

A Six Point Solution for Structure and Motion

F. Schaffalitzky¹, A. Zisserman¹, and R. I. Hartley²

¹ Robotics Research Group, Oxford, UK

² G.E. CRD, Schenectady, NY

Abstract. This paper extends the set of minimal reconstruction algorithms. The paper has two main contributions: The first is a “quasi-linear” method for computing structure and motion for $m \geq 3$ views of 6 points. It is shown that an algebraic image error over all views may be computed as the solution of a cubic in one variable. A minor point is that this enables a more concise algorithm, than any given previously, for the reconstruction of 3 views of 6 points. It is then shown that a geometric image error over all views can be minimized by a simple three parameter numerical optimization.

The second contribution is an m view n point robust reconstruction algorithm which uses the 6 point method as a search engine. The algorithm can cope with missing data and mis-matched 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 correspondences tracked through image sequences. Their characteristics vary from the so-called *minimal* methods [17, 18, 24] which work with the least data necessary to compute structure and motion, through intermediate methods [5, 20] which may perform mis-match (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 2 view 7 point solution is used in the RANSAC estimation of the fundamental matrix in [24], and the 3 view 6 point solution in the RANSAC estimation of the trifocal tensor in [23]. It would seem natural then to use a minimal solution as a search engine in 4 or more views. The problem is that in 4 or more views a solution is forced to include a minimization to account for measurement error (noise). In the ‘2 view 7 point’ and ‘3 view 6 point’ cases there are the same number of measurement constraints as degrees of freedom in the tensor. In both cases 1 or 3 real solutions result (and the duality explanation for this equivalence was given by [3]). However, in four views six points provide one more constraint than the number of degrees of freedom in the four view geometry (the quadrifocal tensor). This means than 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 view case. Instead it is necessary to minimize a measurement error whether algebraic or geometric. The poor estimate which results by using an approach based on minimizing algebraic distance and a standard projective basis for the image is described and demonstrated in section 2.

Here we develop a novel quasi-linear solution for the 6 point $m \geq 3$ case. This solution involves only a SVD and the evaluation of a cubic polynomial in a single variable. This is described in section 3. We also describe a sub-optimal method (compared to bundle-adjustment) which minimizes geometric error at the cost of only a 3 parameter minimization.

A second part of the paper describes (yet another) 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 mis-matches.** 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 [10] 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 mis-matches. However, it is expensive to carry out and most significantly requires a good initial estimate.

In the special case of affine cameras, factorization methods [21] minimize reprojection error [19] and so give the optimal solution found by bundle adjustment. However, factorization cannot cope with mis-matches, and methods to overcome missing data [14] 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 [13, 20]. The problems of missing data and mis-matches remain though.

Bundle-adjustment will almost always be the final step of a reconstruction algorithm. However, achieving good sub-optimal estimates prior to bundle-adjustment is necessary for the latter to be effective (fewer iterations, and less likely to converge to local minimum.) Current methods of initializing a bundle-adjustment include factorization, hierarchical combination of sub-sequences [5], and the Variable State Dimension Filter (VSDF) [16].

In this paper we describe a novel algorithm for computing a reconstruction satisfying the 3 basic objectives above (optimal, missing data, mismatches). It is based on using the 6-pt 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

A method suggested by Carlsson and Weinshall [2, 3] for reconstruction from three views involves a certain duality between points and cameras. 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 $\mathbf{E}_1, \dots, \mathbf{E}_4$, the camera matrix may be seen to be of the form

$$\mathbf{P} = \begin{bmatrix} a_i & -d_i \\ b_i & -d_i \\ c_i & -d_i \end{bmatrix} \quad (1)$$

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

$$\begin{bmatrix} a & -d \\ b & -d \\ c & -d \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ T \end{pmatrix} = \begin{bmatrix} x & -T \\ y & -T \\ z & -T \end{bmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \quad (2)$$

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 method is the distortion of the image measurement error distributions by the projective image mapping as illustrated in figure 1. A circular Gaussian distribution is transformed by a projective transformation to a distribution that is no longer circular, and not even Gaussian. Common methods of two-view reconstruction are not able to handle such error distributions effectively. 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. This is illustrated in figure 1. The method used for reconstruction from the transformed data was a dualization of one of the best methods available for two-view reconstruction ([9]) - an iterative method that minimizes algebraic error.

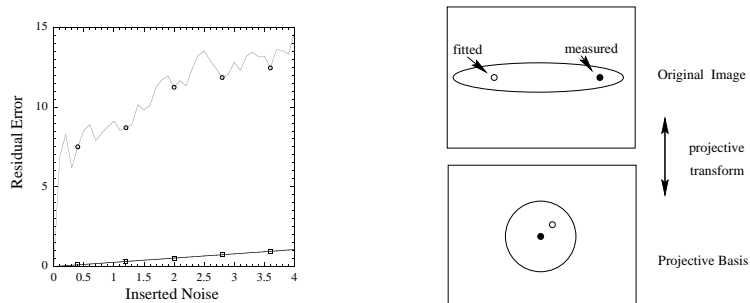


Fig. 1. Left: Residual error as a function of image noise for six points over 20 views. The upper curve is the result of a duality-based reconstruction algorithm, the lower is the result of bundle-adjustment. The method for generating this synthetic data is described in section 3.4. As may be seen the 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 Reconstruction from 6 points over m views

This section describes the main algebraic development of the 6 point method. In essence it is quite similar to the development given by Hartley [11] and Quan [17] 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 6 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 in the first instance.

3.1 A pencil of cameras

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

$$\mathbf{x} \times \mathbf{P}\mathbf{X} = \mathbf{0} \quad (3)$$

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 \mathbf{P} which may be written $\mathbf{M}\mathbf{p} = \mathbf{0}$, where \mathbf{M} is a 10×12 matrix formed from two of the linear equations (3) of each point correspondence, and \mathbf{p} is \mathbf{P} written as a 12-vector. This system of equations has a 2-dimensional null-space and thus results in a pencil of cameras.

Suppose that the five world points are the points of the standard projective frame $\mathbf{E}_1, \dots, \mathbf{E}_5$, so that both \mathbf{X}_i and \mathbf{x}_i ($i = 1, 2, 3, 4, 5$) are now known. Then the null-space

of \mathbf{M} can immediately be computed, and will be denoted from here on by the basis of 3×4 matrices $[\mathbf{A}, \mathbf{B}]$. Then for any choice of the scalars $(s : t) \in \mathbb{P}^1$ the camera in the pencil $\mathbf{P} = s\mathbf{A} + t\mathbf{B}$ exactly projects the standard projective basis to the first five points.

Each camera \mathbf{P} in the pencil has its optical centre located as the null-vector of \mathbf{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 points of the standard projective basis. 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 \mathbf{P} mapping a set of point $\mathbf{X}_1, \dots, \mathbf{X}_6$ to image points $\mathbf{x}_1, \dots, \mathbf{x}_6$. Let $[\mathbf{A}, \mathbf{B}]$ be the pencil of cameras consistent with the projections of the first five points. Since \mathbf{P} lies in the pencil, there are scalars $(s : t) \in \mathbb{P}^1$ such that $\mathbf{P} = s\mathbf{A} + t\mathbf{B}$ and so the projection of the sixth world point \mathbf{X}_6 is $\mathbf{x}_6 = s\mathbf{A}\mathbf{X}_6 + t\mathbf{B}\mathbf{X}_6$. This means that the three points $\mathbf{x}_6, \mathbf{A}\mathbf{X}_6, \mathbf{B}\mathbf{X}_6$ are collinear in the image, so

$$[\mathbf{x}_6, \mathbf{A}\mathbf{X}_6, \mathbf{B}\mathbf{X}_6] = 0 \quad ,$$

which is a quadratic constraint on \mathbf{X}_6 . The 3×3 determinant can be expressed as $\mathbf{X}_6^\top \mathbf{A}^\top [\mathbf{x}_6]_\times \mathbf{B} \mathbf{X}_6 = 0$. To summarize :

Let $[\mathbf{A}, \mathbf{B}]$ 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 surface given by

$$Q = (\mathbf{A}^\top [\mathbf{x}_6]_\times \mathbf{B}) \text{sym} = \mathbf{A}^\top [\mathbf{x}_6]_\times \mathbf{B} - \mathbf{B}^\top [\mathbf{x}_6]_\times \mathbf{A}^\top \quad . \quad (4)$$

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

In the particular case where the five points \mathbf{X}_i are the members \mathbf{E}_i of a projective basis, the condition $\mathbf{X}_i^\top Q \mathbf{X}_i = 0$ allows us to specify the form of Q simply. From $\mathbf{E}_i^\top Q \mathbf{E}_i = 0$ for $i = 1, \dots, 4$, 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

$$Q = \begin{bmatrix} 0 & w_1 & w_2 & -\Sigma \\ w_1 & 0 & w_3 & w_4 \\ w_2 & w_3 & 0 & w_5 \\ -\Sigma & w_4 & w_5 & 0 \end{bmatrix} \quad (5)$$

where $\Sigma = w_1 + w_2 + w_3 + w_4 + w_5$. Let $\mathbf{X}_6 = (p, q, r, s)^\top$ be a point lying on Q . The equation $\mathbf{X}_6^\top Q \mathbf{X}_6$ may be written in a vector form as

$$(w_1, w_2, w_3, w_4, w_5) \begin{pmatrix} pq - ps \\ pr - ps \\ qr - qs \\ qs - ps \\ rs - ps \end{pmatrix} = 0 \quad (6)$$

or more briefly, $\mathbf{W}\mathcal{X} = 0$, where \mathcal{X} is the column vector in (6).

Solving for the point \mathbf{X} . Now consider m views of 6 points and suppose again that the first five world points are in the known positions $\mathbf{E}_1, \dots, \mathbf{E}_5$. 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 $\mathbf{E}_1, \dots, \mathbf{E}_5$ 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 [15].

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 $\mathbf{W}^j \mathcal{X} = 0$ on the 5-vector \mathcal{X} defined in terms of \mathbf{X}_6 by equation (6). The vector \mathbf{W}^j is obtained from the coefficients of the quadric Q^j (see below). The basic method is to solve for $\mathcal{X} = (a, b, c, d, e)^\top \in \mathbb{P}^4$ 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 $\psi : \mathbb{P}^3 \rightarrow \mathbb{P}^4$, given by

$$\psi : \mathbf{X} = \begin{pmatrix} p \\ q \\ r \\ s \end{pmatrix} \rightarrow \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} pq - ps \\ pr - ps \\ qr - ps \\ qs - ps \\ rs - ps \end{pmatrix}$$

which is a (rational) transformation from \mathbb{P}^3 to \mathbb{P}^4 , and maps the quadric $Q \subset \mathbb{P}^3$ into the hyperplane defined in \mathbb{P}^4 by

$$w_1 a + w_2 b + w_3 c + w_4 d + w_5 e = 0 \quad (7)$$

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

Computing \mathbf{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 :

$$\begin{pmatrix} e-d & 0 & 0 & a-b \\ e-c & 0 & a & 0 \\ d-c & b & 0 & 0 \\ 0 & e-b & a-d & 0 \\ 0 & e & 0 & a-c \\ 0 & 0 & d & b-c \end{pmatrix} \quad (8)$$

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 nullity 0 and more care has to be taken to recover a meaningful \mathbf{X} .

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 \mathbf{S} cut out by the cubic equation

$$S(a, b, c, d, e) = abd - abe + ace - ade - bcd + bde = \begin{vmatrix} e & e & b \\ d & c & b \\ d & a & a \end{vmatrix} = 0 \quad (9)$$

This can be verified by direct substitution. Alternatively it can be derived by observing that all 4×4 subdeterminants of (8) 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 \mathbf{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 (7) cut out a line in \mathbb{P}^4 . The line intersects \mathbf{S} in three points (generically) (see figure 2). Thus there are three solutions for \mathbf{X} . This is a well-known [17] minimal solution. Our treatment gives a simpler (than the Quan [17] or Carlsson and Weinshall [3]) algorithm for computing a trifocal tensor from six points (from a projective reconstruction) because it does not involve 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 (6) in the five entries of \mathcal{X} . Since this is a homogeneous set of equations, the scale is immaterial.
2. Obtain a set of solutions of the form $\mathcal{X} = s\mathcal{X}_1 + t\mathcal{X}_2$ 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 (9), form a homogeneous cubic equation in s and t . There will be either one or three real solutions.
4. Once \mathcal{X} is computed (satisfying the cubic constraint (9)), solve for $\mathbf{X}_6 = (p, q, r, s)^\top$

Four or more views. In this case the linear constraints from the hyperplanes alone will (generally) determine a unique solution for \mathcal{X} . In the presence of noise, though, this solution will not satisfy the cubic constraint. That is, it does not lie on \mathbf{S} ; its coordinates do not satisfy $S = 0$. We would like to coerce it to do so. The problem is to perform a “manifold projection” 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 problem.

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 \mathbf{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 \mathbf{S} in three points (S is cubic). We use these three intersections as our candidates for \mathcal{X} . Since they lie exactly on \mathbf{S} , recovering their preimages \mathbf{X} under ψ is not a problem.

This, then, is our heuristic. We overcome our manifold projection problems by projecting in the direction of the singular vector with second-smallest singular value. Note that in the case of 4 views, the smallest singular value will actually be 0.

3.3 Minimizing reprojection error

The previous sub-section has described a quasi-linear method involving the following two steps: first, a linear SVD decomposition of a matrix composed of one hyperplane

from each view; second, intersecting the line in \mathbb{P}^4 (computed from two of the singular vectors) with a cubic surface.

The best use of the given data is full bundle adjustment, minimizing total squared image reprojection error over all camera and structure parameters. This optimization provides the Maximum Likelihood Estimate if the measurement noise is isotropic Gaussian.

In the current case, we have computed cameras which map the first five points exactly to their measured image points, and rather than jump directly to bundle adjustment, an intermediate case is to minimize total squared reprojection error for the sixth point \mathbf{X} over \mathbb{P}^3 . This fits in the middle of a spectrum of possible estimates (degrees of freedom shown in brackets) :

1. **Algebraic fit.** The quasi-linear solution minimizes an “algebraic” error by a direct least squares fit on homogeneous coordinates in \mathbb{P}^4 .
2. **Sub-optimal fit.** Minimizes total squared reprojection error for the sixth point over its position in \mathbb{P}^3 , mapping the first five points exactly [3].
3. **Optimal fit (bundle adjustment).** Minimizes total squared reprojection error for all points, over all structure and camera parameters [$11m + 3$].

The model fitted by the second item is clearly a reduced form of the model fitted by the third item. As will be shown below the cost of executing this reduced minimization is negligible (as the error can be parametrized efficiently by the 3 degrees of freedom of \mathbf{X}).

Geometric error. In each image, fitting error is the distance from the reprojected point $\mathbf{y} = \mathbf{P}\mathbf{X}$ 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 \mathbf{X} and on the choice of camera in the pencil for that image. But for a given world point \mathbf{X} , and choice of camera $\mathbf{P} = s\mathbf{A} + t\mathbf{B}$ in the pencil, the residual is the 2D image vector from \mathbf{x} to the point $\mathbf{y} = \mathbf{P}\mathbf{X} = s\mathbf{A}\mathbf{X} + t\mathbf{B}\mathbf{X}$ on the line \mathbf{l} joining $\mathbf{A}\mathbf{X}$ and $\mathbf{B}\mathbf{X}$. The optimal choice of s, t 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.

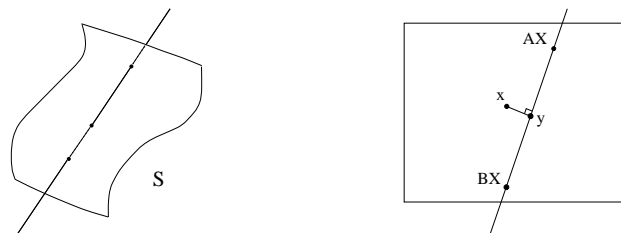


Fig. 2. Left: The diagram shows a line in 3-space intersecting a surface of degree 3. In the case of a line in 4-space and a hyper-surface of degree 3, the number of intersections is also 3. Right: Minimizing reprojection in the reduced model. For a given \mathbf{X} , the best choice $\mathbf{P} = s\mathbf{A}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

We have already seen that the components of the line $l(\mathbf{X}) = \mathbf{A}\mathbf{X} \times \mathbf{B}\mathbf{X}$ are expressible as quadrics in \mathbf{X} , and moreover as linear functions of $\mathcal{X} = \psi(\mathbf{X})$:

$$l(\mathbf{X}) = \mathbf{A}\mathbf{X} \times \mathbf{B}\mathbf{X} = \begin{pmatrix} \mathbf{q}_0 \psi(\mathbf{X}) \\ \mathbf{q}_1 \psi(\mathbf{X}) \\ \mathbf{q}_2 \psi(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} \cdots \mathbf{q}_0 \cdots \\ \cdots \mathbf{q}_1 \cdots \\ \cdots \mathbf{q}_2 \cdots \end{pmatrix} \mathcal{X}$$

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

$$d(\mathbf{x}, l(\mathbf{X}))^2 = \frac{|u\mathbf{q}_0\mathcal{X} + v\mathbf{q}_1\mathcal{X} + \mathbf{q}_2\mathcal{X}|^2}{|\mathbf{q}_0\mathcal{X}|^2 + |\mathbf{q}_1\mathcal{X}|^2}$$

and this is the geometric error which must be minimized in the sub-optimal scheme. Note that this form of the error is amenable to reweighted least squares because, given an initial estimate of \mathbf{X} , we can adjust the scale so as to make the denominator close to 1, while putting the numerator into a least squares problem. This expression shows that the minimization of image error over $\mathbf{X} \in \mathbb{P}^3$ can be carried out as a minimization over $\mathcal{X} \in \mathcal{S}$ instead.

Approximating geometric error. We now compare the first item with the second. The algebraic fitting algorithm which we propose consists of first forming the linear least squares problem which minimizes the sum of squares of $\mathbf{q}_2\mathcal{X}$ over the images. We intersect the 2D SVD nullspace with \mathcal{S} to impose constraints.

As we have presented the algorithm so far, there is an arbitrary choice of scale for each quadric $Q_{\mathbf{A},\mathbf{B}}$, corresponding to the arbitrariness in the choice of representation $[\mathbf{A}, \mathbf{B}]$ of the pencil of cameras (in terms of the equation above the algebraic fitting scheme neglects the denominator and just minimizes the residuals defined by the $u\mathbf{q}_0 + v\mathbf{q}_1 + \mathbf{q}_2$), the scale of which depends on the scale of \mathbf{A}, \mathbf{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. This amounts to (pre)multiplying \mathbf{A}, \mathbf{B} by a 3×3 translation homography and we assume this has been done (so $u, v = 0$ in the above derivation). Thus the geometric error we want to approximate is

$$\frac{|\mathbf{q}_2\mathcal{X}|^2}{|\mathbf{q}_0\mathcal{X}|^2 + |\mathbf{q}_1\mathcal{X}|^2}$$

Making this assumption on the position of the sixth image point means that the normalization is independent of (*ie* is 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). This requirement rules out many obvious candidates, like normalizing the Frobenius norms of \mathbf{A}, \mathbf{B} to 1 or normalizing \mathbf{q}_2 to unit norm.

To describe our choice of normalization, we introduce a dot product which is similar to the Frobenius inner product $(\mathbf{A}, \mathbf{B})_{\text{Frob}} = \text{trace}(\mathbf{A}^\top \mathbf{B})$. The Frobenius inner product can also be computed as the sum of $A_{ij}B_{ij}$ over all indices i, j . Our inner product is the same as the Frobenius inner product, except that the last row is left out :

$$(\mathbf{A}, \mathbf{B})_{\text{Frob}} = \sum_{\substack{i=0,1,2 \\ j=0,1,2,3}} A_{ij}B_{ij} \quad (\mathbf{A}, \mathbf{B})_* = \sum_{\substack{i=0,1 \\ j=0,1,2,3}} A_{ij}B_{ij}$$

The normalization we use can now be described by saying that the choice of basis of the pencil $[\mathbf{A}, \mathbf{B}]$ must be an orthonormal basis wrt $(\cdot, \cdot)_*$. To achieve this, one could start

with any basis of the pencil and use the Gram-Schmidt algorithm [6] to orthonormalize them. Alternatively, if the pencil is computed as the SVD nullspace of a suitably scaled 10×12 design matrix, the orthonormality will be automatic.

Scaling image coordinates corresponds to scaling the first two rows of the basis element matrices, which just scales our dot product (but not the Frobenius product). Rotating image coordinates corresponds to applying an orthogonal transformation to the first two rows of the basis elements, and this preserves our dot product. Finally, choosing a different orthonormal basis corresponds to a certain linear basis change in the pencil and the effect on the \mathbf{q}_i is a scaling by the determinant of that basis change. But that basis change must be orthogonal, so it has determinant 1.

Summary. It has been demonstrated how to pass from m views of six points in the world to a projective reconstruction in a few steps. The positions of the six world points as well as the camera for each view have been computed.

The reconstruction obtained is not the MLE (assuming isotropic Gaussian point localization noise), which optimally distributes measurement error over all the points, but an approximation which puts all the errors on the sixth point.

The steps of the algorithm are :

1. Compute, for each image, 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. and use the transformation $\psi : \mathbb{P}^3 \rightarrow \mathbb{P}^4$ to convert the quadric intersection problem to a hyperplane intersection problem. Use the SVD to compute a pencil of possible values for $\mathcal{X} = \psi(\mathbf{X})$.
3. Intersect that line with the cubic constraint $S = 0$ to get (up to) three solutions for $\mathcal{X} = \psi(\mathbf{X})$ satisfying the constraint and use (8) to recover values for the sixth point \mathbf{X} from \mathcal{X} .
4. (optional) Minimize reprojection error over the 3 degrees of freedom in the position of \mathbf{X} .

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.

3.4 Results I

We will now give results on synthetic and real image sequences of 6 points in m views. The objective is to compare the performance of three algorithms: quasi-linear; minimizing on the 6th point only; and, bundle adjustment. The three performance measures used are reprojection error, registration error to 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.

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 noise with standard deviation σ . For each value of σ , the algorithm is run on 100 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

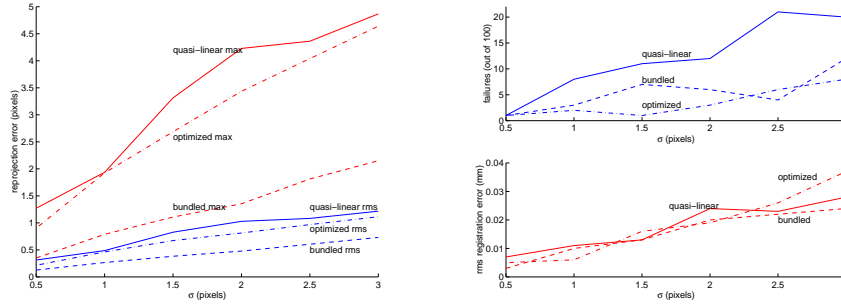


Fig. 3. Summary of experiments on synthetic data. Left: For each of the three estimators (quasi-linear, optimized and bundled), the graph shows the reprojection error (rms and maximum), averaged over 100 randomly generated data sets (6 views of 6 points). Obviously, the rms error is always below the corresponding maximum error. Top right: number of times each estimator failed. Bottom right: the average registration error, for each estimator, into the ground truth frame.

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. A more plausible error model would be isotropic Gaussian error clamped to a circle of radius, say, 2 pixels and indeed, if this modification is made all the failures disappear. The quality of reconstruction degrades gracefully as the noise is turned up from the slightly optimistic 0.5 to the somewhat pessimistic 3.0; 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 algorithm is tested on an image sequence consisting of 10 colour images (JPEG, 768×1024) of a turntable. The three algorithms are compared on this sequence and the results tabulated in figure 4 too, below the image sequence. 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 6 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.

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 is to increase the error in column 1 and to decrease the error in column 2. These figures support our claim that the quasi-linear method gives a very good approximation to the optimized method.

Figure 5 shows the projected reconstruction in a representative view of the sequence.

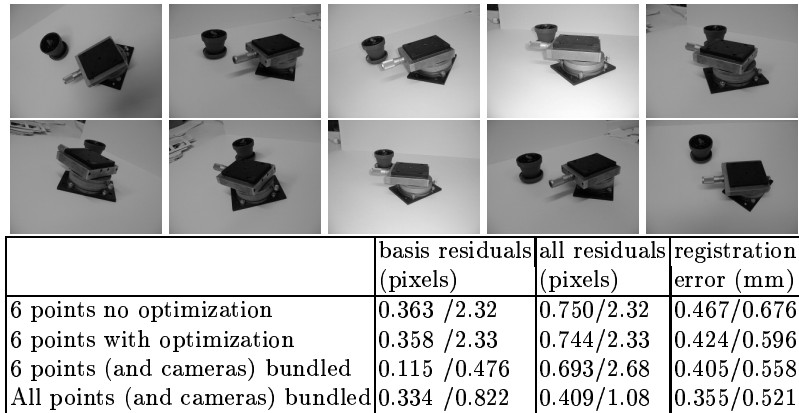


Fig. 4. Results for the 9 tracks over the 10 turntable images. The reconstruction is compared for the three different algorithms, residuals (reported as rms/max) are shown for the 6 points which formed the basis (first column) and for all reconstructed points taken as a whole (second column). The last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

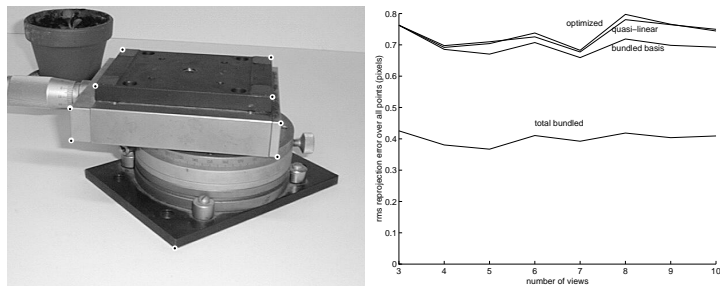


Fig. 5. Left: Reprojected reconstruction in view 3. The large white dots are the input points, measured from the images alone. The smaller, dark points are the reprojected points. Note that the reprojected points lie very close to the centre of each white dot. The reconstruction is computed with the 6-point algorithm, optimizing over the position of the sixth point. 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 [22, 25] and 6 points in 3 views [1, 23].

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 (*ie* 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 6th 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.

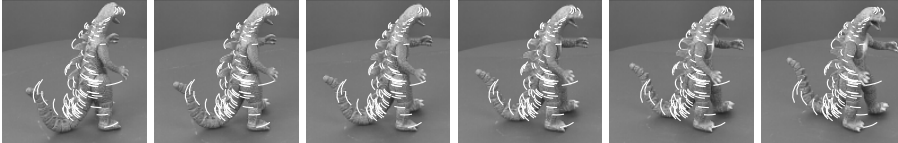
The justification for this algorithm is, as always with RANSAC, that once a “good” basis is found it will (a) score highly and (b) provide a reconstruction against which other points can be tested (to reject mismatches).

4.1 Results II

The second sequence is a turntable sequence (*ie* the camera motion is a turntable motion) of a dinosaur model (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 mis-matched tracks.

For the six point RANSAC basis, linear reconstructions were rejected if any reprojection error exceeded 10 pixels, and the subsequent 3 degrees of freedom optimization 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 backprojection.

Three variations on this algorithm are also compared here. The first is the pure quasi-linear algorithm. The second optimizes only the 6th point. The third applies a full bundle adjustment to the 6-point reconstructions. The errors are summarized in figure 6. The last row shows errors after applying bundle adjustment to the final



Dinosaur sequence results	basis residuals (pixels)	all residuals (pixels)	inlier count
6 points no optimization	0.0443/0.183	0.401/1.24	95
6 points with optimization	0.0443/0.183	0.401/1.24	95
6 points (and cameras) bundled	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. Comparing the three different fitting algorithms (algebraic, reduced, full). 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.

reconstruction (many points, many cameras). 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 no difference to the residuals at this level of precision (3 significant figures). 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).

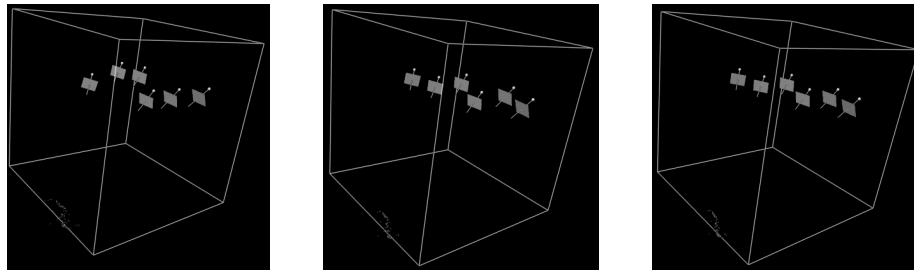


Fig. 7. Dinosaur sequence reconstruction : a view of the reconstructed cameras (and points). Left: quasi-linear model, cameras computed from just 6 tracks. Middle: after resectioning the cameras using the computed structure. Right: after complete full bundle adjustment (the unit cube is for visualization 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 6 point reconstruction over m views by a quasi-linear or sub-optimal method. It has been demonstrated that these recon-

structions provide cameras which are sufficient for a robust reconstruction of $n > 6$ points and cameras over m views from tracks which include mis-matches 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 6 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 7 points over m views.

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. G. H. Golub and C.F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, MD, second edition, 1989.
7. G.-M. Greuel, G. Pfister, and H. Schönemann. Singular version 1.2 user manual. In *Reports On Computer Algebra*, number 21 in Reports On Computer Algebra. Centre for Computer Algebra, University of Kaiserslautern, June 1998. <http://www.mathematik.uni-kl.de/~zca/Singular>
8. R. Hartley. Computation of the quadrfocal tensor. In *Proc. ECCV*, LNCS 1406, pages 20–35. Springer-Verlag, 1998.
9. R. Hartley. Minimizing algebraic error. volume 356, pages 1175–1192, 1998.
10. 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.
11. R. I. Hartley. Projective reconstruction and invariants from multiple images. *IEEE T-PAMI*, 16:1036–1041, October 1994.
12. R. I. Hartley. In defence of the 8-point algorithm. In *Proc. ICCV*, pages 1064–1070, 1995.
13. A. Heyden. Projective structure and motion from image sequences using subspace methods. In *Scandinavian Conference on Image Analysis, Lappenaanta, 1997*, 1997.
14. D. Jacobs. Linear fitting with missing data: Applications to structure from motion and to characterizing intensity images. In *Proc. CVPR*, pages 206–212, 1997.
15. S. J. Maybank and A. Shashua. Ambiguity in reconstruction from images of six points. In *Proc. ICCV*, pages 703–708, 1998.
16. 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.
17. L. Quan. Invariants of 6 points from 3 uncalibrated images. In J. O. Eckland, editor, *Proc. ECCV*, pages 459–469. Springer-Verlag, 1994.
18. L. Quan and F. K. A. Heyden. Minimal projective reconstruction with missing data. In *Proc. CVPR*, 1999.
19. I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *IJCV*, 18(1):41–60, 1996.

20. P. Sturm and W. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. ECCV*, pages 709–720, 1996.
21. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2):137–154, Nov 1992.
22. 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.
23. P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.
24. P. H. S. Torr and A. Zisserman. Robust computation and parameterization of multiple view relations. In *Proc. ICCV*, pages 727–732, Jan 1998.
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.