A Six Point Solution for Structure and Motion

F. Schaffalitzky[1], A. Zisserman[1], R. I. Hartley[2], and P. H. S. Torr[3]

[1] Robotics Research Group, Dept. of Engineering Science, Oxford OX1 3PJ, UK
[2] G.E. CRD, Schenectady, NY 12309, USA
[3] Microsoft Research, 1 Guildhall St, Cambridge CB2 3NH, UK

Abstract. The paper has two main contributions: The first is a set of methods for computing structure and motion for $m \geq 3$ views of 6 points. It is shown that a geometric image error can be minimized over all views by a simple three parameter numerical optimization. Then, that an algebraic image error can be minimized over all views by computing the solution to a cubic in one variable. Finally, a minor point, is that this "quasi-linear" linear solution enables a more concise algorithm, than any given previously, for the reconstruction of 6 points in 3 views.
The second contribution is an m view $n \geq 6$ point robust reconstruction algorithm which uses the 6 point method as a search engine. This extends the successful RANSAC based algorithms for 2-views and 3-views to m views. The algorithm can cope with missing data and mismatched data and may be used as an efficient initializer for bundle adjustment.
The new algorithms are evaluated on synthetic and real image sequences, and compared to optimal estimation results (bundle adjustment).

1 Introduction

A large number of methods exist for obtaining 3D structure and motion from features tracked through image sequences. Their characteristics vary from the so-called *minimal* methods [15, 16, 22] which work with the least data necessary to compute structure and motion, through intermediate methods [5, 18] which may perform mismatch (outlier) rejection as well, to the full-bore *bundle adjustment.*

The minimal solutions are used as search engines in robust estimation algorithms which automatically compute correspondences and tensors over multiple views. For example, the two-view seven-point solution is used in the RANSAC estimation of the fundamental matrix in [22], and the three-view six-point solution in the RANSAC estimation of the trifocal tensor in [21]. It would seem natural then to use a minimal solution as a search engine in four or more views. The problem is that in four or more views a solution is forced to include a minimization to account for measurement error (noise). This is because in the two-view seven-point and three-view six-point cases there are the same number of measurement constraints as degrees of freedom in the tensor; and in both cases one or three real solutions result (and the duality explanation for this equivalence was given by [3]). However, the four-views six-points case provides one more constraint than the number of degrees of freedom of the four-view geometry (the quadrifocal tensor). This means that unlike in the two- and three-view

cases where a tensor can be computed which exactly relates the measured points (and also satisfies its internal constraints), this is not possible in the four (or more) view case. Instead it is necessary to minimize an image measurement error whether algebraic or geometric.

In this paper we develop a novel quasi-linear solution for the 6 point case in three or more views. The solution minimizes an algebraic image error, and its computation involves only a SVD and the solution of a cubic equation in a single variable. This is described in section 3. We also describe a sub-optimal method (compared to bundle adjustment) which minimizes geometric image error at the cost of only a three parameter optimization. Before describing the new solutions, we first demonstrate the poor estimate which results if the error that is minimized is not in the measured image coordinates, but instead in a projectively transformed image coordinate frame. This is described in section 2.

A second part of the paper describes an algorithm for computing a reconstruction of cameras and 3D scene points from a sequence of images. The objectives of such algorithms are now well established:

1. **Minimize reprojection error**. A common statistical noise model assumes that measurement error is isotropic and Gaussian in the image. The Maximum Likelihood Estimate in this case involves minimizing the total squared reprojection error over the cameras and 3D points. This is bundle adjustment.
2. **Cope with missing data**. Structure-from-motion data often arises from tracking features through image sequences and any one track may persist only in few of the total frames.
3. **Cope with mismatches**. Appearance-based tracking can produce tracks of non-features. A common example is a T-junction which generates a strong corner, moving slowly between frames, but which is not the image of any one point in the world.

Bundle adjustment [6] is the most accurate and theoretically best justified technique. It can cope with missing data and, with a suitable robust statistical cost function, can cope with mismatches. It will almost always be the final step of a reconstruction algorithm. However, it is expensive to carry out and, more significantly, requires a good initial estimate in order to be effective (fewer iterations, and less likely to converge to local minimum). Current methods of initializating a bundle adjustment include factorization [10, 12, 18, 23], hierarchical combination of sub-sequences [5], and the Variable State Dimension Filter (VSDF) [14].

In the special case of affine cameras, factorization methods [19] minimize reprojection error [17] and so give the optimal solution found by bundle adjustment. However, factorization cannot cope with mismatches, and methods to overcome missing data [11] lose the optimality of the solution. In the general case of perspective projection iterative factorization methods have been successfully developed and have recently proved to produce excellent results. The problems of missing data and mismatches remain though.

In this paper we describe a novel algorithm for computing a reconstruction satisfying the three basic objectives above (optimal, missing data, mismatches). It is based on using the six-point algorithm as a robust search engine, and is described in section 4.

Notation. The *standard basis* will refer to the five points in $\mathbb{P}^3$ whose homogeneous coordinates are :

$$\mathbf{E}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{E}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{E}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{E}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad \mathbf{E}_5 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

For a 3-vector $\mathbf{v} = (x, y, z)^\top$, we use $[\mathbf{v}]_\times$ to denote the 3×3 skew matrix such that $[\mathbf{v}]_\times \mathbf{u} = \mathbf{v} \times \mathbf{u}$, where $\times$ denotes the vector cross product. For three points in the plane, represented in homogeneous coordinates by $\mathbf{x}, \mathbf{y}, \mathbf{z}$, the incidence relation of collinearity is the vanishing of the bracket $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ which denotes the determinant of the 3×3 matrix whose columns are $\mathbf{x}, \mathbf{y}, \mathbf{z}$. It equals $\mathbf{x} \cdot (\mathbf{y} \times \mathbf{z})$ where $\cdot$ is the vector dot product.

2 Linear Estimation Using a Duality Solution

This section briefly outlines a method proposed by Hartley [8] for computing a reconstruction for six points in three or more views. The method is based on the Carlsson and Weinshall [2, 3] duality between points and cameras. From this duality it follows that an algorithm to compute the fundamental matrix (seven or more points in two views) may be applied to six points in three or more views. This has the advantage that it is a linear method however, as we shall demonstrate, the error distribution that is minimized is transformed in a highly non linear way, leading to a biased estimate. Thus this algorithm is only included here as a warning against minimizing errors in a projectively transformed image frame – we are not recommending it.

The duality proceeds as follows. A projective basis is chosen in each image such that the first four points are

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \mathbf{e}_4 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Assuming in addition that the corresponding 3D points are $\mathbf{E}_1, \ldots, \mathbf{E}_4$, the camera matrix may be seen to be of the form

$$\mathtt{P} = \begin{bmatrix} a_i & & -d_i \\ & b_i & -d_i \\ & & c_i \; -d_i \end{bmatrix} \tag{1}$$

Such a camera matrix is called a *reduced camera matrix*. Now, if $\mathbf{X} = (\mathrm{X}, \mathrm{Y}, \mathrm{Z}, \mathrm{T})^\top$ is a 3D point, then it can be verified that

$$\begin{bmatrix} a & & -d \\ & b & -d \\ & & c \; -d \end{bmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ \mathrm{T} \end{pmatrix} = \begin{bmatrix} \mathrm{X} & & & -\mathrm{T} \\ & \mathrm{Y} & & -\mathrm{T} \\ & & \mathrm{Z} & -\mathrm{T} \end{bmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \tag{2}$$

Note that the rôles of point and camera are swapped in this last equation. This observation allows us to apply the algorithm for projective reconstruction from two views of many points to solve for six point in many views. The general idea is as follows.

1. Apply a transformation to each image so that the first four points are mapped to the points $\mathbf{e}_i$ of a canonical image basis.
2. The two other points in each view are also transformed by these mappings - a total of two points in each image. Swap the rôles of points and views to consider this as a set of two views of several points.
3. Use a projective reconstruction algorithm (based on the fundamental matrix) to solve the two-view reconstruction problem.
4. Swap back the points and camera coordinates as in (2).
5. Transform back to the original image coordinate frame.

The main difficulty with this algorithm is the distortion of the image measurement error distributions by the projective image mapping. One may work very hard to find a solution with minimal residual error with respect to the transformed image coordinates only to find that these errors become very large when the image points are transformed back to the original coordinate system. A circular Gaussian distribution is transformed by a projective transformation to a distribution that is no longer circular, and not even Gaussian. This is illustrated in figure 1. Common methods of two-view reconstruction are not able to handle such error distributions effectively. The method used for reconstruction from the transformed data was a dualization of one of the best methods available for two-view reconstruction – an iterative method that minimizes algebraic error [9].

3 Reconstruction from Six Points over m Views

This section describes the main algebraic development of the six point method. In essence it is quite similar to the development given by Hartley [7] and Quan [15] for a reconstruction of 6 points from 3 views. The difference is that Quan used a standard projective basis for both the image and world points, whereas here the image coordinates are not transformed. As described in section 2 the use of a standard basis in the image severely distorts the error that is minimized. The numerical results that follow demonstrate that the method described here produces a near optimal solution.

In the following it will be assumed that we have six image points $\mathbf{x}_i$ in correspondence over m views. The idea then is to compute cameras for each view such that the scene points $\mathbf{X}_i$ project exactly to their image $\mathbf{x}_i$ for the first five points. Any error minimization required is then restricted to the sixth point $\mathbf{X}_6$, in the first instance, leading to a three parameter optimization problem.

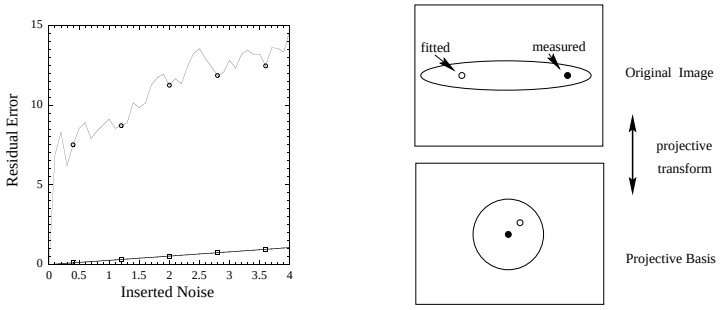


Fig. 1. Left: Residual error as a function of image noise for six points over 20 views. The upper curve is the result of a duality-based reconstruction algorithm, the lower is the result of bundle adjustment. The method for generating this synthetic data is described in section 3.5. As may be seen the residual error of the duality-based algorithm is extremely high, even for quite low noise levels. It is evident that this method is unusable. In fact the results prove to be unsatisfactory for initializing a bundle adjustment in the original coordinate system. Right: Minimizing geometric error (as algebraic error minimization tries to approximate this) in a very projectively transformed space pulls back to a point away from the ellipse centre in the original image.

3.1 A Pencil of Cameras

Each correspondence between a scene point $\mathbf{X}$ and its image $\mathbf{x}$ under a perspective camera $\mathtt{P}$ gives three linear equations for $\mathtt{P}$ whose combined rank is 2. These linear equations are obtained from

$$\mathbf{x} \times \mathtt{P}\mathbf{X} = \mathbf{0} \tag{3}$$

Given only five scene points, assumed to be in general position, it is possible to recover the camera up to a 1-parameter ambiguity. More precisely, the five points generate a linear system of equations for $\mathtt{P}$ which may be written $\mathtt{M}\mathbf{p} = \mathbf{0}$, where $\mathtt{M}$ is a 10×12 matrix formed from two of the linear equations (3) of each point correspondence, and $\mathbf{p}$ is $\mathtt{P}$ written as a 12-vector. This system of equations has a 2-dimensional null-space and thus results in a pencil of cameras.

We are free to choose the position of the five world points (e.g. they could be chosen to be the points of the standard projective frame $\mathbf{E}_1, \ldots, \mathbf{E}_5$) thus both $\mathbf{X}_i$ and $\mathbf{x}_i$ $(i = 1, \ldots, 5)$ are known and the null-space of $\mathtt{M}$ can immediately be computed. The null-space will be denoted from here on by the basis of 3×4 matrices $[\mathtt{A}, \mathtt{B}]$. Then for any choice of the scalars $(\mu : \nu) \in \mathbb{P}^1$ the camera in the pencil $\mathtt{P} = \mu\mathtt{A} + \nu\mathtt{B}$ exactly projects the first five world points to to the first five image points.

Each camera $\mathtt{P}$ in the pencil has its optical centre located as the null-vector of $\mathtt{P}$ and thus a given pencil of cameras gives rise to a 3D curve of possible camera centres. In general (there are degenerate cases) the locus of possible camera centres will be a twisted cubic passing through the five world points. The five points specify 10 of the 12 degrees of freedom of the twisted cubic, the remaining 2 degrees of freedom are specified by the 2 plane projective invariants of the five

image points. If a sixth point in 3-space lies on the twisted cubic then there is a one parameter family of cameras which will exactly project *all* six space points to their images. This situation can be detected (in principle) because if the space point lies on the twisted cubic then all 6 image points lie on a conic.

3.2 The Quadric Constraints

We continue to consider a single camera $\mathtt{P}$ mapping a set of point $\mathbf{X}_1, \ldots, \mathbf{X}_6$ to image points $\mathbf{x}_1, \ldots, \mathbf{x}_6$. Let $[\mathtt{A}, \mathtt{B}]$ be the pencil of cameras consistent with the projections of the first five points. Since $\mathtt{P}$ lies in the pencil, there are scalars $(\mu : \nu) \in \mathbb{P}^1$ such that $\mathtt{P} = \mu\mathtt{A} + \nu\mathtt{B}$ and so the projection of the sixth world point $\mathbf{X}_6$ is $\mathbf{x}_6 = \mu\mathtt{A}\mathbf{X}_6 + \nu\mathtt{B}\mathbf{X}_6$. This means that the three points $\mathbf{x}_6, \mathtt{A}\mathbf{X}_6, \mathtt{B}\mathbf{X}_6$ are collinear in the image, so

$$[\mathbf{x}_6, \mathtt{A}\mathbf{X}_6, \mathtt{B}\mathbf{X}_6] = 0, \tag{4}$$

which is a quadratic constraint on $\mathbf{X}_6$. The 3×3 determinant of (4) can be expressed as ${\mathbf{X}_6}^\top(\mathtt{A}^\top[\mathbf{x}_6]_\times\mathtt{B})\mathbf{X}_6 = 0$. As the skew-symmetric part of the matrix $(\mathtt{A}^\top[\mathbf{x}_6]_\times\mathtt{B})$ does not contribute to this equation, (4) is equivalent to the constraint that $\mathbf{X}_6$ lies on a quadric Q specified by the symmetric part of $(\mathtt{A}^\top[\mathbf{x}_6]_\times\mathtt{B})$. Also, by construction, each of the first five points $\mathbf{X}_i$ $(i = 1, \ldots, 5)$ lies on Q since ${\mathbf{X}_i}^\top Q\mathbf{X}_i = {\mathbf{X}_i}^\top\mathtt{A}^\top[\mathbf{x}_6]_\times\mathtt{B}\mathbf{X}_i = {\mathbf{x}_i}^\top[\mathbf{x}_6]_\times\mathbf{x}_i = 0$. To summarize so far

Let $[\mathtt{A}, \mathtt{B}]$ be the pencil of cameras consistent with the projections of five known points $\mathbf{X}_i$ to image points $\mathbf{x}_i$. Let $\mathbf{x}_6$ be a sixth image point. Then the 3D point $\mathbf{X}_6$ mapping to $\mathbf{x}_6$ must lie on a quadric Q given by

$$Q = (\mathtt{A}^\top[\mathbf{x}_6]_\times\mathtt{B})sym = \mathtt{A}^\top[\mathbf{x}_6]_\times\mathtt{B} - \mathtt{B}^\top[\mathbf{x}_6]_\times\mathtt{A} \ . \tag{5}$$

In addition, the known points $\mathbf{X}_1, \ldots, \mathbf{X}_5$ also lie on Q.

In the particular case case where the five points $\mathbf{X}_i$ are the points $\mathbf{E}_i$ of a projective basis the conditions ${\mathbf{X}_i}^\top Q\mathbf{X}_i = 0$ allow the form of Q (or indeed of any quadric Q which passes through each $\mathbf{E}_i$) to be specified in more detail: from ${\mathbf{E}_i}^\top Q\mathbf{E}_i = 0$ for $i = 1, \ldots, 4$, we deduce that the four diagonal elements of Q vanish. From ${\mathbf{E}_5}^\top Q\mathbf{E}_5$ it follows that the sum of elements of Q is zero. Thus, we may write Q in the following form

$$Q = \begin{bmatrix} 0 & w_1 & w_2 & -\Sigma \\ w_1 & 0 & w_3 & w_4 \\ w_2 & w_3 & 0 & w_5 \\ -\Sigma & w_4 & w_5 & 0 \end{bmatrix} \tag{6}$$

where $\Sigma = w_1 + w_2 + w_3 + w_4 + w_5$. The conclusion we draw from this is that if $\mathbf{X}_6 = (p, q, r, s)^\top$ is a point lying on Q, the equation ${\mathbf{X}_6}^\top Q\mathbf{X}_6 = 0$ may be written in vector form as

$$(w_1, w_2, w_3, w_4, w_5)^\top \begin{pmatrix} pq - ps \\ pr - ps \\ qr - qs \\ qs - ps \\ rs - ps \end{pmatrix} = 0 \tag{7}$$

or more briefly, $\mathbf{w}^\top\mathcal{X} = 0$, where $\mathcal{X}$ is the column vector in (7).

Note, equations (5)-(7) are algebraically equivalent to the equations obtained by Quan [15] and Carlsson-Weinshall [3]. However, they differ in that here the original image coordinate system is used, with the consequence that a different numerical solution is obtained in the over constrained case. It will be seen in section 3.3 that this algebraic solution may be a close approximation of the solution which minimizes geometric error.

Solving for the Point X. Now consider m views of 6 points and suppose again that the first five world points are in the known positions $\mathbf{E}_1, \ldots, \mathbf{E}_5$. To compute projective structure it suffices to find the sixth world point $\mathbf{X}_6$. In the manner described above, each view provides a quadric on which $\mathbf{X}_6$ must lie. For two views the two associated quadrics intersect in a curve, and consequently there is a one parameter family of solutions for $\mathbf{X}_6$ in that case. The curve will meet a third quadric in a finite number of points, so 3 views will determine a finite number (namely $2 \times 2 \times 2 = 8$ by Bézout's theorem) of solutions for $\mathbf{X}_6$. However, five of these points are the points $\mathbf{E}_1, \ldots, \mathbf{E}_5$ which must lie on all three quadrics. Thus there are up to three possible solutions for $\mathbf{X}_6$. With more than three views, a single solution will exist, except for critical configurations [13].

The general strategy for finding $\mathbf{X}_6$ is as follows: For each view j, the quadratic constraint ${\mathbf{X}_6}^\top Q^j \mathbf{X}_6 = 0$ on $\mathbf{X}_6$ can be written as the linear constraint ${\mathbf{w}^j}^\top \mathcal{X} = 0$ on the 5-vector $\mathcal{X}$ defined in terms of $\mathbf{X}_6$ by equation (7). The vector $\mathbf{w}^j$ is obtained from the coefficients of the quadric Q^j (see below). The basic method is to solve for $\mathcal{X} \in \mathbb{P}^4$ by intersecting hyperplanes in $\mathbb{P}^4$, rather than to solve directly for $\mathbf{X} \in \mathbb{P}^3$ by intersecting quadrics in $\mathbb{P}^3$.

In more abstract terms there is a map $\psi : \mathbb{P}^3 - \to \mathbb{P}^4$, given by $\psi : \mathbf{X} \mapsto \mathcal{X}$ which is a (rational) transformation from $\mathbb{P}^3$ to $\mathbb{P}^4$, and maps any quadric $Q \subset \mathbb{P}^3$ through the five basepoints $\mathbf{E}_i$ into the hyperplane defined in $\mathbb{P}^4$ by

$$w_1 \mathcal{X}_1 + w_2 \mathcal{X}_2 + w_3 \mathcal{X}_3 + w_4 \mathcal{X}_4 + w_5 \mathcal{X}_5 = 0 \tag{8}$$

where the (known) coefficients w_i of $\mathbf{w}$ are $Q_{12}, Q_{13}, Q_{23}, Q_{24}, Q_{34}$.

Computing X from $\mathcal{X}$. Having solved for $\mathcal{X} = (a, b, c, d, e)^\top$ we wish to recover $\mathbf{X} = (p, q, r, s)^\top$. By considering ratios of a, b, c, d, e and their differences, various forms of solution can be obtained. In particular it can be shown that $\mathbf{X}$ is a right nullvector of the following 6×4 design matrix:

$$\begin{pmatrix} e-d & 0 & 0 & a-b \\ e-c & 0 & a & 0 \\ d-c & b & 0 & 0 \\ 0 & e-b & a-d & 0 \\ 0 & e & 0 & a-c \\ 0 & 0 & d & b-c \end{pmatrix} \tag{9}$$

This will have nullity ≥ 1 in the ideal noise-free case where the point $\mathcal{X} = (a, b, c, d, e)^\top$ really does lie in the range of ψ. When the point $\mathcal{X}$ does not lie exactly in the image of ψ, the matrix may have full rank, i.e. no nullvector. In the following we determine a solution such that the matrix always has a nullvector.

A Cubic Constraint. The fact that $\dim \mathbb{P}^3 = 3 < 4 = \dim \mathbb{P}^4$ implies that the image of ψ is not all of $\mathbb{P}^4$. In fact the image is the hypersurface $\mathtt{S}$ cut out by the cubic equation

$$S(a,b,c,d,e) = abd - abe + ace - ade - bcd + bde = \begin{vmatrix} e & e & b \\ d & c & b \\ d & a & a \end{vmatrix} = 0 \qquad (10)$$

This can be verified by direct substitution. Alternatively it can be derived by observing that all 4×4 subdeterminants of (9) must vanish, since it is rank deficient. These subdeterminants will be quartic algebraic expressions in a, b, c, d, e, but are in fact all multiples of the cubic expression S.

The fact that the image $\psi(\mathbf{X})$ of $\mathbf{X}$ must lie on $\mathtt{S}$ introduces the problem of enforcing this constraint ($S = 0$) numerically. This will be dealt with below.

Solving for 3 Views of Six Points. The linear constraints defined by the three hyperplanes (8) cut out a line in $\mathbb{P}^4$. The line intersects $\mathtt{S}$ in three points (generically) (see figure 2). Thus there are three solutions for $\mathbf{X}$. This is a well-known [15] minimal solution. Our treatment gives a simpler (than the Quan [15] or Carlsson and Weinshall [3]) algorithm for computing a reconstruction from six points (and thereby computing a trifocal tensor for the minimum number of point correspondences as in [21]) because it does not require changing basis in the images. To be specific, the algorithm for three views proceeds as follows:

1. From three views, obtain three equations of the form (7) $\mathbf{w}^{j\top}\mathcal{X} = 0$ in the five entries of $\mathcal{X}$. Collecting together the $\mathbf{w}^{j\top}$ as the rows of a 3×5 matrix $\mathtt{W}$, this may be written $\mathtt{W}\mathcal{X} = \mathbf{0}$, which is a homogeneous linear system.
2. Obtain a set of solutions of the form $\mathcal{X} = \alpha\mathcal{X}_1 + \beta\mathcal{X}_2$ where $\mathcal{X}_1$ and $\mathcal{X}_2$ are generators of the null space of the 3×5 linear system.
3. By expanding out the constraint (10), form a homogeneous cubic equation in α and β. There will be either one or three real solutions.
4. Once $\mathcal{X}$ is computed (satisfying the cubic constraint (10)), solve for $\mathbf{X}_6 = (p, q, r, s)^\top$. This could be computed as the null-space of the matrix (9), or more directly, as a vector of suitably chosen 4×4 minors of that matrix.

3.3 Four or More Views

We extend the solution above from three to m views as follows: an equation of the form (7) $\mathbf{w}^{j\top}\mathcal{X} = 0$ is obtained for each view, and these may be combined into a single equation of the form $\mathtt{W}\mathcal{X} = \mathbf{0}$, where $\mathtt{W}$ is a $m \times 5$ matrix for m views.

Now, in the case of perfect data (no measurement error) $\mathtt{W}$ has rank 4 for $m \geq 4$ views, and the nullvector is the unique (linear) solution for $\mathcal{X}$. The point $\mathbf{X}_6$ is then obtained from $\mathcal{X}$, e.g. as the null-vector of (9) ($\mathcal{X}$ satisfies the cubic constraint (10)).

However, if there is measurement error (noise) then there are two problems. First, for $m = 4$ views, although $\mathtt{W}$ has rank 4 the (linear) solution $\mathcal{X}$ to $\mathtt{W}\mathcal{X} = \mathbf{0}$ may not satisfy the cubic constraint (10), i.e. the linear solution may not lie on

$\mathtt{S}$ (and so a unique value of $\mathbf{X}_6$ cannot be obtained as a null-vector from (9) because that matrix will have full rank). Second, and worse still, in the case of $m > 4$ views the matrix $\mathtt{W}$ will generally have full rank, and there is not even an exact linear solution for $\mathcal{X}$.

Thus for $m \geq 4$, we require another method to produce a solution which satisfies the cubic constraint $S = 0$. The problem is to perform a "manifold projection" of the least-squares solution to $\mathtt{W}\mathcal{X} = \mathbf{0}$ onto the constraint manifold, but in a non-Euclidean space with the usual associated problem that we don't know in which direction to project. We will now give a novel solution to this algebraic problem.

Algebraic Error. An (over)determined linear system of equations is often solved using Singular Value Decomposition, by taking as null-vector the singular vector with the smallest singular value. The justification for this is that the SVD elicits the "directions" of space in which the solution is well determined (large singular values) and those in which it is poorly determined (small singular values). Taking the singular vector with smallest singular value is the usual "linear" solution, but as pointed out, it does not in general lie on $\mathtt{S}$. However, there may still be some information left in the second-smallest singular vector, and taking the space spanned by the two smallest singular vectors gives a line in $\mathbb{P}^4$, which passes through the "linear" solution and must also intersect $\mathtt{S}$ in three points (S is cubic). We use these three intersections as our candidates for $\mathcal{X}$. Since they lie exactly on $\mathtt{S}$, recovering their preimages $\mathbf{X}$ under ψ is not a problem.

Geometric Error. In each image, fitting error is the distance from the reprojected point $\mathbf{y} = \mathtt{P}\mathbf{X}$ to the measured image point $\mathbf{x} = (u, v, 1)^\top$. The reprojected point will depend both on the position of the sixth world point and on the choice of camera in the pencil for that image. But for a given world point $\mathbf{X}$, and choice of camera $\mathtt{P} = \mu\mathtt{A} + \nu\mathtt{B}$ in the pencil, the residual is the 2D image vector from $\mathbf{x}$ to the point $\mathbf{y} = \mathtt{P}\mathbf{X} = \mu\mathtt{A}\mathbf{X} + \nu\mathtt{B}\mathbf{X}$ on the line $\mathbf{l}$ joining $\mathtt{A}\mathbf{X}$ and $\mathtt{B}\mathbf{X}$. The optimal choice of μ, ν for given $\mathbf{X}$ is thus easy to deduce; it must be such as to make $\mathbf{y}$ the perpendicular projection of $\mathbf{x}$ onto this line (figure 2). What this means is that explicit minimization over camera parameters is unnecessary and so only the 3 degrees of freedom for $\mathbf{X}$ remain.

Due to the cross-product, the components $l_i(\mathbf{X})$ of the line $\mathbf{l}(\mathbf{X}) = \mathtt{A}\mathbf{X} \times \mathtt{B}\mathbf{X}$ are expressible as homogeneous quadratic functions of $\mathbf{X}$, and we note that these are expressible as linear functions of $\mathcal{X} = \psi(\mathbf{X})$. This is because the quadratic function $\mathtt{A}\mathbf{X} \times \mathtt{B}\mathbf{X}$ vanishes at each $\mathbf{E}_i$ and so, as was noted in section 3.2 (in particular, equation (7)), has the form derived earlier for such quadrics. Thus:

$$\mathbf{l}(\mathbf{X}) = \mathtt{A}\mathbf{X} \times \mathtt{B}\mathbf{X} = \begin{pmatrix} l_1(\mathbf{X}) \\ l_2(\mathbf{X}) \\ l_3(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} \cdots \mathbf{q}_1 \cdots \\ \cdots \mathbf{q}_2 \cdots \\ \cdots \mathbf{q}_3 \cdots \end{pmatrix} \mathcal{X}$$

for some 3×5 matrix with rows $\mathbf{q}_i$ whose coefficients can be determined from those of $\mathtt{A}$ and $\mathtt{B}$. If the sixth image point is $\mathbf{x} = (u, v, 1)^\top$ as before, then the squared geometric image residual becomes:

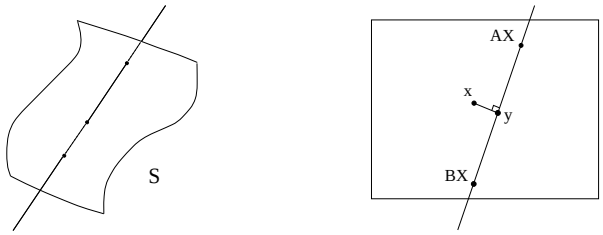


Fig. 2. Left: The diagram shows a line in 3-space intersecting a surface of degree 3. In the case of a line in 4-space and a hyper-surface of degree 3, the number of intersections is also 3. Right: Minimizing reprojection in the reduced model. For a given $\mathbf{X}$, the best choice $\mathtt{P} = \mu\mathtt{A} + \nu\mathtt{B}$ of camera in the pencil corresponds to the point $y = \mu\mathtt{A}\mathbf{X} + \nu\mathtt{B}\mathbf{X}$ on the line closest to the measured image point $\mathbf{x}$. Hence the image residual is the vector joining $\mathbf{x}$ and $\mathbf{y}$.

$$|d(\mathbf{x}, \mathbf{l}(\mathbf{X}))|^2 = \frac{|u\ l_1(\mathbf{X}) + v\ l_2(\mathbf{X}) + l_3(\mathbf{X})|^2}{|l_1(\mathbf{X})|^2 + |l_2(\mathbf{X})|^2} = \frac{|u\ \mathbf{q}_1\mathcal{X} + v\ \mathbf{q}_2\mathcal{X} + \mathbf{q}_3\mathcal{X}|^2}{|\mathbf{q}_1\mathcal{X}|^2 + |\mathbf{q}_2\mathcal{X}|^2} \tag{11}$$

and this is the geometric error (summed over each image) which must be minimized over $\mathbf{X}$. We can now compare the algebraic cost to the geometric cost. The algebraic error minimized is $\mathtt{W}\mathcal{X}$, which corresponds to summing an algebraic residual $|(u\ \mathbf{q}_1 + v\ \mathbf{q}_2 + \mathbf{q}_3)\mathcal{X}|^2$ over each image. Thus, the algebraic cost neglects the denominator of the geometric cost (11).

Invariance of Algebraic Error. As we have presented the algorithm so far, there is an arbitrary choice of scale for each quadric $Q_{\mathtt{A},\mathtt{B}}$, corresponding to the arbitrariness in the choice of representation $[\mathtt{A}, \mathtt{B}]$ of the pencil of cameras, the scale of which depends on the scale of $\mathtt{A}, \mathtt{B}$. Which normalization is used matters, and we address that issue now.

Firstly, by translating coordinates, we may assume that the sixth point is at the origin. The assumption $u, v = 0$ on the position of the sixth image point makes our method invariant to translations of image coordinates. It is desirable that the normalization should be invariant to scaling and rotation as well since these are the transformations which preserve our error model (isotropic Gaussian noise, see below). Our choice of normalization is most simply described by introducing a dot product similar to the Frobenius inner product $(\mathtt{A}, \mathtt{B})_{\text{Frob}} = \mathbf{trace}(\mathtt{A}^\top\mathtt{B}) = \sum_{ij} A_{ij}B_{ij}$. Our inner product simply leaves out the last row:

$$(\mathtt{A}, \mathtt{B})_{\text{Frob}} = \sum_{\substack{i=1,2,3 \\ j=1,2,3,4}} A_{ij}B_{ij} \qquad (\mathtt{A}, \mathtt{B})_* = \sum_{\substack{i=1,2 \\ j=1,2,3,4}} A_{ij}B_{ij}$$

The normalization we use can now be described by saying that the choice of basis of the pencil $[\mathtt{A}, \mathtt{B}]$ must be an orthonormal basis with respect to $(\cdot, \cdot)_*$. To achieve this, one could start with any basis of the pencil and use the Gram-Schmidt algorithm to orthonormalize them. It can be shown that with this normalization the algebraic error is invariant to scaling and rotation of the image coordinate system.

3.4 Algorithm Summary

It has been demonstrated how to pass from $m \geq 3$ views of six points in the world to a projective reconstruction in a few steps. These are:

1. Compute, for each of m views, the pencil of cameras which map the five standard basis points in the world to the first five image points, using the recommended normalization to achieve invariance to image coordinate changes.
2. Form from each pencil [A,B] the quadric constraint on the sixth world point $\mathbf{X}$ as described in section 3.2. i.e. form (7) $\mathbf{w}^{j\top}\mathcal{X} = 0$ in the five entries of $\mathcal{X}$.
3. Collect together the $\mathbf{w}^{j\top}$ as the rows of a $m \times 5$ matrix W.
4. Obtain the singular vectors corresponding to the two smallest singular values of W via the SVD. Let these be $\mathcal{X}_1$ and $\mathcal{X}_2$.
5. The solution $\mathcal{X}$ lies in the one-parameter family $\mathcal{X} = \alpha\mathcal{X}_1 + \beta\mathcal{X}_2$.
6. By expanding out the constraint (10), form a homogeneous cubic equation in α and β. There will be either one or three real solutions.
7. Once $\mathcal{X}$ is computed (satisfying the cubic constraint (10)), solve for $\mathbf{X}_6$ from the null-space of (9).
8. (optional) Minimize reprojection error (11) over the 3 degrees of freedom in the position of $\mathbf{X}_6$.

In practice, for a given set of six points, the quality of reconstruction can vary depending on which point is last in the basis. We try all six in turn and choose the best one.

Related Work. Yan *et al* [24] describe a linear method for reconstruction from $m \geq 4$ views of six points. Both our method and theirs turn the set of m quadratic equations in $\mathbf{X}$ into a set of m linear equations in some auxiliary variables ($\mathcal{X}$ here), and then impose constraints on a resulting null-space. There are two problems with their method when measurement error is present: first, their solution may not satisfy (both) the constraints on the auxiliary variable and second, their method uses projectively transformed image coordinates, and so potentially suffers from the bias described in section 2.

3.5 Results I

We have computed cameras which map the first five points exactly to their measured image points, and then minimize either an algebraic or geometric error on the sixth point. As discussed in the introduction, the Maximum Likelihood Estimate of the reconstruction, assuming isotropic Gaussian measurement noise, is obtained by bundle adjustment in which reprojection error (squared geometric image residual) is minimized over all points $\mathbf{X}_i$ and all cameras. This will is the optimal reconstruction. Minimizing error on only the sixth image point is thus a sub-optimal method.

We now give results on synthetic and real image sequences of 6 points in m views. The objective is to compare the performance of four algorithms:

Quasi-linear: minimizes algebraic error on sixth point only (as in the algorithm above).

Sub-optimal: minimizes reprojection error on the sixth point only (as in (11)) by optimizing over $\mathbf{X}_6$. This is a three parameter optimization problem. It is initialized by the quasi-linear algorithm.

Factorization: a simple implementation of projective factorization (the projective depths are initialized as all 1s and ten iterations performed).

Bundle Adjustment: minimizes reprojection error for all six points (varying both cameras and points $\mathbf{X}_i$). This is a $11m + 18$ parameter optimization problem for m views and six points. For synthetic data, it is initialized by whichever of the above three gives the smallest reconstruction error. For real data, it is initialized with the sub-optimal algorithm.

The three performance measures used are reprojection error, reconstruction error (the registration error between the reconstruction and ground truth), and stability (the algorithm converges). The claim is that the quasi-linear algorithm performs as well as the more expensive variants and can safely be used in practice.

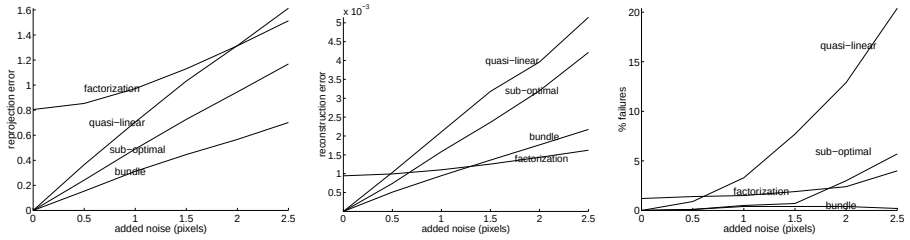


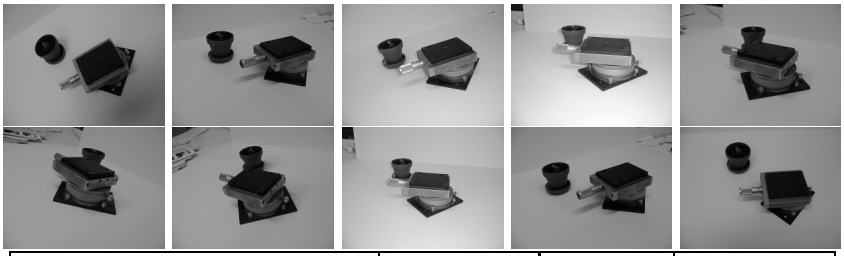
Fig. 3. Summary of experiments on synthetic data. 1000 data sets were generated randomly (7 views of 6 points) and each algorithm tried on each data set. Left: For each of the four estimators (quasi-linear, sub-optimal, factorization and bundle adjustment), the graph shows the average rms reprojection error over all 1000 data sets. Middle: the average reconstruction error, for each estimator, into the ground truth frame. Right: the average number of times each estimator failed (i.e. gave a reprojection error greater than 10 pixels).

Synthetic data. We first show results of testing the algorithm on synthetic data with varying amounts of pixel localisation noise added; our noise model is isotropic Gaussian with standard deviation σ. For each value of σ, the algorithm is run on 1000 randomly generated data sets. Each data set is produced by choosing six world points at random uniformly in the cube $[-1, +1]^3$ and six cameras with centres between 4 and 5 units from the origin and principal rays passing through the cube. After projecting each point under each chosen camera, artificial noise is added. The images are 512×512, with square pixels, and the principal point is at the centre of the image. Figure 3 summarizes the results.

The "failures" refer to reconstructions for which some reprojection error exceeded 10 pixels. The quality of reconstruction degrades gracefully as the noise

is turned up from the slightly optimistic 0.5 to the somewhat pessimistic 2.5; the rms and maximum reprojection error are highly correlated, with correlation coefficient 0.999 in each case (which may also be an indicator of graceful degradation).

Real Data. The image sequence consists of 10 colour images (JPEG, 768×1024) of a turntable, see figure 4. The algorithms from before, except factorization, are compared on this sequence and the results tabulated also in figure 4. Points were entered and matched by hand using a mouse (estimated accuracy is 2 pixels standard deviation). Ground truth is obtained by measuring the turntable with vernier calipers, and is estimated to be accurate to $0.25mm$. There were 9 tracks, all seen in all views. Of course, in principle any six tracks could be used to compute a projective reconstruction, but in practice some bases are much better than others. Examples of poor bases include ones which are almost coplanar in the world or which have points very close together.



	basis residuals (pixels)	all residuals (pixels)	reconstruction error (mm)
6 points quasi-linear	0.363 /2.32	0.750/2.32	0.467/0.676
6 points sub-optimal	0.358 /2.33	0.744/2.33	0.424/0.596
6 points bundle adjustment	0.115 /0.476	0.693/2.68	0.405/0.558
All points (and cameras) bundled	0.334 /0.822	0.409/1.08	0.355/0.521

Fig. 4. Results for the 9 tracks over the 10 turntable images. The reconstruction is compared for the three different algorithms, residuals (reported as rms/max) are shown for the 6 points which formed the basis (first column) and for all reconstructed points taken as a whole (second column). The last row shows the corresponding residuals after performing a full bundle adjustment.

Bundle adjustment achieves the smallest reprojection error over all residuals, because it has greater freedom in distributing the error. Our method minimizes error on the sixth point of a six point basis. Thus it is no surprise that the effect of applying bundle adjustment to *all* points is to increase the error on the basis point (column 1) but to decrease the error over all points (column 2). These figures support our claim that the quasi-linear method gives a very good approximation to the optimized methods. Figure 5 shows the reprojected reconstruction in a representative view of the sequence.

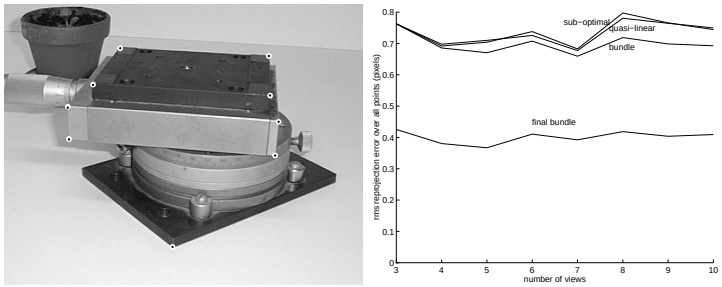


Fig. 5. Left: Reprojected reconstruction in view 3. The large white dots are the input points, measured from the images alone. The smaller, dark points are the reprojected points. Note that the reprojected points lie very close to the centre of each white dot. The reconstruction is computed with the 6-point sub-optimal algorithm. Right: The graph shows for each algorithm, the rms reprojection error for all 9 tracks as a function of the number of views used. For comparison the corresponding error after full-bore bundle adjustment is included.

4 Robust Reconstruction Algorithm

In this section we describe a robust algorithm for reconstruction built on the 6-point engine of section 3. The input to the algorithm is a set of point tracks, some of which will contain mismatches. Robustness means that the algorithm is capable of rejecting mismatches, using the RANSAC [4] paradigm. It is a straightforward generalization of the corresponding algorithm for 7 points in 2 views [20, 25] and 6 points in 3 views [1, 21].

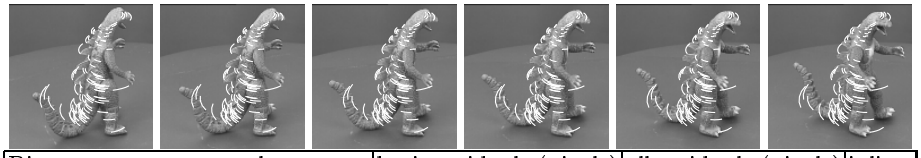
Algorithm Summary. The input is a set of measured image projections. A number of world points have been tracked through a number of images. Some tracks may last for many images, some for only a few (i.e. there may be missing data). There may be mismatches. Repeat the following steps as required:

1. From the set of tracks which appear in all images, select six at random. This set of tracks will be called a *basis*.
2. Initialize a projective reconstruction using those six tracks. This will provide the world coordinates (of the six points whose tracks we chose) and cameras for all the views (either quasi-linear or with 3 degrees of freedom optimization on the sixth point – see below).
3. For all remaining tracks, compute optimal world point positions using the computed cameras by minimizing the reprojection error over all views in which the point appears. This involves a numerical minimization.
4. Reject tracks whose image reprojection errors exceed a threshold. The number of tracks which pass this criterion is used to score the reconstruction.

The justification for this algorithm is, as always with RANSAC, that once a "good" basis is found it will (a) score highly and (b) provide a reconstruction against which other points can be tested (to reject mismatches).

4.1 Results II

The second sequence is of a dinosaur model rotating on a turntable (figure 6). The image size is 720×576. Motion tracks were obtained using the fundamental matrix based tracker described in [5]. The robust reconstruction algorithm is applied using 100 samples to the subsequence consisting of images 0 to 5. For these 6 views, there were 740 tracks of which only 32 were seen in all views. 127 tracks were seen in 4 or more views. The sequence contains both missing points and mismatched tracks.



Dinosaur sequence results	basis residuals (pixels)	all residuals (pixels)	inliers
6 points quasi-linear	0.0443/0.183	0.401/1.24	95
6 points sub-optimal	0.0443/0.183	0.401/1.24	95
6 points bundle adjustment	0.0422/0.127	0.383/1.181	97
All points (and cameras) bundled	0.313 /0.718	0.234/0.925	95

Fig. 6. The top row shows the images and inlying tracks used from the dinosaur sequence. The table in the bottom row summarizes the result of comparing the three different fitting algorithms (quasi-linear, sub-optimal, bundle adjustment). There were 6 views. For each mode of operation, the number of points marked as inliers by the algorithm is shown in the third column. There were 127 tracks seen in four or more views.

For the six point RANSAC basis, a quasi-linear reconstruction was rejected if any reprojection error exceeded 10 pixels, and the subsequent 3 degrees of freedom sub-optimal solution was rejected if any reprojection error exceeded a threshold of 5 pixels. These are very generous thresholds and are only intended to avoid spending computation on very bad initializations. The real criterion of quality is how much support an initialization has. When backprojecting tracks to score the reconstruction, only tracks seen in 4 or more views were used and tracks were rejected as mismatches if any residual exceed 1.25 pixels after back-projection.

The algorithms of section 3.5 (except factorization) are again compared on this sequence. The errors are summarized in figure 6. The last row shows an additional comparison where bundle adjustment is applied to all the points and cameras of the final reconstruction. Figure 6 also shows the tracks accepted by the algorithm. Figure 7 shows the computed model.

Remarks entirely analogous to the ones made about the previous sequence apply to this one, but note specifically that optimizing makes very little difference to the residuals. This means that the quasi-linear algorithm performs almost as well as the sub-optimal one. Applying bundle adjustment to each initial 6-point reconstruction improves the fit somewhat, but the gain in accuracy and support is rather small compared to the extra computational cost (in this example, there was a 7-fold increase in computation time).

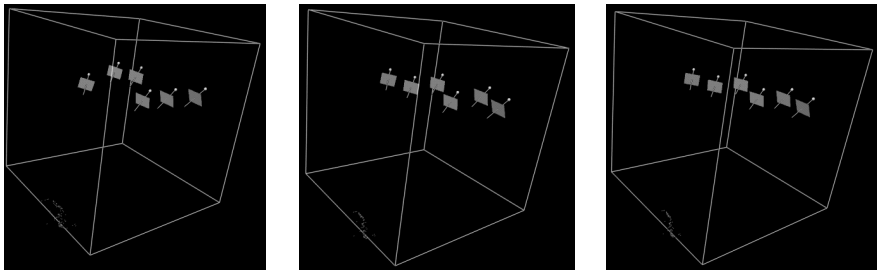


Fig. 7. Dinosaur sequence reconstruction: a view of the reconstructed cameras (and points). Left: quasi-linear model, cameras computed from just 6 tracks. Middle: after resectioning the cameras using the computed structure. Right: after bundle adjustment of all points and cameras (the unit cube is for visualization only).

The results shown for view 0 to 5 are typical of results obtained for other segments of 6 consecutive views from this sequence. Decreasing the number of views used has the disadvantage of narrowing the baseline, which generally leads to both structure and cameras being less well determined. The advantage of using only a small number of points (i.e. 6 instead of 7) is that there is a higher probability that sufficient tracks will exist over many views.

5 Discussion

Algorithms have been developed which estimate a six point reconstruction over m views by a quasi-linear or sub-optimal method. It has been demonstrated that these reconstructions provide cameras which are sufficient for a robust reconstruction of $n > 6$ points and cameras over m views from tracks which include mismatches and missing data. This reconstruction can now form the basis of a hierarchical method for extended image sequences. For example, the hierarchical method in [5], which builds a reconstruction from image triplets, could now proceed from extended sub-sequences over which at least six points are tracked.

We are currently investigating whether the efficient 3 degree of freedom parametrization of the reconstruction can be extended to other multiple view cases, for example seven points over m views.

Acknowledgements We are grateful for the tracked dinosaur points provided by Andrew Fitzgibbon. This work was supported by an EPSRC studentship and EC Esprit Project Improofs. Thanks also to Nicolas Dano for coding and running the baseline duality-based method.

References

1. P. Beardsley, P. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *Proc. ECCV*, LNCS 1064/1065, pages 683–695. Springer-Verlag, 1996.
2. S. Carlsson. Duality of reconstruction and positioning from projective views. In *IEEE Workshop on Representation of Visual Scenes, Boston*, 1995.

3. S. Carlsson and D. Weinshall. Dual computation of projective shape and camera positions from multiple images. *IJCV*, 1998. in Press.
4. M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.
5. A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. ECCV*, pages 311–326. Springer-Verlag, Jun 1998.
6. R. I. Hartley. Euclidean reconstruction from uncalibrated views. In J.L. Mundy, A. Zisserman, and D. Forsyth, editors, *Proc. 2nd European-US Workshop on Invariance, Azores*, pages 187–202, 1993.
7. R. I. Hartley. Projective reconstruction and invariants from multiple images. *IEEE T-PAMI*, 16:1036–1041, October 1994.
8. R. I. Hartley. Multilinear relationships between coordinates of corresponding image points and lines. In *Proceedings of the Sophus Lie Symposium, Nordfjordeid, Norway (*not published yet*)*, 1995.
9. R. I. Hartley. Minimizing algebraic error. *Phil. Trans. R. Soc. Lond. A*, 356(1740):1175–1192, 1998.
10. A. Heyden. Projective structure and motion from image sequences using subspace methods. In *Scandinavian Conference on Image Analysis, Lappenraanta*, 1997.
11. D. Jacobs. Linear fitting with missing data: Applications to structure from motion and to characterizing intensity images. In *Proc. CVPR*, pages 206–212, 1997.
12. S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *Proc. CVPR*, 2000.
13. S. J. Maybank and A. Shashua. Ambiguity in reconstruction from images of six points. In *Proc. ICCV*, pages 703–708, 1998.
14. P. F. McLauchlan and D. W. Murray. A unifying framework for structure from motion recovery from image sequences. In *Proc. ICCV*, pages 314–320, 1995.
15. L. Quan. Invariants of 6 points from 3 uncalibrated images. In J. O. Eckland, editor, *Proc. ECCV*, pages 459–469. Springer-Verlag, 1994.
16. L. Quan, A. Heyden, and F. Kahl. Minimal projective reconstruction with missing data. In *Proc. CVPR*, pages 210–216, Jun 1999.
17. I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *IJCV*, 18(1):41–60, 1996.
18. P. Sturm and W. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. ECCV*, pages 709–720, 1996.
19. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2):137–154, Nov 1992.
20. P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.
21. P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.
22. P. H. S. Torr and A. Zisserman. Robust computation and parameterization of multiple view relations. In *Proc. ICCV*, pages 727–732, Jan 1998.
23. W. Triggs. Factorization methods for projective structure and motion. In *Proc. CVPR*, pages 845–851, 1996.
24. X. Yan, X. Dong-hui, P. Jia-xiong, and D. Ming-yue. The unique solution of projective invariants of six points from four uncalibrated images. *Pattern Recognition*, 30(3):513–517, 1997.
25. Z. Zhang, R. Deriche, O. D. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, 1995.