi)$.

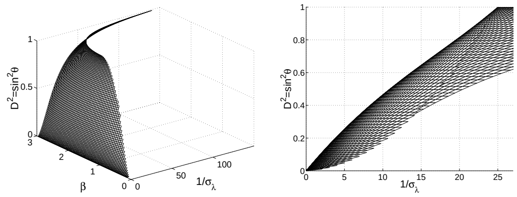

Figure 2: Mesh Surface. (Left): General view of β and $\sin^2\theta$ vs. their corresponding reliability quantity $1/\sigma_\lambda$. (Right): Side view of $\sin^2\theta$ vs. $1/\sigma_\lambda$ over all β. Approximate linearity is observed between $\sin^2\theta$ and $1/\sigma_\lambda$.

Observe that the main part of the mesh ($D^2 < 0.5$) in Fig2(left) is almost planar. In other words, D^2 varies almost linearly as a function of $1/\sigma_\lambda$ for all angles β. There is a small portion in the plot where $1/\sigma_\lambda$ blows up and $D^2 \approx 1$. It corresponds to the situation $\beta \approx \pi, \theta \approx \pi/2$ which is impossible in practice and should be ignored. Fig2(right) shows a side view of the mesh surface in the (D^2,$1/\sigma_\lambda$) plane. A roughly linear dependence between D^2 and $1/\sigma_\lambda$ is most noticeable on this plot. The narrower band and sharp end of the plot around 0 indicates that the more unreliable the triangulation is, the better D^2 represents $1/\sigma_\lambda$ up to an overall scale. It fits the criteria of a weight function to 'reject' correctly the unreliable points. Therefore, D^2 is a good approximation of the reliability quantity $1/\sigma_\lambda$ up to an overall scale for the triangulation. By computing the first order derivative of $1/\sigma_\lambda$ as a function of $\sin^2\theta$ at $\sin^2\theta = 0$, we derive a closed form expression for the proportionality factor between the two quantities: $D^2 \approx \sqrt{2}t\sigma|\sin\beta|\frac{1}{\sigma_\lambda}$. In the case $t = 4, \sigma = 0.01$, the maximum slope shown in Fig2(right) is about 0.05.

In general, we can assume the motions are smooth so that $\sigma_{\lambda_a} \approx \sigma_{\lambda_b}, D_{\lambda_a} \approx D_{\lambda_b}$. Then we have $\omega_{opt} \approx 1/(\sqrt{2}\sigma_{\lambda_a})$ and $\omega_{tri} \approx D^2_{\lambda_a}$. Therefore, ω_{tri} is roughly linear to ω_{opt}. Consequently, ω_{tri} is a good substitute to the optimal weight for depth matching constraint.

5 Scale Propagation Scheme

To show the different performances of depth matching constraint under different weighting conditions in the next experiment section, we only leave the relative scale $s = \frac{\|T''\|}{\|T'\|}$ as the unknown motion parameter which makes it easier for comparison and also demonstrates the problem well enough. The other motion parameters, i.e., motions between each two views with unit length translation are given by the standard 2view motion estimation scheme. Denote the depth of points in view1 frame triangulated between two pairs of views with the unit length motions by $\lambda_{au}, \lambda_{bu}$, and their variances by $\sigma^2_{\lambda_{au}}, \sigma^2_{\lambda_{bu}}$. The depth matching constraint Eq.(8) is modified as:

$$\lambda_{au} - s\lambda_{bu} = 0 \tag{14}$$

The relative scale s is invariant to all the points $\mathbf{P}$ while the depth $\lambda_{au}, \lambda_{bu}$ are functions of points $\mathbf{P}$. The least squares solution of s which minimizes $\sum_{\mathbf{P}}(\lambda_{au} - s\lambda_{bu})^2$ is:

$$s^* = \frac{\sum_{\mathbf{P}} \lambda_{bu}\lambda_{au}}{\sum_{\mathbf{P}} \lambda^2_{bu}} \tag{15}$$

However, this is an asymmetric solution because we can also write Eq.(14) as a function of $1/s$: $\frac{1}{s}\lambda_{au} - \lambda_{bu} = 0$. Then we obtain the least squares solution of $1/s$ which minimizes $\sum_{\mathbf{P}}(\lambda_{au}/s - \lambda_{bu})^2$:

$$\frac{1}{s^*} = \frac{\sum_{\mathbf{P}} \lambda_{au}\lambda_{bu}}{\sum_{\mathbf{P}} \lambda^2_{au}} \tag{16}$$

Both Eq.(15) and Eq.(16) give reasonable estimations though they minimize different cost functions. A practical solution is obtained by $\sqrt{Eq.(15)/Eq.(16)}$:

$$s^* = \sqrt{\frac{\sum_{\mathbf{P}} \lambda^2_{au}}{\sum_{\mathbf{P}} \lambda^2_{bu}}} \tag{17}$$

Trilinear constraints are also linear functions of the scale s: $\mathcal{E}1^{ijlm} = N_{\lambda_{au}}D_{\lambda_{bu}} - sN_{\lambda_{bu}}D_{\lambda_{au}} = 0$. And the similar solution of s using trilinear constraints is:

$$s^*_{tri} = \sqrt{\frac{\sum_{\mathbf{P}}\sum_{ijlm}(N_{\lambda_{au}}D_{\lambda_{bu}})^2}{\sum_{\mathbf{P}}\sum_{ijlm}(N_{\lambda_{bu}}D_{\lambda_{au}})^2}} \tag{18}$$

The solution for s using the optimal weighted cost function according to Eq.(12) is:

$$\begin{aligned} s^* &= arg\min_s \sum_{\mathbf{P}}\sum_{ijlm}\left(\frac{\lambda_{au} - s\lambda_{bu}}{\sqrt{\sigma^2_{\lambda_{au}} + s^2\sigma^2_{\lambda_{bu}}}}\right)^2 \\ &\approx arg\min_s \sum_{\mathbf{P}}\sum_{ijlm}\left(\frac{\lambda_{au} - s\lambda_{bu}}{\sqrt{\sigma^2_{\lambda_{au}} + \frac{\lambda^2_{au}}{\lambda^2_{bu}}\sigma^2_{\lambda_{bu}}}}\right)^2 \\ &= \frac{\sum_{\mathbf{P}}\sum_{ijlm}\frac{\lambda^2_{bu}}{\lambda^2_{bu}\sigma^2_{\lambda_{au}}+\lambda^2_{au}\sigma^2_{\lambda_{bu}}}\lambda_{bu}\lambda_{au}}{\sum_{\mathbf{P}}\sum_{ijlm}\frac{\lambda^2_{bu}}{\lambda^2_{bu}\sigma^2_{\lambda_{au}}+\lambda^2_{au}\sigma^2_{\lambda_{bu}}}\lambda_{bu}\lambda_{bu}} \end{aligned} \tag{19}$$

During the derivation, the scale s appearing in the weight function is approximated by the individual scale $\lambda_{au}/\lambda_{bu}$, which gives exactly individual weight function and generates a linear estimator. The same solution can also be reached by another method: weight the individual scale by the inverse of its variance, $\frac{\sum_{\mathbf{P}}\sum_{ijlm}\left(s(\mathbf{P},ijlm)/\sigma^2_{s(\mathbf{P},ijlm)}\right)}{\sum_{\mathbf{P}}\sum_{ijlm}\left(1/\sigma^2_{s(\mathbf{P},ijlm)}\right)}$. Interested readers can easily verify it.

The solution for $1/s$ using the optimum weighted constraint can be similarly derived. Unlike the Eq.(15) and Eq.(16), to find the s^*_{opt}, there are no common terms that can cancel each other, because two weights $\omega_{opt,s}, \omega_{opt,1/s}$ are different:

$$\begin{aligned} \omega_{opt,s} &= \sqrt{\frac{\lambda^2_{bu}}{\lambda^2_{bu}\sigma^2_{\lambda_{au}} + \lambda^2_{au}\sigma^2_{\lambda_{bu}}}} \\ \omega_{opt,1/s} &= \sqrt{\frac{\lambda^2_{au}}{\lambda^2_{bu}\sigma^2_{\lambda_{au}} + \lambda^2_{au}\sigma^2_{\lambda_{bu}}}} \end{aligned} \tag{20}$$

The optimal solution s^*_{opt} is much more complicated than s^* in Eq.(17) and s^*_{tri} in Eq.(18).

$$s^*_{opt} = \tag{21}$$
$$\sqrt{\frac{(\sum_{\mathbf{P}}\sum_{ijlm}\omega^2_{opt,s}\lambda_{bu}\lambda_{au})(\sum_{\mathbf{P}}\sum_{ijlm}\omega^2_{opt,1/s}\lambda_{au}\lambda_{au})}{(\sum_{\mathbf{P}}\sum_{ijlm}\omega^2_{opt,s}\lambda_{bu}\lambda_{bu})(\sum_{\mathbf{P}}\sum_{ijlm}\omega^2_{opt,1/s}\lambda_{bu}\lambda_{au})}}$$

6 Experiments

6.1 Simulated Translation

This experiment is designed to demonstrate the basic results in Section 4. A total of 500 features are uniformly distributed inside a 60 × 60 × 60 cm^3 cubic which is put right in front of the first view 70 cm away from the camera center. Therefore, the point coordinates in the camera frame of view1 are in the range of $X \in [-30, 30], Y \in [-30, 30], Z \in [40, 100]$ which corresponds to 73.7^o viewing angle. The focal length is assumed to be 1 for feature normalization. The camera translates in the $X - Z$ plane to its view2 and view3 positions respectively with $T' = \begin{bmatrix}1.5 & 0 & 1.5\end{bmatrix}^T$ and $T'' = \begin{bmatrix}-3 & 0 & -3\end{bmatrix}^T$. The projection matrix are

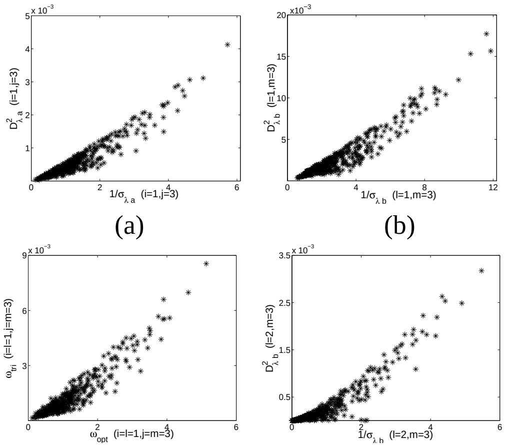

Figure 3: Characteristic of Weight Functions of Simulated Translation Experiment.

$M' = \begin{bmatrix} I & T' \end{bmatrix}, M'' = \begin{bmatrix} I & T'' \end{bmatrix}$. We make the $\|T''\|$ twice larger than $\|T'\|$ here so that the triangulation from (view1,view3) is much more reliable than that from (view1,view2). Points are projected onto the 3 image planes and Gaussian noise $\mathcal{N}(0, 0.002)$ is added to feature measurements.

Fig3(a) and (b) show the relationship D^2_λ vs. $1/\sigma_\lambda$ of all the points $\mathbf{P}$ when the $X - Z$ is chosen as the projection plane for both triangulation. Each plot is closely linear which confirms that D^2_λ does represent well the triangulation reliability. Notice from the axis of the plots, the two triangulation have different levels of reliability and also different approximated slopes due to their different baselines. This case is more complicated than the smooth motion that we assumed in the last paragraph of Section4 to infer the relationship of ω_{tri} and ω_{opt}. However, Fig3(c) shows that ω_{tri} still varies linearly as a function of the optimal weight ω_{opt}. In this experiment, the $X - Z$ plane is the best for projection (Fig3(a),(b)). If we project the triangles in space onto the $Y - Z$ plane, the triangles will be deformed, especially for the points whose Y components are close to zero. Therefore the reliability of the depth estimation from such triangles will decrease dramatically. Fig3(d) illustrates that effect. Projecting the triangles in (view1,view3) onto the $Y - Z$ plane, reliability shown in Fig3(d) is four times smaller than that on Fig3(b) which chooses the $X - Z$ plane to project the same triangles. The dominant plane varies from case to case between the $X - Z$ and $Y - Z$, but in general will not be the $X - Y$ plane.

6.2 Real Corridor Navigation

When the camera is moving forward in an environment, the epipoles of any two frames are close to the image centers. Therefore, the triangulation is ill-conditioned for the feature points that appear close to the image center. Therefore, it is necessary in that case to appropriately reject those points by assigning to them a scalar weight close to zero. We test the strategy on the real sequence taken with a calibrated CCD video camera mounted on a cart moving forward along a building corridor (see Fig6 left). In this 400 frames long experiment, we choose a baseline of 4 frames for motion estimation. The sequence is difficult in the sense that the features in the far end of the corridor appear almost around the image center and their depth triangulation are not reliable at all.

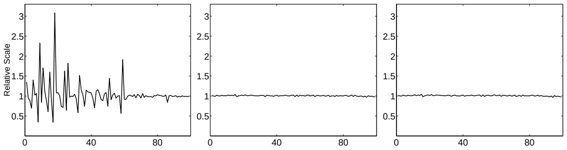

Figure 4: Relative Scale s over Frames for Navigation Sequence (Left): s recovered by direct depth matching. (Center): s recovered by depth matching with optimal weight. (Right): s recovered from trilinear constraints.

The motion parameters with unit length of translation between each two consecutive frames are estimated using the recursive Newton-Raphson method. Given those parameters, the relative scale between $T(t-1,t)$ and $T(t,t+1)$ for all $t \in (1, 100)$ is recovered by three different schemes stated in Section5: direct depth matching Eq.(17), optimal weighted depth matching Eq.(21), and trilinear constraints Eq.(18). A few implementation details: the depth λ and its variance σ^2_λ appearing in Eq.(17) and Eq.(21) are calculated using all 3 triangulation equations. Therefore there is no summation $\sum_{ijlm}$ necessary in Eq.(21). For the trilinear constraints, only 4 equations with combinations of $i = (1,2), l = (1,2), j = m = 3$ are used. Fig4 shows the recovered scale s over time. Although we do not have ground truth for s, it should be around 1 since the cart was moving very smoothly. The first method does not give acceptable result because the estimated s oscillates dramatically between 0.5 and 3. However scale estimation becomes much more stable in the final phase, the reason is that the camera is approaching to the end of the hallway and features located there are projected far away from the image center now. A big improvement is achieved by the second method which uses ω_{opt} as the weight of the depth matching. The result shown on the third plot demonstrates that the trilinear scheme is very competitive to the optimal solution.

Let us pick three frames within the whole sequence (frames 16,17 and 18 in this example) and draw some relationship in detail. Choosing $X - Z$ as projection plane for both triangulation $(i = l = 1, j = m = 3)$, Fig5(a) shows that D^2_λ is almost proportional to the optimal reliability measurement $1/\sigma_\lambda$, Fig5(b) shows the trilinear weight ω_{tri} is also roughly linear to the optimal weight $\omega_{opt,s}$. The property that both of them assign very small weight to the unreliable triangulation is a crucial point that makes the scale recovery schemes successful. The other two projection planes are less informative since they are less important and

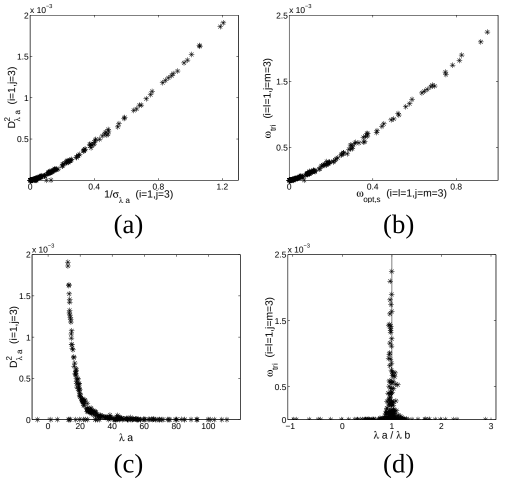

Figure 5: Characteristic of the Trilinear Weight Functions of 3 views in Navigation Sequence.

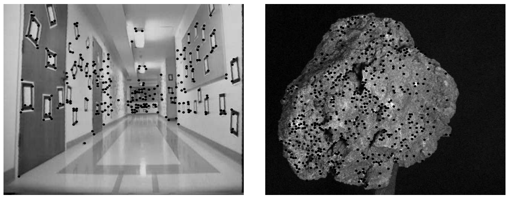

Figure 6: Sample Images. (Left): Image in the real corridor sequence. (Right): Image in the rock sequence. Black dots indicate the features tracked on the image.

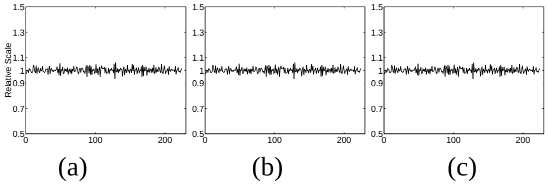

Figure 7: Relative Scale s over Frames for Rock Sequence (Left): s recovered by direct depth matching. (Center): s recovered by depth matching with optimal weight. (Right): s recovered from trilinear constraints. Three plots are similar because all points are weighted almost equally.

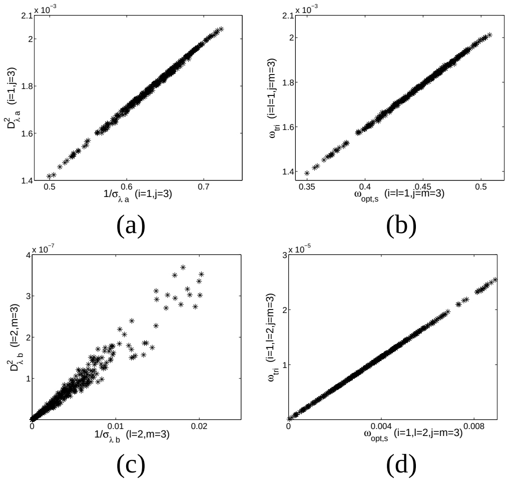

Figure 8: Characteristic of the Weight Functions of 3 Sample Views in Rotating Rock Sequence

less reliable than the $X - Z$ projection in this example. Fig5(c) and (d) show how the weight relates to the depth and individual scale. The depth λ_a, λ_b here are calculated using all 3 triangulation equations, therefore no $(ijlm)$ option necessary. According to Fig5(c), as λ_a increases, $D^2_{\lambda_a}$ decreases. That provides a supporting argument to our analysis since a larger λ_a means that the point is closer to the image center, and therefore unreliable for triangulation. That justifies the small value for $D^2_{\lambda_a}$. Fig5(d) indicates that larger weights are assigned to features with individual scales λ_a/λ_b close to the believed true scale 1, which makes the estimation of scale s using weighted schemes reliable.

6.3 Rotating Rock Experiment

Another experiment is done for a rotating sequence. A textured rock is placed on a turn table. Between two consecutive images, the rotating stage is turned by 2 degrees. A total of 225 frames are acquired.

For this sequence, the relative scale s is recovered reliably with or without weighting factor. Three plots in Fig7 are almost identical. The scale s varies in the range 1 ± 0.04, which is quite acceptable. The reason for the similarity can be found in Fig8(a) and (b). As we expected, linear relationship is observed between $D^2_{\lambda_a}$ and $1/\sigma_{\lambda_a}$ for single 2D triangulation, and between ω_{tri} and $\omega_{opt,s}$ for depth matching. Different from the previous navigation experiment which has most features weighted close to 0, the ranges of the weight functions in this experiment shown in Fig8 are very small, $\omega_{tri} \in [1.4, 2.1] \times 10^{-3}$ and $\omega_{opt,s} \in [0.35, 0.5]$. This indicates that all the features are almost equally reliable. Therefore, three different scale recovery schemes return almost identical results. Fig8(c) shows the reliability of the triangulation if the $Y - Z$ is chosen as the projection plane. The order of D^2 goes to 10^{-7}, which is neglectable compared to the values of $D^2, (i = 1, j = 3)$ in Fig8(a) (order of 10^{-3}). If we pick one triangulation in the $X - Z$ plane and another one in the $Y - Z$ (i=1,l=2,j=m=3) to form the depth matching (which is one of the 9 trilinear constraints), the corresponding weight functions ω_{opt} and ω_{tri} are shown in Fig8(d) to be 100 times smaller than that in Fig8(b). Therefore, among the 9 trilinear constraints, only a few dominate the estimation.

7 Extension to Uncalibrated Case

For an uncalibrated sequence, let us simply assume the unknown calibration matrix $K' = K''$ to be:

$$K' = \begin{bmatrix} f & 0 & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{bmatrix}$$

The projection equation in view2 becomes $\lambda' \mathbf{u}'_{new} = M'_{new}\mathbf{x}$, where $\mathbf{u}'_{new} = K'\mathbf{u}'$ and $M'_{new} = K'M'$. Keeping the projection equation of view1 $\lambda\mathbf{u} = [I, 0]\mathbf{x}$, we get the same triangulation equation as Eq.(6) except that $\mathbf{a}, \mathbf{v}'$ and $\mathbf{u}'$ are substituted by $K'\mathbf{a}, K'\mathbf{v}'$ and $\mathbf{u}'_{new}$ respectively. By choosing different combination (i, j) of 2 equations to estimate the depth, the denominator associated to the uncalibrated case is related to that of the calibrated case as follows:

$$
\begin{aligned}
D_{\lambda_a,new} =& f D_{\lambda_a} \qquad\qquad i = (1,2), j = 3 && (22)\\
D_{\lambda_a,new} =& f^2 D_{\lambda_a} + f(c_2 {u'}^1 a_k^3 u^k + c_1 a_k^2 u^k - c_2 a_k^1 u^k \\
& - c_1 {u'}^2 a_k^3 u^k) \qquad (i,j) = (1,2) && (23)
\end{aligned}
$$

The derivations are straightforward and omitted in this paper. Eq.(22) and Eq.(23) give the hint as for why using trilinear constraints in the uncalibrated case it is better to pre-normalize the feature measurements approximately (see [5], [6]). If we assume $c_1 = c_2 = 0$, the principal point is assured to be at the center of the image, then $D_{\lambda_a,new} = f^2 D_{\lambda_a}$ for $(i,j) = (1,2)$, which is weighted f times larger than the other two triangulation combinations. The scalar f is usually around $500 - 2000$ pixels. When we form the trilinear constraint, the depth matching weight is $\omega_{tri} = D_{\lambda_a} D_{\lambda_b}$. The weight ω_{tri} of the constraint with $(i = l = 1, j = m = 2)$ will be f^4 times larger than that coming from the calibrated case while the weight of the constraint with $(i = l = 1, j = m = 3)$ will only be f^2 times larger than that of its corresponding calibrated case . Such unbalanced weight may affect the whole estimation using trilinear constraints. The influence of c_1, c_2 is not as significant as f, however the closer they are to zero, the better the results will be. This is also a reason why usually only 4 trilinearities $i = (1,2), l = (1,2), j = m = 3$ out of the total 9 constraints are used for 3 views motion estimation with uncalibrated scenes [10].

8 Conclusion

This paper gives a new geometrical interpretation of the algebraic trilinear constraints. It is shown that minimizing the trilinear algebraic equations in a least squares sense is equivalent to minimizing an error in Euclidean space with appropriate weighting applied on every point. This fundamental result provides a explanation as for why, in practice, there is no need of adding extra weighting coefficients to each scalar constraints in order to achieve satisfactory motion estimates. In other words, the sum of the squares of every algebraic equation is a very valid cost function to minimize for estimating motion parameters. All the theoretical statements are supported by experimental results. One additional contribution of this paper is a robust scheme to propagate scale information which is significantly insensitive to noise in the measurement data. Although most of the derivations are done assuming calibrated camera, natural extensions of the results to the uncalibrated case are also provided.

References

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

[2] O. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between n images. *Proc. 5th Int. Conf. Computer Vision*, pages 951–956, 1995.

[3] O. Faugeras and T. Papadopoulo. A nonlinear method for estimating the projective geometry of 3 views. *Proc. 6th Int. Conf. Computer Vision*, pages 477–484, 1998.

[4] R. Hartley. Minimizing algebraic error in geometric estimation problems. *Proc. 6th Int. Conf. Computer Vision*, pages 469–476, 1998.

[5] R.I. Hartley. In defence of the 8-point algorithm. *Proc. 5th Int. Conf. Computer Vision*, pages 1064–1070, 1995.

[6] R.I. Hartley. Lines and points in three views and the trifocal tensor. *Int. J. of Computer Vision*, 22(2):125–140, 1997.

[7] R.I. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68:2:146–157, 1997.

[8] T. Papadopoulo and O. Faugeras. A new characterization of the trifocal tensor. *Proc. 5th Europ. Conf. Comput. Vision, LNCS-Series Vol. 1406-1407, Springer-Verlag*, pages 109–123, 1998.

[9] A. Shashua. Trilinearity in visual recognition by alignment. *Proc. 3rd Europ. Conf. Comput. Vision, J.-O. Eklundh (Ed.), LNCS-Series Vol. 800-801, Springer-Verlag*, I:479–484, 1994.

[10] A. Shashua. Algebraic functions for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):779–789, Aug 1995.

[11] A. Shashua and M. Werman. Trilinearity of three perspective views and its associated tensor. *Proc. 5th Int. Conf. Computer Vision*, pages 920–925, 1995.

[12] M.E. Spetsakis and Y. Aloimonos. A multiframe approach to visual motion perception. *Int. J. of Computer Vision*, 6:245–255, 1991.

[13] G.P. Stein and A. Shashua. On degeneracy of linear reconstruction from three views: linear line complex and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(3):244–251, Mar. 1999.

[14] P. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, pages 591–605, 1997.